Contents

Viewpoint

SOMAScience: A Novel Platform for Multidimensional, Longitudinal Pain Assessment (e47177)
Chloe Gunsilius, Joseph Heffner, Sienna Bruinsma, Madison Corinha, Maria Cortinez, Hadley Dalton, Ellen Duong, Joshua Lu, Aisulu Omar, Lucy Owen, Bradford Roarr, Kevin Tang, Frederike Petzschner ................................. 41

Original Papers

Application of eHealth Tools in Anticoagulation Management After Cardiac Valve Replacement: Scoping Review Coupled With Bibliometric Analysis (e48716)
Ying Wu, Xiaohui Wang, Mengyao Zhou, Zhuoer Huang, Lijuan Liu, Li Cong ................................................................. 83

A Smartphone Food Record App Developed for the Dutch National Food Consumption Survey: Relative Validity Study (e50196)
Marga Ocké, Ceciel Dinnissen, Coline van den Bogaard, Marja Beukers, José Drijvers, Eline Sanderman-Nawijn, Caroline van Rossum, Ido Toxopeus .......................................................... 260

Feasibility Study on Menstrual Cycles With Fitbit Device (FEMFIT): Prospective Observational Cohort Study (e50135)
Anna-Lena Lang, Rosa-Lotta Bruhn, Maya Fehling, Anouk Heidenreich, Jonathan Reisdorf, Ifrah Khanyaree, Maike Henningsen, Cornelius Remschmidt ............................................. 277

Investigating Citizens’ Acceptance of Contact Tracing Apps: Quantitative Study of the Role of Trust and Privacy (e48700)
Grace Fox, Lisa van der Werff, Pierangelo Rosati, Theo Lynn ................................................................. 290

Investigating the Integration and the Long-Term Use of Smart Speakers in Older Adults’ Daily Practices: Qualitative Study (e47472)
Fangyuan Chang, Lin Sheng, Zhenyu Gu ................................................................. 303

Documentation Completeness and Nurses’ Perceptions of a Novel Electronic App for Medical Resuscitation in the Emergency Room: Mixed Methods Approach (e46744)
Kin Cheung, Chak-Yip ................................................................. 329

Health Care Professionals’ Experiences With a Mobile Self-Care Solution for Low Complex Orthopedic Injuries: Mixed Methods Study (e51510)
Jelle Spierings, Gijs Willinge, Marike Kokke, Bas Twigt, Wendela de Lange, Thijs Geerdink, Detlef van der Velde, Sjoerd Repping, Carel Goslings ................................................................. 339
The Effectiveness of a Cell Phone eHealth App in Changing Knowledge, Stigmatizing Attitudes, and Intention to Seek Help Associated With Obsessive-Compulsive Disorder: Pilot Questionnaire Study (e48027)
Antonio Chaves, Sandra Arnáez, Gemma García-Soriano ................................................................. 353
Evidence of How Physicians and Their Patients Adopt mHealth Apps in Germany: Exploratory Qualitative Study (e48345)
Tanja Schroeder, Maximilian Haug, Andrew Georgiou, Karla Seaman, Heiko Gewald ........................................ 367
Engagement With a Remote Symptom-Tracking Platform Among Participants With Major Depressive Disorder: Randomized Controlled Trial (e44214)
Katie White, Ewan Carr, Daniel Leightley, Faith Matcham, Pauline Conde, Yatharth Ranjan, Sara Simblett, Erin Dawe-Lane, Laura Williams, Claire Henderson, Matthew Hotopf ................................................................. 382
Effectiveness of a Nurse-Led Mobile-Based Health Coaching Program for Patients With Prostate Cancer at High Risk of Metabolic Syndrome: Randomized Waitlist Controlled Trial (e47102)
Kyounjin Lee, Jeongok Park, Eui Oh, JuHee Lee, Chang Park, Young Choi ................................................................. 402
Patient Engagement With and Perspectives on a Mobile Health Home Spirometry Intervention: Mixed Methods Study (e51236)
Andrew Liu, William Brown, Ill, Ndubuisi Madu, Ali Maiorano, Olivia Bigazzi, Eli Medina, Christopher Sorric, Steven Hays, Anobel Odisho ................................................................. 424
Mining User Reviews From Hypertension Management Mobile Health Apps to Explore Factors Influencing User Satisfaction and Their Asymmetry: Comparative Study (e55199)
Yunfan He, Wei Zhu, Tong Wang, Han Chen, Junyi Xin, Yongcheng Liu, Jianbo Lei, Jun Liang ................................................................. 436
Use and Engagement With Low-Intensity Cognitive Behavioral Therapy Techniques Used Within an App to Support Worry Management: Content Analysis of Log Data (e47321)
Paul Farrand, Patrick Raue, Earlise Ward, Dean Repper, Patricia Areán ................................................................. 476
A Behaviorally Informed Mobile App to Improve the Nutritional Quality of Grocery Shopping (SwapSHOP): Feasibility Randomized Controlled Trial (e45854)
Carmen Piernas, Charlotte Lee, Alice Hobson, Georgina Harmer, Sarah Payne Riches, Michaela Noreik, Susan Jebb ................................................................. 489
Effectiveness of Metaverse Space–Based Exercise Video Distribution in Young Adults: Randomized Controlled Trial (e46397)
Rami Mizuta, Noriaki Maeda, Tsubasa Tashiro, Yuta Suzuki, Sayo Kuroda, Ayano Ishida, Sakura Oda, Tomoya Watanabe, Yuki Tamura, Makoto Komiya, Yukio Urabe ................................................................. 503
Mobile App Intervention of a Randomized Controlled Trial for Patients With Obesity and Those Who Are Overweight in General Practice: User Engagement Analysis Quantitative Study (e45942)
Vera Buss, Margo Barr, Sharon Parker, Alamgir Kabir, Annie Lau, Siaw-Teng Liaw, Nigel Stocks, Mark Harris ................................................................. 516
The Association of Macronutrient Consumption and BMI to Exhaled Carbon Dioxide in Lumen Users: Retrospective Real-World Study (e56083)
Shlomo Yeshurun, Tomer Cramer, Daniel Souroujon, Merav Mor ................................................................. 528
The Roles of Trust in Government and Sense of Community in the COVID-19 Contact Tracing Privacy Calculus: Mixed Method Study Using a 2-Wave Survey and In-Depth Interviews (e48986)
Hyunjin Kang, Jeong Lee, Edmund Lee, Cindy Toh ................................................................. 537
Evaluation of Chinese HIV Mobile Apps by Researchers and Patients With HIV: Quality Evaluation Study (e52573)
Peng Liu, Lingmeng Wang, Fuzhi Wang ................................................................. 636
mHealth Apps for the Self-Management of Low Back Pain: Systematic Search in App Stores and Content Analysis (e53262)
Tianyu Zhou, David Salman, Alison McGregor. ................................................................. 654

Evaluation of Patient-Facing Mobile Apps to Support Physiotherapy Care: Systematic Review (e55003)
Mark Merolli, Jill Francis, Patrick Vallance, Kim Bennell, Peter Malliaras, Rana Hinman. ................................................................. 666

Short-Term Effect of a Health Promotion Intervention Based on the Electronic 12-Hour Dietary Recall (e-12HR) Smartphone App on Adherence to the Mediterranean Diet Among Spanish Primary Care Professionals: Randomized Controlled Clinical Trial (e49302)
Luis Béjar, Pedro Mesa-Rodriguez, María García-Perea. ................................................................. 693

Quality Assessment of Smartphone Medication Management Apps in France: Systematic Search (e54866)
Mickael Toïgo, Julie Marc, Maurice Hayot, Lionel Moulis, Francois Carbonnel. ................................................................. 708

The Association of eHealth Literacy Skills and mHealth Application Use Among US Adults With Obesity: Analysis of Health Information National Trends Survey Data (e46656)
George Shaw Jr, Bianca Castro, Laura Gunn, Keith Norris, Roland Thorpe Jr. ................................................................. 719

Barriers and Implications of 5G Technology Adoption for Hospitals in Western China: Integrated Interpretive Structural Modeling and Decision-Making Trial and Evaluation Laboratory Analysis (e48842)
Linyun Zhou, Minghuan Jiang, Ran Duan, Feng Zuo, Zongfang Li, Songhua Xu. ................................................................. 736

Views of Service Users, Their Family or Carers, and Health Care Professionals on Telerehabilitation for People With Neurological Conditions in Ghana: Qualitative Study (e49501)
Lorna Paul, Katie Thomson, Shadrack Asibey, Marian Brady, Frederike van Wijck, Derrick Antwi, Eric Opoku, Fred Sarfo. ................................................................. 753

Reflective Engagement With a Digital Physical Activity Intervention Among People Living With and Beyond Breast Cancer: Mixed Methods Study (e51057)
Michael Robertson, Emily Cox-Martin, Karen Basen-Engquist, Elizabeth Lyons. ................................................................. 765

Using In-Shoe Inertial Measurement Unit Sensors to Understand Daily-Life Gait Characteristics in Patients With Distal Radius Fractures During 6 Months of Recovery: Cross-Sectional Study (e55178)
Akiko Yamamoto, Eriku Yamada, Takuya Ibata, Fumiyuki Nihey, Takuma Inai, Kazuya Tsukamoto, Tomohiko Waki, Toshitaka Yoshii, Yoshiyuki Kobayashi, Kentaro Nakahara, Koji Fujita. ................................................................. 781

Effectiveness of Telemonitoring in Reducing Hospitalization and Associated Costs for Patients With Heart Failure in Finland: Nonrandomized Pre-Post Telemonitoring Study (e51841)
Jorma Kokkonen, Pirjo Mustonen, Eija Heikkilä, Rikka-Leena Leskelä, Paula Pennanen, Kati Krühn, Arto Jalkanen, Jussi-Pekka Laakso, Jari Kempers, Sami Väisänen, Paulus Torkkéli. ................................................................. 790

Investigating Receptivity and Affect Using Machine Learning: Ecological Momentary Assessment and Wearable Sensing Study (e46347)
Zachary King, Han Yu, Thomas Vaessen, Inez Myin-Germeys, Akane Sano. ................................................................. 815

Reviews

Dissemination Strategies for mHealth Apps: Systematic Review (e50293)
Henri Moungui, Huges Nana-Djeunga, Che Anyiang, Mireia Cano, Jose Ruiz Postigo, Carme Carrion. ................................................................. 98

Effects of mHealth-Based Lifestyle Interventions on Gestational Diabetes Mellitus in Pregnant Women With Overweight and Obesity: Systematic Review and Meta-Analysis (e49373)
Yirong He, Chuanhua Huang, Qiuyang He, Shujuan Liao, Biru Luo. ................................................................. 118
Effectiveness of Telecare Interventions on Depression Symptoms Among Older Adults: Systematic Review and Meta-Analysis (e50787)
Man Wu, Chaoyang Li, Ting Hu, Xueyang Zhao, Guiyuan Qiao, Xiaolian Gao, Xinhong Zhu, Fen Yang. .......................... 137

Effects of Digital Physical Health Exercises on Musculoskeletal Diseases: Systematic Review With Best-Evidence Synthesis (e50616)
Johanna Nagel, Florian Wegener, Casper Grim, Matthias Hoppe. .......................... 153

Mobile and Computer-Based Applications for Rehabilitation Monitoring and Self-Management After Knee Arthroplasty: Scoping Review (e47843)
Sabhya Pritwani, Purnima Shrivastava, Shruti Pandey, Ajit Kumar, Rajesh Malhotra, Ralph Maddison, Niveditha Devasenapathy. .......................... 171

User Perception of Smart Home Surveillance Among Adults Aged 50 Years and Older: Scoping Review (e48526)
Jessica Percy Campbell, Jacob Buchan, Charlene Chu, Andria Bianchi, Jesse Hoey, Shehroz Khan. .......................... 195

Mobile Apps for COVID-19 Detection and Diagnosis for Future Pandemic Control: Multidimensional Systematic Review (e44406)
Mehdi Gheisari, Mustafa Ghaderzadeh, Huxiong Li, Tania Taami, Christian Fernández-Campusano, Hamidreza Sadeghsalehi, Aaqif Afzaal Abbasi. .......................... 207

Lessons and Untapped Potential of Smartphone-Based Physical Activity Interventions for Mental Health: Narrative Review (e45860)
Emily Bernstein, Emma Wolfe, Brynn Huguenel, Sabine Wilhelm. .......................... 224

Functionality and Quality of Asthma mHealth Apps and Their Consistency With International Guidelines: Structured Search and Evaluation (e47295)
Billy Robinson, Eleni Proimos, Daniel Zou, Enying Gong, Brian Oldenburg, Katharine See. .......................... 600

Developing a Comprehensive List of Criteria to Evaluate the Characteristics and Quality of eHealth Smartphone Apps: Systematic Review (e48625)
Janette Ribaut, Annette DeVito Dabbs, Fabienne Dobbels, Alexandra Teynor, Elisabeth Mess, Theresa Hoffmann, Sabina De Geest. .......................... 619

Quality and Accessibility of Home Assessment mHealth Apps for Community Living: Systematic Review (e52996)
Jung-hye Shin, Rachael Shields., Jenny Lee, Zachary Skrove, Ross Tredinnick, Kevin Ponto, Beth Fields. .......................... 679

Corrigenda and Addenda

Correction: Mobile App Intervention of a Randomized Controlled Trial for Patients With Obesity and Those Who Are Overweight in General Practice: User Engagement Analysis Quantitative Study (e58507)
Vera Buss, Margo Barr, Sharon Parker, Alamgir Kabir, Annie Lau, Siaw-Teng Liaw, Nigel Stocks, Mark Harris. .......................... 804
An Introduction to Smart Home Ward–Based Hospital-at-Home Care in China

Weibin Cheng¹, MD; Xiaowen Cao¹, MSc; Wanmin Lian², MSc; Junzhang Tian¹, MD

¹Corresponding Author:
Junzhang Tian, MD

Abstract

Hospital-at-home has been gaining increased attention as a potential remedy for the current shortcomings of our health care system, allowing for essential health services to be provided to patients in the comfort of their own homes. The advent of digital technology has revolutionized the way we provide medical and health care, leading to the emergence of a "hospital without walls." The rapid adoption of novel digital health care technologies is revolutionizing remote health care provision, effectively dismantling the conventional boundary separating hospitals from the comfort of patients’ homes. The Guangdong Second Provincial General Hospital has developed a 5G-powered Smart Home Ward (SHW) that extends medical care services to the home setting and is tailored to meet the needs and settings of each patient’s household. The SHW was initially tested for its suitability for treating 4 specialized diseases, including cardiovascular disease, stroke, Parkinson disease, and Alzheimer disease. Understanding and addressing the potential challenges and risks associated with SHWs is essential for the successful implementation and maintenance of safe and effective home hospitalization.

(JMIR Mhealth Uhealth 2024;12:e44422) doi:10.2196/44422

KEYWORDS
smart home ward; telemonitoring; telemedicine; home care; hospital at home; healthcare delivery; implementation; smart ward; medical monitoring; medical care; rehabilitation; health care

The Future of Older Adult Care in China: Innovations in Health Care Delivery

China has emerged as a rapidly aging society. Aging contributes significantly to the health care burden in China due to the increased prevalence of chronic diseases and disabilities among older adults. The health care burden for the population of older individuals in China is further compounded by inadequate health care infrastructure and limited access to health care services in rural areas. The Chinese government has implemented various policies and initiatives to address these challenges, including expanding health care coverage, promoting preventative care, and increasing investment in health care infrastructure. Currently, the top-level design of China's older adult care model is “9073,” which means that 90% of older people receive older adult care services at home, 7% of them receive short-term care in the community, and 3% of them receive institutional care. This implies that the age-friendly and livable environment at home affects 90% of the population of older adults.

Hospital-at-home, which seeks to provide essential health services to patients in the comfort of their own homes, has been gaining increasing attention as a feasible solution for at-home older adult care and medical services [1-3]. The COVID-19 pandemic has underscored the risks of overreliance on physical medical institutions, emphasizing the continued need to develop a decentralized medical service ecosystem that revolves around patients’ families and communities. Hospital-at-home programs have been implemented in many high-income countries for years [4-7]; however, successful programs are limited [8]. Experience translating from high-income countries to low- and middle-income countries can be challenging, where medical resources are limited and public health literacy is low [2,3,9,10].

The Growth of Emerging Technologies for Home Hospitalization

New digital health care technologies are being rapidly adopted for remote health care provision, which is breaking down the traditional barrier between the hospital and home. The emergence of 5G technology posed the potential to further enhance home hospitalization by enabling remote patient monitoring and real-time communication between health care providers and patients. The introduction of sensor-based wearables and devices has changed the way that clinical data are collected and stored, leading to groundbreaking advancements in how care is provided [11,12]. Digital biomarkers—a set of objectives, quantifiable measures of physiological and behavioral characteristics that are acquired via wearables, implants, and other devices—are becoming increasingly essential in this process [13,14]. Home hospitalization can greatly benefit from the use of digital
biomarkers, providing care teams with a more comprehensive understanding of a patient’s health. These biomarkers enable the tracking of vital signs such as heart rate and respiration rate, as well as changes in sleep patterns, activity levels, and dietary habits [12,15,16].

**Smart Home Ward for Hospital-at-Home Care**

In 2021, Guangdong Second Provincial General Hospital developed the “Hospital Intelligent Twins,” a 5G-powered smart hospital that integrates the Internet of Things (IoT), artificial intelligence (AI), cloud computing, and 5G applications to create all-scenario intelligence for health care and hospital management [17,18]. Leveraging this 5G smart hospital infrastructure, Guangdong Second Provincial General Hospital further explored the Smart Home Ward (SHW), which seeks to break through the “wall” of centralized hospital-based health care services by extending care services to the home setting (Figure 1). The SHW is a seamless hospital unit managed by health professionals in the hospital and is designed to provide patients with equivalent hospital-level services at home, including medical monitoring, ward rounds, consultation, and medical care, and so on. SHW can support 2 hospital-at-home models, early-supported discharge and admission avoidance, offering integrated health monitoring, medical care, and rehabilitation services all within the comfort of their homes [19,20].

**Figure 1.** Demonstration of the Smart Home Ward in the hospital.
The management of patients in the SHW is handled by an interdisciplinary care team overseen by responsible specialists from relevant clinical departments. At present, the Departments of Cardiology and Neurology are pilot testing SHWs for patients with cardiovascular conditions and those with cerebrovascular conditions, respectively. Each specialty establishes a dedicated home ward team comprising physicians from that department, along with nurses, pharmacists, physical therapists, occupational therapists, and social workers. For example, the post–coronary heart disease treatment team includes cardiologists, cardiac nurses, and rehabilitation therapists. Patients undergo an initial clinical assessment by their specialist physician to determine their suitability for remote home care. The inclusion and exclusion criteria for specific conditions such as coronary heart disease and stroke are provided in Multimedia Appendix 1. If deemed feasible based on the assessment and if the patient provides consent, the interprofessional SHW staff conduct a home environment evaluation. Necessary modifications are made and monitoring equipment or devices installed to safely support care at home with remote specialist oversight.

In the SHW, series of digital biomarkers are monitored, including blood pressure, heart rate, respiratory rate, blood oxygen saturation, body temperature, and electrocardiographic signals, with patients’ daily activities, treatment compliance, and risk occurrences monitored via contact or contactless devices. These data can be leveraged to tailor treatments or notify care providers of any deviations from expected parameters.

Data generated from the SHW are transmitted securely to the hospital’s electronic medical record (EMR) system in real time. The integration follows privacy and security guidelines set forth by our hospital’s Health Insurance Portability and Accountability Act (HIPAA)–compliant policies. Only deidentified data points that are relevant for clinical care, such as vital signs, activity levels, and medication adherence data, are integrated into the EMR. Data transmission from the SHW to the EMR is one-way to ensure the security of sensitive hospital information. Authorized care team members, including attending physicians, nurses, pharmacists, physical therapists, and case managers, have access to consolidated patient data reports within the EMR system. This allows them to monitor trends, recognize any deviations from normal ranges, and act accordingly without needing to use separate systems.

The primary care team, led by the patient’s attending physician, is responsible for reviewing the daily reports and contacting patients if follow-up care is needed, based on the remote patient monitoring data. Family members and caregivers participate in care conferences to stay updated on the patient’s progress but do not have access to the EMR. We aim to expand access to aggregated reports to allow for greater caregiver involvement while maintaining privacy and security.

A pretraining ward has been established in the hospital to help patients transit from hospital to home living. Patients are trained on the use of smart home devices and receive guidance on remote rehabilitation training. In addition, the pretraining ward includes medical assessments to ensure that the patient is eligible for admission to SHW care and will benefit from the treatments provided at home. Specialized staff consisting of nurses, physical therapists, and occupational therapists are dedicated to supporting the SHW program. These clinicians provide personalized patient education, demonstrations, and skills training. They ensure that patients and caregivers are comfortable with the remote monitoring system and therapy program prior to discharge. The goal is to maximize treatment adherence and outcomes through empowering patients and families with knowledge. A dedicated tech support team is also available to assist families in successfully setting up the in-home system and addressing any technical issues that may arise post discharge. Readily available support from clinical and technical experts further enables safe and independent living at home with remote care and monitoring.

### Characteristics of the SHW

The SHW constructed in this project not only integrates advanced medical IoT technology and products, but also involves the renovation of the environment to make it suitable for older adults and patients, extending homogeneous medical services provided in the hospital wards to patients’ homes. Its characteristics are mainly reflected in the following 4 aspects.

First, the SHW has a multitude of functions including ambient sensing, medical monitoring, rehabilitation training, exercise and diet guidance, psychological counseling, and sleep management. To achieve these capabilities, a variety of cutting-edge technologies such as 5G or Wi-Fi 6, Internet of Medical Things (IoMT), smart wearable devices, smart home appliances, and health monitoring equipment have been deployed in the home setting. For instance, an intelligent closenowl can detect urine levels and monitor heart rate, body fat, and the length of time spent on the toilet [21]. Additionally, an intelligent mattress and pillow monitor heart rate, respiration rate, and body movement frequency during sleep [22,23]. Applications such as those for fall detection in the bathroom and the amount of stillness in daily activities can trigger an automatic alarm for an emergency, which connects to the hospital [24].

Second, to facilitate hospital-at-home management, a platform was developed with the integration of cross-system and cross-ecological IoMT devices (Figure 2; see the screenshot of the SHW Medical Management Information System). Thus, IoMT devices and ambient sensors are interconnected, and data are generated, gathered, managed, and processed by various terminals, including data from a hospital and a SHW (including data for the home environment, daily activities, treatment, rehabilitation, etc). This enables interoperability for home ward management, allowing patients, health professionals, and caregivers to use mobile terminals such as tablets to view the patient’s physiological data in real time and manage the home ward setting with ease. As medical teams have the capability to monitor a patient’s health from a distance, fewer in-person visits are needed, thus decreasing the cost of care [25].
Third, deployment of an SHW is tailored to cater the needs and settings of a patient’s household. To promote patient autonomy in the home setting, the SHW undergoes various levels of renovations suitable for aging individuals, creating a barrier-free, safe, and secure living space. Furthermore, ambient sensors enable automatic configuration to ensure the comfort of the environment [26]. These sensors continuously detect the air, temperature, and humidity in the ward. Whenever the indoor concentrations of particulate matter of diameter $\leq 2.5 \mu m$, formaldehyde, and carbon dioxide exceed standard levels, the SHW automatically adjusts its equipment to improve indoor air quality.

Fourth, the extension of hospital care services to the home setting is achieved through the integration of various technologies and resources. This challenge is addressed by the SHW through the deployment of advanced IoMT technology and products, which allow for the passive monitoring of patients’ health and the implementation of timely medical interventions. Additionally, a comprehensive EMR system has been implemented, enabling seamless sharing of patients’ medical records between the hospital and home health care systems. Furthermore, a robust 5G-based telemedicine infrastructure has been established, enabling real-time communication between patients and doctors, including digital consultations, remote monitoring, and telemedicine-enabled home visits, as well as providing on-demand home care services [27].
**SHW Care Plans for Specialized Diseases**

The SHW’s feasibility of treating 4 specialized diseases, namely Parkinson disease, cardiovascular disease, stroke, and Alzheimer disease, is being tested. For each disease, corresponding care plans have been developed, and necessary monitoring and treatment facilities have been suggested.

In the case of Parkinson disease, exercise rehabilitation treatment is mainly used as the primary treatment modality in the home ward. Patients are guided by doctors remotely to implement various forms of internet-based rehabilitation exercise training, such as relaxation training, joint range of motion training, muscle training, breathing training, gait training, balance training, and cognitive training. During exercise, the patient’s heart rate is monitored using a smart watch to prevent excessive exercise intensity. A smart lunch box is provided for daily monitoring of the frequency and amplitude of the patient’s hand tremors while eating to evaluate the effectiveness of rehabilitation treatment [28].

For patients with cardiovascular disease, the home ward caters to those with ST-segment elevation myocardial infarction, non–ST-segment elevation acute coronary syndrome, stable angina pectoris, ischemic cardiomyopathy, chronic systolic heart failure, and sudden cardiac death syndrome. Cardiac rehabilitation exercise therapy is guided by health professionals from the hospital, which includes aerobic exercise, resistance training, and neuromuscular training. A treadmill with heart and lung function monitoring is set up as a rehabilitation tool at home, which records and monitors the patient’s exercise status and heart rate, blood oxygen saturation, and calorie consumption. This allows doctors to adjust the program promptly [29].

For stroke, given that individuals who have experienced a stroke often face difficulties with mobility, the adaptation of living spaces to suit their needs is crucial to promote independent living and facilitate rehabilitation. To this end, we have implemented appropriate accessibility retrofits within the home setting to establish a barrier-free environment for patients. Such efforts can contribute to not only increased mobility but also heightened patient confidence. To evaluate the efficacy of functional rehabilitation, regular assessments are conducted through both home visits and hospital evaluations.

In the case of Alzheimer disease, a comprehensive treatment approach is adopted, including cognitive training, task training, and music therapy. These treatments are supported by the augmented reality and virtual reality applications that allow doctors to simulate complex scenarios to stimulate patients’ brain activities in a safe and controlled environment. Several physiological signals, such as heart rate variability, eye movements, and sleep patterns, are recorded and are used to train models for treatment evaluation. Furthermore, an AI-powered camera is installed in the home environment to monitor the patient’s activity status, such as sitting and lying time. A smart watch with GPS positioning and communication functions is also used to prevent patients from getting lost [30].

**Challenges for the Scale-Up of the SHW Care Program in China**

The widespread adoption of SHWs has been faced with several challenges. First, authoritative guidelines for the implementation of medical service standards for home ward care and those for setting up technology-enabled digital wards in China are lacking. Second, a sustainable fee-based model for home ward care has yet to be established. Third, the effective operation of home ward care requires close collaboration among hospitals, community health service organizations, and family members. However, there are currently no unified regulations to delineate the responsibilities of various stakeholders involved in providing home ward care. Fourth, the popularity and reliability of the technology needs improvement. SHWs require specialized equipment and technologies, a secure and trust-based environment, and staffing competencies to ensure patient safety and privacy. A unique standard needs to be established to ensure the interconnectedness and interoperability of various devices. Meanwhile, potential problems associated with the cost of equipment, access to necessary data, and data privacy also exist.

**Implications of Hospital-at-Home Care Using SHWs for Practitioners, Researchers, and Policy Makers**

The implementation of SHWs has the potential to extend the quantity and quality of hospital care services in response to the increasing demand for medical care in China’s aging society. For health care practitioners, the integration of smart home digital health technologies can facilitate remote patient monitoring and management, enabling timely and effective care delivery in patients’ homes. This approach can reduce the burden on hospitals and clinics, while improving patient outcomes and satisfaction. To enable the successful implementation of SHWs, new team structures with corresponding workflows must be created in clinical settings to optimize health care systems and patients’ usage of this technology. Clinical physicians must work closely with technicians in developing workflows and integrated AI tools and in the process of care provision, such as remote monitoring.

For researchers, the implementation of SHWs presents a unique opportunity to study the impact of digital health technology on health care delivery, patient outcomes, and cost-effectiveness. However, ethical considerations must be navigated when accessing the vast amounts of data gathered at SHWs to ensure patient privacy and security [31]. All stakeholders must understand and adapt to these implications to fully harness their potential benefits while mitigating associated risks.

Policy makers can benefit from the use of SHW technologies, as they show promising potential in reducing health care costs and improving access to care, particularly for patients in rural or underserved areas. However, policy makers must establish regulations to safeguard patient privacy and security, ensure equitable access to technology, and enhance digital infrastructure for widespread adoption. With careful consideration of these factors, the implementation of SHW can revolutionize...
hospital-at-home care, improving the quality of care, reducing health care costs, and ultimately benefiting patients and the health care system as a whole.

**In Summary**

The advent of digital technology has revolutionized the way we provide medical and health care, leading to the emergence of “hospitals without walls.” This SHW is an innovative care model that promises to bring a host of improvements to health care by providing convenient access to “boundless ambulatory care,” as well as “boundless inpatient care.” Despite the potential benefits, these digital projects are still faced with some challenges, such as public acceptance and adoption of the technology, willingness to pay for services, and encouraging medical insurance uptake. Therefore, it is essential for governments to promote public awareness of the advantages of digital technology and introduce incentives that motivate people to take up medical insurance in order for these services to be widely available.

**Acknowledgments**

This study was funded by the National Key R&D (research and development) Program of China (number 2021YFC2009400).

**Authors' Contributions**

WC, XC, and WL conceived the study and drafted the manuscript. JT provided professional support and made critical revisions to the manuscript. All authors read and approved the final manuscript.

**Conflicts of Interest**

None declared.

**Multimedia Appendix 1**

Inclusion/exclusion criteria for specific conditions. [DOCX File, 18 KB - mhealth_v12i1e44422_app1.docx]

**References**


Abbreviations

- **AI**: artificial intelligence
- **EMR**: electronic medical record
- **HIPAA**: Health Insurance Portability and Accountability Act
- **IoMT**: Internet of Medical Things
- **IoT**: Internet of Things
- **SHW**: Smart Home Ward
Advances and Opportunities of Mobile Health in the Postpandemic Era: Smartphonization of Wearable Devices and Wearable Deviceization of Smartphones

Wonki Hong, PhD
Department of Digital Healthcare, Daejeon University, Daejeon, Republic of Korea

Corresponding Author:
Wonki Hong, PhD

Abstract
Mobile health (mHealth) with continuous real-time monitoring is leading the era of digital medical convergence. Wearable devices and smartphones optimized as personalized health management platforms enable disease prediction, prevention, diagnosis, and even treatment. Ubiquitous and accessible medical services offered through mHealth strengthen universal health coverage to facilitate service use without discrimination. This viewpoint investigates the latest trends in mHealth technology, which are comprehensive in terms of form factors and detection targets according to body attachment location and type. Insights and breakthroughs from the perspective of mHealth sensing through a new form factor and sensor-integrated display overcome the problems of existing mHealth by proposing a solution of smartphonization of wearable devices and the wearable deviceization of smartphones. This approach maximizes the infinite potential of stagnant mHealth technology and will present a new milestone leading to the popularization of mHealth. In the postpandemic era, innovative mHealth solutions through the smartphonization of wearable devices and the wearable deviceization of smartphones could become the standard for a new paradigm in the field of digital medicine.

(JMIR Mhealth Uhealth 2024;12:e48803) doi:10.2196/48803

KEYWORDS
mobile health; mHealth; smartphonization; wearable deviceization; new form factor; sensor-integrated display

Background
In the postpandemic era, the significance of mobile health (mHealth) has been highlighted, and explosive growth in this area is expected to continue [1,2]. Cutting-edge technologies are converging with health care, and mHealth, based on hyperconnected intelligence, is leading the paradigm shift in medical care [3,4]. Many countries have already entered a superaged society, and the proportion of gross domestic product expenditures for medical care is increasing due to an upsurge in the number of people with chronic diseases. In addition, the excessive demand compared to the available supply, the lack of health care infrastructure, and the unbalanced distribution of medical staff are also problems. Therefore, prediction, prevention, and management through artificial intelligence (AI)–based medical big data analysis are required, and for this purpose, ubiquitous and accessible medical services using personalized devices must be provided [5]. mHealth is a strong candidate to make this possible, and the ultimate goal is to dramatically improve the standard and satisfaction of living by providing quality services at affordable prices [6,7].

Wearable electronics and smartphones are representative types of mobile systems optimized for personalized health care sensing. As shown in Figure 1 [8-19], wearable devices that cover the human body and smartphones, a necessity for modern people, enable comprehensive health management in real time.
Figure 1. Application and placement schematic illustration of wearable devices by body part and smartphone for mHealth management. The images were reprinted from Shin et al [8,19], Kim et al [9,19], Escobedo et al [10,19], Hwang et al [11], Nakamura et al [12,19], Hua et al [13,19], Moon et al [14,19], Zhao et al [15,19], Kim et al [16,19], Liu et al [17,19], and Chan et al [18,19]. e-Skin: electronic skin; mHealth: mobile health.

However, the pace of the development and popularization of mHealth technology is progressing more slowly than expected. From the perspective of a paradigm shift from the smartphonization of wearable devices and the wearable deviceization of smartphones, this viewpoint aimed to propose ways to unleash the potentiality of mHealth in the postpandemic era. The smartphonization of wearable devices and the wearable deviceization of smartphones do not simply mean that current smartphones become wearable devices and that current wearable devices maintain the functions of current smartphones. The smartphonization of wearable devices is to completely replace the smartphone function with a wearable device, while upgrading health care performance by embedding the current smartphone’s computational power and sensor-integrated display, including large-area panels and user interaction, in the wearable system. In addition, the wearable deviceization of smartphones refers to a change in the form factor so that health care sensing can be performed by switching from the current rigid form to a form that can be attached to a curved skin surface. The new form factor, which features both wearable computer and smartphone functions, will improve detection performance through large-area sensing and increase the penetration rate.

This viewpoint investigated recent trends in health care sensing methods using wearable devices and smartphones, which are
the central axis of mHealth. In the case of wearable devices, the form factor for each detailed location on the body and the corresponding detection target technology was described. In the case of smartphones, it covered the detection target and principles of health care according to the application of internal and external sensors, materials, and software. This viewpoint also analyzed the prospects of and current challenges in existing mHealth systems and considered new health care solutions using flexible displays for the convergence form factor of smartphones and wearables. The differentiating point was to consider the direction of mHealth from the perspective of a sensor-integrated and new form factor display. Ultimately, from a display perspective, solutions for the smartphonization of wearable devices and the wearable deviceization of smartphones will provide insight into the health care paradigm shift.

Recent Progress in the Development of Wearable Electronics for Health Care

The primary classification of wearable electronics based on the attachment position can be divided into the face, upper body, limbs, and whole body. Wearable clothes all over the body can also be classified separately.

Face

Head

The face, which is closest to the brain, is significant from a sensory point of view because it is where the 5 senses are concentrated. Face-wearable devices with various form factors, such as bands, caps, headsets, lenses, glasses, tattoos, mouthguards, and masks, may be distributed at each part of the head, eyes, nose, mouth, and ears to sense critical biosignals. In the case of the head, a wearable system that can analyze brain waves and psychological states can be applied [8,20,21]. Figure 2A [8] shows a wireless wearable electroencephalogram (EEG) measurement device based on a tattoo. AI can enhance decision-making by deep learning classification of received EEG data. Namely, it advances the decision performance of AI by feedback through brain waves. Additionally, it would be possible to grasp the degree of brain activation and mental condition of the frontal lobe and temporal lobe through the measurement of biosignals, such as brain waves.
Eyes and Nose

System form factors worn on the eye may be divided mainly into lenses and glasses. In the case of lenses, eye health factors, such as glucose, intraocular pressure, and electroretinographic measurements, can be determined using noninvasive methods [9,22,24-26]. For example, a corneal sensor embedded in a disposable soft contact lens can be deployed for electroretinography based on electrochemical anchoring, as shown in Figure 2B [22]. These corneal lenses are functional sensors tailored for ophthalmic electroretinographic testing in human eyes via a user-friendly interface and a design that can be deployed noninvasively. Glasses for health care are prescribed by doctors as an auxiliary tool for surgery and can also analyze the electrolyte and metabolite content of sweat flowing from the head [27-29]. In addition, a wearable system placed on the nose in the form of a nose pad on the glasses can sense the pulse wave, respiratory rate, and electrooculographic measurements [30,31].
**Mouth and Ears**

Wearable electronics related to the mouth take the form of mouth guards, tooth sensors, and masts and can analyze saliva and nutrients and monitor air quality [10,23,32-35]. For example, a small stretchable circuit and sensor that can be inserted into the human oral cavity may be integrated into a breathable, flexible microporous membrane for a tissue-friendly design, as shown in Figure 2C [23]. Such a device may be used in research to study the prevention of hypertension by facilitating continuous quantification analysis of sodium intake. Figure 2D [10] shows a sensing platform for detecting gaseous CO$_2$ inside a face mask via stable inorganic phosphors whose luminescence is controlled by a pH indicator. A mask combining a battery-free printed near-field communication (NFC) tag and a photochemical sensor for noninvasive CO$_2$ measurement was used to achieve detection performance with a resolution of 103 ppm. Practicality in physical activity has been increased through the compensation of the temperature noise and characterized analytical specifications of measurement systems. Moreover, health care wearable systems attached to the ears use earbuds to perform heart rate and sleep monitoring functions [36,37].

**Upper Body**

In addition to the face, wearable systems can be applied to the neck, chest, abdomen, internal organs, back, and waist to extract significant health values.

**Neck**

In the case of the neck, wearable devices with a necklace and patch form factor can record an electrocardiogram (ECG) and voice pressure and monitor the diet through an electroglottogram (EGG) using a neckband [38-40]. For example, a neck-attached wearable device incorporating a cross-linked polymer film and hole-patterned diaphragm structure detects and quantifies voice with an excellent sensitivity of 5.5 V Pa$^{-1}$ over the voice frequency range, as shown in Figure 3A [39]. This device can be used for voice health management and security authentication by eliminating vibration distortions on the curved skin surface through excellent skin compatibility via using ultrathin profiles of ≥5 µm.
Figure 3. Wearable electronics mounted onto the upper body. (A) Vibration-responsive patch for sensing voice pressure. The images were reprinted from Lee et al [19,39]. (B) Epidermal cardiopulmonary patch based on laser fabrication. The images were reprinted from Rachim et al [41]. (C) Air-silicon composite transducer (ASiT) for breathing pattern monitoring. The images were reprinted from Cotur et al [19,42]. (D) Spine tracker sensor system. The image was reprinted from Stollenwerk et al [19,43]. (E) A belt for waistline measurement. The images were reprinted from Nakamura et al [12,19]. EPE: electrophysiological electrode; MES: mechano-acoustic sensor; PCB: printed circuit board.

Thorax

Thorax-related wearable electronics, such as patches, chest belts, and brassieres, enable ECG recording, temperature measurement, sleep monitoring, posture analysis, and galvanic skin response (GSR) assessment [11,41,44-48]. Figure 3B [41] shows a sensor designed for continuous monitoring of the cardiopulmonary biosignal via a CO$_2$ laser–based manufacturing process. The epidermal patch consisting of a mechnoacoustic sensor and electrophysiological electrodes provides advanced functionality through a gas-permeable and biocompatible layer.

Abdomen

Abdomen-attached mHealth systems can sense glucose and breathing patterns through patches and straps [42,49]. For instance, an air-silicon composite transducer monitors respiratory activity by continuously measuring the force applied to the air channel embedded in the silicon-based elastomer, as shown in Figure 3C [42]. The system, which uses a pressure sensor and mixed-signal radio electronics, follows the principle of sensing the air pressure change inside the channel when breathing force is applied to the transducer surface. In particular, tactile sensing, including pressure sensing, is critical in health...
care. This is because tactile sensors attached to the skin detect physical stimuli, such as breathing patterns, heart rate, pulse, muscle activity, and body temperature, linked to biological signals. Skin, the most widely distributed organ among the five sense organs in the human body, is a tactile sensor with receptors that detect pressure, delicate movements, and temperature and is also an actuating organ that emits the same physical stimulation. Flexibility is a crucial element for the tactile sensor to be conformally attached to the skin to detect minute physical changes in detail and increase user convenience [50-52].

Furthermore, digestible pills check medication compliance. Management of medication adherence can prevent patients with severe mental illness from experiencing relapses and hospitalizations [53]. In addition, capsule endoscopy can monitor the colon health or bladder pressure state [54,55].

**Back**
A wearable system attached to the back can be used to analyze changes in the spine’s shape during training. A spine tracker device shown in Figure 3D consists of 5 sensors, with each sensor attached to the lumbar spine, and can correct posture by providing real-time feedback [43].

**Waist**
In addition, a waist belt can be useful for obesity management [12,56]. The belt automatically measures waist circumference with high accuracy, with an $F_1$-score of 0.95, and monitors the daily lifestyle using a magnetometer, an accelerometer, and a gyroscope, as shown in Figure 3E [12].

**Limbs**
In the case of the limbs, the main categories include the hands, arms, legs, and feet by attachment location.

**Hands**
The measurable health factors in a hand-related wearable device, such as a patch, ring, or glove, include rehabilitation evaluation analysis, ECG characteristics, oxygen saturation, dietary monitoring, pulse wave, and temperature [13,57-61]. For instance, a multisensory electronic skin integrated into a polyimide network simultaneously detects physical properties, such as temperature, strain, humidity, light, magnetic field, pressure, and proximity, in real time, as shown in Figure 4A [13]. It can also be used for rehabilitation evaluation using personalized intelligent prostheses.
**Arms**

mHealth systems of various form factors related to the arm can also be useful for health management. Among them, wristwatches, bands, and bracelet devices can detect health factors, such as the heart rate, oxygen saturation, number of steps, blood pressure, ECG characteristics, glucose, blood sugar, and sweat metabolites [14,15,65-75]. Figure 4B [15] shows a magnetoeelastic generator that provides the power to drive a wearable biosensor system. This generator can help measure cardiovascular parameters underwater without encapsulation for telemedicine and has excellent water vapor transmission characteristics.

A patch sensor attached to the arm can measure the pH, sweat rate, lactate, heart rate, temperature, electromyogram (EMG) and ECG characteristics, blood pressure, and water content and...
can also be applied for wound treatment and rehabilitation evaluation [62-64,76-85]. For instance, a stand-alone organic skin patch for health care with an organic light-emitting display with sufficient pixels reports the heart rate via a stretchable photoplethysmogram (PPG) sensor, as shown in Figure 4C [62]. An ultrathin patch of 15 μm is configured on a soft elastomer substrate and can operate stably at 30% strain using a combination of a stress relief layer and deformable microcracks. Figure 4D [63] shows a wireless graphene patch that simultaneously provides thermal sensing and thermotherapy capabilities. This thermal patch consists of a graphene-based capacitive sensor, a graphene thermal pad, and a flexible wireless communication module to continuously monitor temperature changes with high resolution and sensitivity and perform thermal treatment through a graphene-based heater. Beyond the existing complex multisensor structure, skin patches alone may decode movements of 5-finger gestures by detecting microdeformation using the laser-induced crack structure, as shown in Figure 4E [64]. Based on the same principle, it can be attached to various body parts to track physical movements.

Furthermore, ECG, EMG, temperature, sweat, and interstitial fluid analyses can be performed following health care monitoring through arm tattoos [16,86]. For instance, a noninvasive epidermal biosensing system includes physically separated electrochemical biosensors for the extraction of interstitial fluid at the cathode and sweat stimulus extraction at the anode, as shown in Figure 4F [16]. Namely, this biomarker monitoring system is a single wearable epidermal platform that simultaneously samples and analyses different biofluids.

**Legs and Feet**

Figure 5 describes a wearable health care device that may be applied to the legs, feet, or whole body. The mobile form factors applicable to the legs include patches, wearable robots, and straps, which perform moisture analysis at the wound area, gait analysis, ECG measurement, and rehabilitation evaluation [17,87-92]. For instance, appropriate dressing changes for exudative wounds are essential. Using a moisture sensor mounted on the bandage, as shown in Figure 5A [89], the change in the amount of dressing on the wound can be detected and the replacement time determined, increasing patient convenience. A motion capture device can accurately measure the movement of limbs during daily activities, strenuous exercise, and long-term exercise, as shown in Figure 5B [17]. Existing drift and instability problems are solved by integrating microtriaxis inertial and microtriaxis flow sensors. Additionally, it is possible to evaluate gait performance on irregular and uneven surfaces using a wearable sensor in the form of a strap with 6 inertial measurement units (IMUs) and an analysis algorithm, as shown in Figure 5C [92]. It is possible to implement edema measurement, gait analysis, and ulcer detection via plantar pressure analysis using wearable sensors attached to the shoes, socks, or soles of the feet [93-97].
Whole Body

Furthermore, clothes worn on the whole body are also a type of wearable device. **Figure 5D** [98] shows a personalized and conformable suit of an electronics-based textile for multimodal health care sensing. The platform’s elasticity ensures intimate contact between the electronic device and the skin, and it can detect the skin temperature, heart rate, and respiration with high accuracy and precision. The suit with electronic textiles can measure the body temperature, respiratory rate, heart rate, oxygen saturation, and EMG and ECG characteristics and can also perform phototherapy [98-103]. As described before, form factor and detection targets by body part on wearable devices are summarized in Tables 1-4.
Table. Summary of form factor and detection targets on wearable devices for the face.

<table>
<thead>
<tr>
<th>Body position and form factor</th>
<th>Target(s) of detection</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Head</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tattoo</td>
<td>EEG&lt;sup&gt;a&lt;/sup&gt;</td>
<td>[8]</td>
</tr>
<tr>
<td>Band, cap, headset</td>
<td>Mental stress through EEG</td>
<td>[20]</td>
</tr>
<tr>
<td>Band, cap, headset</td>
<td>EEG</td>
<td>[21]</td>
</tr>
<tr>
<td><strong>Eyes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lenses</td>
<td>Glucose</td>
<td>[24,25]</td>
</tr>
<tr>
<td>Lenses</td>
<td>Intracocular pressure</td>
<td>[9,26]</td>
</tr>
<tr>
<td>Lenses</td>
<td>Electoretinogram</td>
<td>[22]</td>
</tr>
<tr>
<td>Glasses</td>
<td>Auxiliary surgical tool</td>
<td>[27,28]</td>
</tr>
<tr>
<td>Glasses</td>
<td>Sweat electrolytes, metabolites</td>
<td>[29]</td>
</tr>
<tr>
<td><strong>Nose</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nose pad</td>
<td>Pulse wave, respiration rate</td>
<td>[30]</td>
</tr>
<tr>
<td>Nose pad</td>
<td>Electrooculogram</td>
<td>[31]</td>
</tr>
<tr>
<td><strong>Mouth</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouthguard</td>
<td>Saliva monitoring</td>
<td>[32-34]</td>
</tr>
<tr>
<td>Mouthguard</td>
<td>Nutrition analysis</td>
<td>[23]</td>
</tr>
<tr>
<td>Tooth sensor</td>
<td>Nutrition analysis</td>
<td>[35]</td>
</tr>
<tr>
<td>Mask</td>
<td>Air quality monitoring</td>
<td>[10]</td>
</tr>
<tr>
<td><strong>Ears</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Earbuds</td>
<td>Heart rate</td>
<td>[36]</td>
</tr>
<tr>
<td>Earbuds</td>
<td>Sleep monitoring using EEG</td>
<td>[37]</td>
</tr>
</tbody>
</table>

<sup>a</sup>EEG: electroencephalogram.
<table>
<thead>
<tr>
<th>Body position and form factor</th>
<th>Target(s) of detection</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Neck</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Necklace</td>
<td>ECG&lt;sup&gt;a&lt;/sup&gt;</td>
<td>[38]</td>
</tr>
<tr>
<td>Patch</td>
<td>Voice pressure</td>
<td>[39]</td>
</tr>
<tr>
<td>Band</td>
<td>EGG&lt;sup&gt;b&lt;/sup&gt;</td>
<td>[40]</td>
</tr>
<tr>
<td><strong>Thorax</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patch</td>
<td>ECG</td>
<td>[11,41,44]</td>
</tr>
<tr>
<td>Patch</td>
<td>ECG, temperature</td>
<td>[45]</td>
</tr>
<tr>
<td>Patch</td>
<td>Sleep monitoring</td>
<td>[46]</td>
</tr>
<tr>
<td>Chest belt</td>
<td>Trunk posture</td>
<td>[47]</td>
</tr>
<tr>
<td>Brassiere</td>
<td>Galvanic skin response</td>
<td>[48]</td>
</tr>
<tr>
<td><strong>Abdomen</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patch</td>
<td>Glucose</td>
<td>[49]</td>
</tr>
<tr>
<td>Strap</td>
<td>Respiratory patterns</td>
<td>[42,50-52]</td>
</tr>
<tr>
<td><strong>Internal organs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ingestible pill/capsule</td>
<td>Medication compliance</td>
<td>[53]</td>
</tr>
<tr>
<td>Ingestible pill/capsule</td>
<td>Intravesical pressure and colon monitoring</td>
<td>[54,55]</td>
</tr>
<tr>
<td><strong>Back</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strap</td>
<td>Spine monitoring</td>
<td>[43]</td>
</tr>
<tr>
<td><strong>Waist</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Belt</td>
<td>Obesity management</td>
<td>[12,56]</td>
</tr>
</tbody>
</table>

<sup>a</sup>ECG: electrocardiogram.

<sup>b</sup>EGG: electroglottogram.
<table>
<thead>
<tr>
<th>Body position and form factor</th>
<th>Target(s) of detection</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hands</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patch</td>
<td>Rehabilitation</td>
<td>[13]</td>
</tr>
<tr>
<td>Ring</td>
<td>ECG&lt;sup&gt;a&lt;/sup&gt;</td>
<td>[57]</td>
</tr>
<tr>
<td>Ring</td>
<td>SpO&lt;sub&gt;2&lt;/sub&gt;&lt;sup&gt;b&lt;/sup&gt;</td>
<td>[58]</td>
</tr>
<tr>
<td>Ring</td>
<td>Dietary management</td>
<td>[59]</td>
</tr>
<tr>
<td>Ring</td>
<td>Pulse wave, temperature</td>
<td>[60]</td>
</tr>
<tr>
<td>Glove</td>
<td>Rehabilitation</td>
<td>[61]</td>
</tr>
<tr>
<td><strong>Wrist</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Watch/bracelet</td>
<td>Heart rate, step number</td>
<td>[65]</td>
</tr>
<tr>
<td>Watch/bracelet</td>
<td>SpO&lt;sub&gt;2&lt;/sub&gt;</td>
<td>[66]</td>
</tr>
<tr>
<td>Watch/bracelet</td>
<td>SpO&lt;sub&gt;2&lt;/sub&gt;, heart rate, energy expenditure</td>
<td>[67,68]</td>
</tr>
<tr>
<td>Watch/bracelet</td>
<td>Blood pressure</td>
<td>[14,69]</td>
</tr>
<tr>
<td>Watch/bracelet</td>
<td>Pulse management</td>
<td>[15]</td>
</tr>
<tr>
<td>Watch/bracelet</td>
<td>ECG</td>
<td>[70,71]</td>
</tr>
<tr>
<td>Watch/bracelet</td>
<td>Diagnosis of Parkinson disease</td>
<td>[72]</td>
</tr>
<tr>
<td>Watch/bracelet</td>
<td>Glucose</td>
<td>[73,74]</td>
</tr>
<tr>
<td>Watch/bracelet</td>
<td>Sweat metabolites (glucose, lactate)</td>
<td>[75]</td>
</tr>
<tr>
<td>Patch</td>
<td>Sweat rate, pH, lactate, glucose, chloride</td>
<td>[76]</td>
</tr>
<tr>
<td>Patch</td>
<td>Heart rate</td>
<td>[62]</td>
</tr>
<tr>
<td>Patch</td>
<td>Wound management</td>
<td>[77,78]</td>
</tr>
<tr>
<td>Patch</td>
<td>Temperature, thermotherapy</td>
<td>[63]</td>
</tr>
<tr>
<td>Patch</td>
<td>ECG, EMG&lt;sup&gt;c&lt;/sup&gt;</td>
<td>[79-81]</td>
</tr>
<tr>
<td>Patch</td>
<td>EMG</td>
<td>[82,83]</td>
</tr>
<tr>
<td>Patch</td>
<td>Blood pressure, skin hydration, temperature</td>
<td>[84]</td>
</tr>
<tr>
<td>Patch</td>
<td>Biometrics</td>
<td>[85]</td>
</tr>
<tr>
<td>Patch</td>
<td>Rehabilitation</td>
<td>[64]</td>
</tr>
<tr>
<td>Tattoo</td>
<td>ECG, EMG, temperature</td>
<td>[86]</td>
</tr>
<tr>
<td>Tattoo</td>
<td>Sweat and Interstitial fluid analysis</td>
<td>[16]</td>
</tr>
<tr>
<td><strong>Legs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patch</td>
<td>ECG</td>
<td>[87]</td>
</tr>
<tr>
<td>Patch</td>
<td>Moisture analysis at the wound area</td>
<td>[88,89]</td>
</tr>
<tr>
<td>Wearable robot</td>
<td>Rehabilitation</td>
<td>[17,90,91]</td>
</tr>
<tr>
<td>Strap</td>
<td>Gait analysis</td>
<td>[92]</td>
</tr>
<tr>
<td><strong>Feet</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patch</td>
<td>Edema</td>
<td>[93]</td>
</tr>
<tr>
<td>Shoes</td>
<td>Gait analysis</td>
<td>[94-96]</td>
</tr>
<tr>
<td>Socks</td>
<td>Foot pressure ulcer</td>
<td>[97]</td>
</tr>
</tbody>
</table>

<sup>a</sup>ECG: electrocardiogram.

<sup>b</sup>SpO<sub>2</sub>: oxygen saturation.

<sup>c</sup>EMG: electromyogram.
### Table. Summary of form factor and detection targets on wearable devices for the whole body (clothes using electronic textiles).

<table>
<thead>
<tr>
<th>Target(s) of detection</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature, respiration, heart rate</td>
<td>[98]</td>
</tr>
<tr>
<td>SpO\textsubscript{2}\textsuperscript{a}, heart rate, temperature</td>
<td>[99]</td>
</tr>
<tr>
<td>Phototherapy, temperature, heart rate</td>
<td>[100,101]</td>
</tr>
<tr>
<td>EMG\textsuperscript{b}</td>
<td>[102]</td>
</tr>
<tr>
<td>ECG\textsuperscript{c}</td>
<td>[103]</td>
</tr>
</tbody>
</table>

\textsuperscript{a}SpO\textsubscript{2}: oxygen saturation.  
\textsuperscript{b}EMG: electromyogram.  
\textsuperscript{c}ECG: electrocardiogram.

#### Recent Progress in the Development of Smartphone-Based Health Care Apps

In addition to wearable devices, health care delivery is also possible using smartphones through built-in sensors, smartphone-interlocked gadgets, display-related materials, and apps.

**CMOS Only**

Smartphones have built-in 20-30 sensors; in particular, complementary metal-oxide-semiconductor (CMOS) image sensors may be used to monitor heart, eye, and skin-related diseases [18,104-109]. As shown in Figure 6A [18], the atrial fibrillation screening ability using PPG pulse analysis based on a smartphone camera and a commercialized app showed a similar performance level to that of patches used for single-lead ECG monitoring. It has been proven that prodromal stroke symptoms can be detected using only a smartphone in a primary care setting. In addition, the fingertip motion signal and color intensity signal, both heterogeneous signals, are acquired and analyzed using a camera to remove finger movement and optical noise, as shown in Figure 6B [104]. In this way, a clean heart rhythm signal with high accuracy can be extracted via smartphone monitoring, while minimizing noise artifacts.
Figure 6. Health care apps using built-in smartphone sensors. (A) Smartphone built-in camera and app-based atrial fibrillation diagnosis. The images were reprinted from Chan et al [18,19]. (B) Heart rhythm analysis using CMOS image sensor. The images were reprinted from Tabei et al [19,104]. (C) Smartphone-based blood pressure measurement through the oscillometric finger-pressing method. The images were reprinted from Chandrasekhar et al [19,110]. (D) Set and acquisition graph of smartphone and 3D-printed mouthpiece adapter for spirometry. The images were reprinted from Thap et al [19,111]. CMOS: complementary metal-oxide-semiconductor; PPG: photoplethysmogram.

Hybrid Including CMOS

New functions, such as blood pressure measurement and temperature and dietary monitoring, can be established by combining pressure sensors, temperature sensors, and the phone microphone instead of CMOS alone [110,112-115]. For instance, as shown in Figure 6C [110], absolute blood pressure is measured via a blood flow oscillometric signal through finger pressure using a strain gauge on the front of the smartphone, in addition to CMOS. A light-emitting display may also be added to this, so it is possible to measure blood pressure ultimately with pure smartphone components.

IMU/Microphone/Ultrasonic Sensor

In addition, sleep position monitoring and treatment can be performed by detecting body movements through an IMU of the smartphone, and the gait of patients with Parkinson disease can also be analyzed [116-118]. The smartphone’s built-in microphone sensor can also assess lung capacity and breathing sounds and monitor sleep [111,119-121]. Figure 6D [111] reports lung capacity and function parameter measurements.
following smartphone microphone–based, high-resolution time-frequency spectral analysis. A moisture-resistant ultrasonic sensor using polyvinylidene fluoride can be used for biometric authentication through fingerprinting [122].

**Touch Sensor/Digitizer**

Moreover, general user interfaces, such as a touch sensor and digitizer, can also be used for health care purposes. For example, the heart rate can be checked by assessing capacitance changes according to the heartbeat with a capacitive touch sensor. The touch sensor is also helpful in diagnosing Parkinson disease through touch accuracy analysis [123,124]. In addition, a digitizer for writing can be applied to biometric authentication through handwritten signature recognition [125,126].

**Interlocked Gadgets**

There is a case of combining various mHealth sensing techniques, such as pesticide analysis, otitis media diagnosis, malaria infection detection, and ECG measurement, by adding a separate gadget rather than using just the smartphone itself [127-133]. The platform shown in Figure 7A [128] performs a visual, quantitative analysis of pesticides using an optical system that combines a dark cavity and an ultraviolet lamp with a smartphone. In other words, integrating a smartphone and a gadget-based paper strip enables real-time and on-site food evaluation. Additionally, it was confirmed that the diagnosis of acute otitis media is possible with the same level of accuracy as that attained with existing otoscopes through the combination of a commercialized optical system and a camera in a smartphone, as shown in Figure 7B [130]. Figure 7C [132] shows a smartphone-based immunodiagnostic platform that performs a chemiluminescence-based enzyme-linked immunosorbent assay using a lyophilized chemiluminescence reagent. This hand-held point-of-care-testing analyzer can detect active malaria infections with a sensitivity of 8 ng/mL.
Figure 7. Health care apps using gadgets mounted on smartphones. (A) Smartphone platform for pesticide evaluation of food, integrated with an ultraviolet lamp and a dark cavity by 3D printing. The images were reprinted from Chu et al [128]. (B) Smartphone otoscope for diagnosis of acute otitis media. The images were reprinted from Mousseau et al [130]. (C) Smartphone-based immunodiagnosis using microfluidic assays. The images were reprinted from Ghosh et al [19,132]. (D) Antibacterial touchscreen for preventing contamination. The images were reprinted from Ippili et al [134]. (E) Digital biomarkers that reflect users’ moods, behaviors, and cognitions using text logs, browser history, human-computer interactions, and various sensors. The images were reprinted from Chen et al [19,135].

Display Materials
Health care delivery can also be achieved through materials used in manufacturing smartphones, such as window coatings for antireflection and display processes. An ecofriendly antibacterial coating with Zn-doped silicon oxide thin films can prevent infectious diseases caused by microbial contamination of touch events, as shown in Figure 7D [134]. In addition, it is possible to reduce the deformation of retinal cells by decreasing the blue light of the display through the material development of organic light-emitting or color filters [136].

Apps
Furthermore, health care sensing is possible through apps incorporating digital phenotypes and digital therapeutics [135,137-140]. A digital phenotype refers to a disease or health condition that is unintentionally reflected in patterns of use of digital devices. Mobile apps can collect human-smartphone interaction data to monitor smartphone usage and construct
long-term patterns and trend changes. As shown in Figure 7E
[135], analyzing a digital biomarker that reflects human effects,
moods, behaviors, and cognition can predict psychiatric
conditions, such as depression and smartphone addiction. In
addition, digital therapeutics delivered through games,
education, coaching, and counselling are based on cognitive
behavioral therapy and can treat insomnia, alcohol addiction,
drug addiction, panic disorder, and attention deficit hyperactivity
disorder. Additionally, it effectively improves physical diseases,
such as obesity and high blood glucose. Table 5 summarizes
the sensing methods and targets using smartphones.

<table>
<thead>
<tr>
<th>Type and sensing methods</th>
<th>Target(s) of detection</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Built-in sensors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMOSa</td>
<td>Atrial fibrillation</td>
<td>[18]</td>
</tr>
<tr>
<td>CMOS</td>
<td>Heart rate</td>
<td>[104,105]</td>
</tr>
<tr>
<td>CMOS</td>
<td>Diabetic retinopathy</td>
<td>[106]</td>
</tr>
<tr>
<td>CMOS</td>
<td>Skin cancer</td>
<td>[107-109]</td>
</tr>
<tr>
<td>CMOS + microphone</td>
<td>Heart rate, ( \text{SpO}_2 ), blood pressure</td>
<td>[112]</td>
</tr>
<tr>
<td>CMOS + microphone + speaker</td>
<td>Diet management</td>
<td>[113]</td>
</tr>
<tr>
<td>CMOS + strain gauge + display</td>
<td>Blood pressure</td>
<td>[110,114]</td>
</tr>
<tr>
<td>CMOS + temperature sensor</td>
<td>Temperature, heart rate</td>
<td>[115]</td>
</tr>
<tr>
<td>IMUf</td>
<td>Sleep monitoring</td>
<td>[116,117]</td>
</tr>
<tr>
<td>IMU</td>
<td>Gait analysis</td>
<td>[118]</td>
</tr>
<tr>
<td>Microphone</td>
<td>Spirometry</td>
<td>[111,119]</td>
</tr>
<tr>
<td>Microphone</td>
<td>Breathing sound analysis</td>
<td>[120]</td>
</tr>
<tr>
<td>Microphone</td>
<td>Sleep monitoring</td>
<td>[121]</td>
</tr>
<tr>
<td>Ultrasonic sensor</td>
<td>Biometric using fingerprint</td>
<td>[122]</td>
</tr>
<tr>
<td>Touch sensor</td>
<td>Heart rate</td>
<td>[123]</td>
</tr>
<tr>
<td>Touch sensor</td>
<td>Parkinson disease</td>
<td>[124]</td>
</tr>
<tr>
<td>Digitizer</td>
<td>Biometrics using signature</td>
<td>[125,126]</td>
</tr>
<tr>
<td><strong>Gadgets interlocked with smartphones</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optical platform</td>
<td>Pesticide evaluation in food</td>
<td>[127-129]</td>
</tr>
<tr>
<td>Smartphone CMOS + lens</td>
<td>Otoscopy</td>
<td>[130,131]</td>
</tr>
<tr>
<td>Microfluidic platform</td>
<td>Malaria infection</td>
<td>[132]</td>
</tr>
<tr>
<td>Patch electrode</td>
<td>ECGd</td>
<td>[133]</td>
</tr>
<tr>
<td><strong>Materials</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Window coating</td>
<td>Antibacterial</td>
<td>[134]</td>
</tr>
<tr>
<td>Light emitting</td>
<td>Blocking of blue light</td>
<td>[136]</td>
</tr>
<tr>
<td><strong>Apps</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digital phenotyping</td>
<td>Addiction, attention deficit hyperactivity disorder</td>
<td>[135,137,138]</td>
</tr>
<tr>
<td>Digital therapeutics</td>
<td>Mental health</td>
<td>[139,140]</td>
</tr>
</tbody>
</table>

aCMOS: complementary metal-oxide-semiconductor.
b\( \text{SpO}_2 \): oxygen saturation.
cIMU: inertial measurement unit.
dECG: electrocardiogram.

**Prospects for mHealth**

The industry of mHealth is expected to grow explosively in the future. In particular, the third generation of medicine and therapies that rely on novel solutions are emerging beyond the existing state of mHealth. Among them, bioelectronic medicine is a nonpharmacological treatment category that stimulates nerve functions with energy, such as electricity, light, and
ultrasonic waves. This approach uses an electronic device that controls metabolic function to maintain homeostasis by regulating hormones [141]. To date, electroceuticals have been used for obesity, asthma, sleep apnea, brain tumors, epilepsy, and Parkinson disease and have shown substantial and significant therapeutic effects [142-144]. It is also one of the most innovative fields in medicine because it has significant advantages when considering the development time and cost of existing drugs.

Using digital therapeutics, also referred to as “software as a medical device,” it is possible to manage and treat not only physical diseases but also psychiatric conditions, such as posttraumatic stress disorder and schizophrenia [145,146]. It is of great significance in terms of patient convenience that personal and sensitive mental health conditions can be diagnosed in real life, not in hospitals, through digital phenotypic analysis, such as smartphone usage patterns and uploaded social networking service (SNS) content.

From the point of view of the wearable form factor, since much of health care sensing is possible on the wrist, the smartwatch is currently playing a pivotal role in health care. The finger (as well as the wrist) is a body part to focus on as it can be used to assess health factors, such as the heart rate, oxygen saturation, ECG characteristics, blood pressure, blood sugar, biometric authentication, body temperature, and dietary monitoring. Therefore, it is expected that in the future, the ring type of device for health care will pair with the smartwatch as the 2 main pillars.

**Challenges for mHealth**

This viewpoint investigated comprehensive health care sensing technology using wearable electronics and smartphones. However, mHealth technology is less widely used than expected, unfortunately. Wearable devices are relatively more optimized for continuous and real-time health care sensing compared to smartphones [147,148]. However, the penetration rate compared to smartphones worldwide is sluggish [149-151]. A smartwatch, a representative wearable device, needs to be connected to a smartphone to operate, so users do not recognize the wearable device as an independent entity. Independent use is required to be fully positioned as a separate device. These devices lack effectiveness due to reduced user convenience because of their small screens, poor battery performance, low usage rate, clunky design, and high price. Wearable devices are recognized as a kind of subdevice rather than an essential and leading product because they do not have as much impact as smartphones. Therefore, in the case of wearable devices, innovative solutions are required to make them universal necessities for human beings, such as smartphones.

However, in the case of smartphones, the penetration rate is high worldwide, including low- and middle-income countries [152]. In the case of current smartphones, the fundamental value in terms of user experience as well as utility is high. However, it is not such a great solution from the perspective of health care. It is challenging to conduct biosignal sensing using a smartphone while being in close contact with human skin all day long, so it is challenging to implement continuous real-time big data–based predictive and preventive medical care using smartphones from the health care perspective. Smartphones desperately require a breakthrough that can allow them to monitor health in real time continuously, 24 hours a day, through a form more closely adherent to the skin, while maintaining the current phone function.

**Breakthroughs for mHealth**

The display is a crucial component of a health care system. In other words, smartphones and wearable devices, as central axes of the mHealth system, are inseparable from their displays. In addition, displays and sensors in mobile devices are closely related. To improve the convenience of user interaction, the proportion of the active area of mobile displays is increasing. However, the increase in the active area has a limitation that reduces the sensing performance, including sensitivity. To overcome this, the upper part of the sensor covers the display by lowering the resolution of the display to prevent the deterioration of the sensing transmittance. A typical example is under-panel camera (UPC) technology that covers the camera with the display by reducing the display resolution on the top of the CMOS image sensor to increase light transmittance.

**Sensor-Integrated Display Solution**

However, the ultimate and ideal method is a sensor-integrated display solution. A sensor-integrated display has many advantages from a health care sensing point of view. This is because (1) many mHealth sensors use an optical approach, (2) it is relatively easy to manufacture large-area sensors, and (3) the application of a new form factor display can lead to an increase in the body contact area.

First, the majority of mHealth sensing approaches are optical methods. Various health care parameters, such as the heart rate, oxygen saturation, blood pressure, blood sugar, body temperature, environmental monitoring, and ECH characteristics, can be measured optically. A display is an optical system that already has the means to transmit light. Therefore, a sensor-integrated display could be an optimized health care solution. To implement health care devices using optical systems, in addition to optical transmitters, receiver systems must also be equipped. For advanced performance, the light-emitting wavelength band needs to be expanded and supplemented, including infrared as well as visible light, through the development of materials for the light-emitting layer.

Second, since the sensing area and detection performance are proportional, health care ability can be improved through a sensor embedded in a wide display area. It enables health care sensing in a large area over the entire display area when the built-in optical system is applied, considering design rules. In addition, it is more advantageous for wearability because of a reduction in volume due to the implementation of microlevel thickness because of the sensor-integrated display. Additionally, compared to the number of photomasks needed to manufacture a conventional display, the number of additional photomasks required to implement a display health care system with built-in sensors is far less. It can contribute to popularization due to the low manufacturing price according to the integral type.
Ultrathin, low-cost health care devices with relatively simple processes have significant benefits over conventional, bulky, and expensive wearable computers.

Finally, the new form factor device, such as a stretchable sensor-integrated display, increases the area of contact with the body and improves detection capability through health care sensing in close contact with the skin. Flexible panels with user convenience could be applied to the human skin, considering ergonomic factors [153-159]. The flexibility of not only the active matrix backplane and core of the panel but also the touch sensor, fingerprint sensor, and pressure sensor must be ensured, as shown in Figure 8A [153]. In a complete sensor-integrated display, the flexibility of the backplane allows the sensor part to gain flexibility naturally.

Figure 8. New form factor display and principle. (A) Wearable display with flexible and ultrathin active matrix backplane, touch screen panel, and fingerprint sensor components. The image was reprinted from Park et al [19,153]. (B) A flexible active matrix organic light-emitting diode (AMOLED) with large-area MoS2-based backplane for human skin display. The images were reprinted from Choi et al [154]. (C) A graph of compressive and tensile strength as the thickness increases in a single-layer structure (solid blue line) and laminated structure (dotted red line).

Furthermore, Figure 8B [154] shows a wearable full-color organic light-emitting diode (OLED) display using a 2D material–based backplane transistor suitable for complex skin shapes. The 18×18 thin-film transistor array was fabricated on ultrathin MoS2 film and then transferred to Al2O3 (30 nm)/polyethene terephthalate (6 μm), providing mechanical flexibility beyond conventional OLED technology.
New Form Factor Display

The left picture of Figure 8C simulates a multilayered display, and when this display is bent, tensile strength is applied at the top and compressive strength is applied at the bottom. Assuming that it is formed with only a single layer of the same thickness rather than a laminated structure, extreme tensile and compressive forces occur on the upper and lower surfaces, resulting in cracks in the display, as shown by the solid blue line in the right graph. However, in the stacked structure, a pressure-sensitive adhesive (PSA) between the display layers continues to create new neutral planes, as shown by the dotted red line. In response, the magnitude of the tension and compression force at the top and bottom surfaces does not increase, even if the thickness of the display increases. In other words, using the PSA, it is possible to implement a flexible display without cracks.

No part of the human body is flat. When the health care system and the skin conformally adhere, sensing performance improves. Display technology based on PSA with the harmony of creep and recovery characteristics induces form changes in wearable devices and smartphones. A new form factor with flexibility based on PSA technology that creates a new neutral plane will facilitate a critical conversion of the mHealth system.

Standard of the Medical Paradigm in the Postpandemic Era

A new form factor display for health care with flexibility and display convergence technology using an optical method attaches a large-area health care system to the human skin conformally and continuously detects health care factors in real time, thereby providing a framework for collecting big data. As a result, the existing smartphone becomes a wearable device attached to the body, and the existing wearable device is equipped with smartphone functions suitable for user convenience. Namely, convergence health care technology with the sensor-integrated and new form factor display is an indispensable element that enables the smartphonization of a wearable device and the wearable deviceization of the smartphone. Of course, health care systems with new form factors and sensor-integrated displays do not solve all mHealth problems. In other words, advances in big data AI software analysis and medical security should go hand in hand with the smartphonization of wearable devices and the wearable deviceization of smartphones. Furthermore, it will be necessary to supplement the medical system policy so that these benefits do not become the exclusive property of the upper class of the economy and so that people from lower social classes can also benefit. Advanced and popularized mHealth system technology could ensure universal health coverage so that everyone can use essential, high-quality medical services without discrimination. In other words, the authentic democratization of health care could become a reality, and a standard for a future health care paradigm in the post-pandemic era could arise.

Conclusion

Personalized platforms, such as wearable devices and smartphones, can be applied to AI-based disease prediction, prevention, and treatment. This viewpoint researched the latest technology trends in mHealth regarding form factors and detection targets according to body attachment location and type. In particular, the sensor convergence technology of the new form factor display provides a framework to analyze health factors in real time by conformally adhering a large-area system to the skin. Innovation in form factors in sensor-integrated displays and convergence health care solutions enable the smartphonization of wearable devices and the wearable deviceization of smartphones. In addition, the strategy for the smartphonization of wearable devices and the wearable deviceization of smartphones can accelerate the development of mHealth, realizing the democratization of medical care so that anyone can use essential services of high quality. Furthermore, it is expected to create a new milestone for the medical paradigm shift in the postpandemic era.


19. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2024-01-11]


66. Phillips C, Liaqat D, Gabel M, de Lara E. Wristo2: reliable peripheral oxygen saturation readings from wrist-worn pulse oximeters. Presented at: 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops); Mar 22-26, 2021; Kassel, Germany. [doi: 10.1109/PerComWorkshops51409.2021.9430986]


Abbreviations

AI: artificial intelligence
CMOS: complementary metal-oxide-semiconductor
ECG: electrocardiogram
EEG: electroencephalogram
EGG: electroglottogram
EMG: electromyogram
IMU: inertial measurement unit
mHealth: mobile health
OLED: organic light-emitting diode
PPG: photoplethysmogram
PSA: pressure-sensitive adhesive

© Wonki Hong. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 22.1.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.

https://mhealth.jmir.org/2024/1/e48803 JMIR Mhealth Uhealth 2024 | vol. 12 | e48803 | p.40
SOMAScience: A Novel Platform for Multidimensional, Longitudinal Pain Assessment

Chloe Zimmerman Gunsilius, Joseph Heffner, Sienna Bruinsma, Madison Corinha, Maria Cortinez, Hadley Dalton, Ellen Duong, Joshua Lu, Aisulu Omar, Lucy Long Whittington Owen, Bradford Nazario Roarr, Kevin Tang, Frederike H Petzschner

Abstract

Chronic pain is one of the most significant health issues in the United States, affecting more than 20% of the population. Despite its contribution to the increasing health crisis, reliable predictors of disease development, progression, or treatment outcomes are lacking. Self-report remains the most effective way to assess pain, but measures are often acquired in sparse settings over short time windows, limiting their predictive ability. In this paper, we present a new mobile health platform called SOMAScience. SOMAScience serves as an easy-to-use research tool for scientists and clinicians, enabling the collection of large-scale pain datasets in single- and multicenter studies by facilitating the acquisition, transfer, and analysis of longitudinal, multidimensional, self-report pain data. Data acquisition for SOMAScience is done through a user-friendly smartphone app, SOMA, that uses experience sampling methodology to capture momentary and daily assessments of pain intensity, unpleasantness, interference, location, mood, activities, and predictions about the next day that provide personal insights into daily pain dynamics. The visualization of data and its trends over time is meant to empower individual users’ self-management of their pain. This paper outlines the scientific, clinical, technological, and user considerations involved in the development of SOMAScience and how it can be used in clinical studies or for pain self-management purposes. Our goal is for SOMAScience to provide a much-needed platform for individual users to gain insight into the multidimensional features of their pain while lowering the barrier for researchers and clinicians to obtain the type of pain data that will ultimately lead to improved prevention, diagnosis, and treatment of chronic pain.

(JMIR Mhealth Uhealth 2024;12:e47177) doi:10.2196/47177
KEYWORDS
acute pain; acute-chronic pain transition; chronic pain; clinical outcome measurement; digital health; ecological momentary assessment; EMA; ESM; experience sampling methodology; mHealth; mobile health; pain management; pain self-management; patient reported outcomes; smartphone app

Introduction

More of us are in chronic pain than you might think—20% of adults in the United States reported pain on most or several days in 2019 [1]. This presents a substantial burden on society, costing up to US $635 billion annually [2]. In 2016, chronic back and neck pain alone accounted for the highest amount of US health care spending across 154 conditions, including diabetes and heart disease [3]. Moreover, pain is the leading cause of health care use across all illnesses [4-6]. Against this backdrop, it has never been more important to develop accurate pain symptom assessment and prediction methods to help patients, caregivers, and other stakeholders make informed decisions about treatment and care.

Accurately measuring pain is crucial for predicting an individual’s pain trajectory [7]. Methods to identify objective biomarkers of pain intensity [8-10] are still in their infancy and have yet to be proven effective in predicting future self-reported pain [11]. To date, the most common way to assess if someone is in pain is to simply ask them [12]. Typically, this is done using an 11-point pain intensity scale, where individuals are asked to rate their pain from 0 (no pain) to 10 (worst pain imaginable) [12-14]. Since the 1980s, this simple pain intensity scale has played a significant role in the clinical assessment of pain by enabling defined targets for pain management and the dosing of pain-relieving medications [15,16]. If a person consistently reports pain intensity as more than 3 out of 10 for more than 3 months, the scale becomes part of the diagnostic criteria for chronic pain [17]. For patients, self-reporting their pain on the scale can validate and quantify their pain experience, leading to improved shared decision-making and enhanced communication with health care providers [18].

Despite its ease of use, the pain intensity scale has not led to significant advancements in pain management or patient satisfaction [18-23]. When measured in medical settings, people tend to over- or underreport their pain intensity depending on difficulties with recall, expected treatments, care standards, or other subjective factors such as mood [24,25]. Additionally, regulatory approval for pharmaceutical companies to promote “titration to effect” practices, whereby physicians were encouraged to increase opioid doses to achieve continued reductions on the pain intensity scale, contributed to opioid overprescribing that fueled the opioid epidemic [23,25]. These limitations highlight the need for more nuanced self-report measures of pain.

In recent years, clinical approaches to pain have sought to better assess the multidimensional experience of pain from a biopsychosocial perspective [12,26]. Multidisciplinary and individualized assessment and treatment of pain with both pharmacologic and nonpharmacologic interventions is now considered the ideal way to treat both acute and chronic pain [27]. To overcome the limitations of the unidimensional pain intensity scale, expert panelists at the Food and Drug Administration (FDA) and National Institutes of Health (NIH) have established “core outcome sets” of multidimensional questionnaires for use in research studies [13,28,29]. Dimensions assessed include pain, unpleasantness, interference, and impacts on mood and activity [25,30]. However, completing multiple questionnaires is time-consuming, and therefore not ideal for daily longitudinal studies. As a result, a large barrier remains in the ability of researchers to collect comprehensive, multidisciplinary pain data sets [31]. It is therefore imperative in pain research to implement a reliable method for multidimensional pain measurements that accurately captures the most important dimensions of pain symptoms and treatments as they evolve over time in the context of people’s daily lives.

A comprehensive approach to pain assessment necessitates the acquisition of both deep and wide pain data. Deep data involve in-depth evaluations of the multidimensional aspects of pain within individuals over extended periods, while wide data refer to data sets that encompass a large number of individuals across different demographic factors like age, geographic location, race, ethnicity, and socioeconomic status, as well as across different pain diagnoses (e.g., arthritis, fibromyalgia, and postsurgical pain). The acquisition of deep data enables a deeper understanding of the mechanisms that trigger and sustain pain in individuals, while wide data provide the foundation for generalizing findings and developing biomarkers for pain persistence or recovery. Smartphone apps can provide large-scale platforms for data collection while also helping users track their daily symptom experience [32]. Such digital tools provide a promising solution for acquiring deep and wide data sets that enable new behavioral and scientific insights into the dynamics and evolution of pain.

Here, we introduce a novel mobile health (mHealth) platform for longitudinal pain assessment, called SOMA. This platform has been conceptualized to comprehensively capture multiple facets of pain through its smartphone app, SOMA. Unlike conventional pain assessment tools that focus primarily on daily pain intensity, SOMA offers an enriched multidimensional pain assessment. This includes factors like daily pain intensity, unpleasantness, interference, mood, expectations, and activities, based on experience sampling methodologies (ESMs) [33]. Our choice of measures is in accordance with the current Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) recommendations for pain assessment [28,34]. In addition, the design and validation of the SOMA app have been executed following the guidelines set forth by the American Psychiatric Association (APA) [35,36].

SOMA was designed with users in mind, offering a free and user-friendly tracking feature that facilitates real-time tracking of pain, medications, and treatment regimens. A “Trends” section distills the multidimensional data to visualize pain trends.
This feature aids users in self-managing their pain, recognizing patterns, and discerning between effective pain management and areas needing improvement. With these insights, users can communicate more efficiently with their health care providers, positioning SOMA as a valuable “companion” tool to be used alongside standard medical advice.

SOMAScience has been engineered to provide researchers with the ability to carry out independent single- and multicenter studies. Its pain assessments can be used as end points in research studies and within clinical practice. To support expansive research initiatives, the SOMAScience platform ensures seamless transfer of app-sourced data to researchers and affiliated institutions and offers open-source code to streamline data preprocessing.

Contemporary pain research standards underscore the significance of multidimensional pain assessments through established batteries of pain questionnaires [13,29]. However, we identified a scarcity of available tools that effectively capture such assessments in a format meaningful for researchers aiming to develop better measures of pain outcomes or symptom burden over time. While there are smartphone apps, like Manage My Pain and Pain Scale-Pain Tracker App, they primarily cater to users as self-management tools [37,38]. The foundational architecture of such apps and the nature of the data they procure do not typically align with the specifications researchers need for conducting or analyzing comprehensive, large-scale studies that meaningfully contribute to pain research. A further limitation is the proprietary nature of many of these apps, which can impede data quality, accessibility, and transparency.

To our knowledge, no other tools currently exist that are specifically designed to capture multidimensional and longitudinal pain metrics in compliance with the recommended standards and are simultaneously tailored for extensive open-source academic research. This distinctiveness sets SOMAScience apart as an unparalleled mHealth platform.

In this paper, we discuss the choices and technological considerations for the development of SOMAScience as well as the scientific rationale behind the selection of measures. Our aim is to outline how SOMAScience can be used by scientists and institutions to acquire large-scale, longitudinal, multidimensional pain data in single and multicenter studies in order to gain new insights into pain that will benefit patients.

**Section 1: SOMAScience Platform**

**Overview**

SOMAScience aims to address the current shortage of platforms for acquiring both deep (multidimensional and longitudinal) and wide (cross-spectrum and large-scale) pain data to create novel insights into the dynamics and evolution of acute and chronic pain. The term SOMA stems from the Greek word σώμα (meaning body or entire person), signifying that it takes a holistic approach to pain.

SOMAScience represents the combination of 3 branches: the SOMA smartphone app used for data acquisition, an application programming interface for data transfer, and open-source analysis code distributed through GitHub [39] (Figure 1).

**Figure 1.** Schematic overview of the SOMAScience platform. Data will be acquired using the SOMA app, a user-friendly smartphone app available on Google Play and the Apple App Store. Encrypted data are sent to our application programming interface (currently located at Brown University) and shared in a study-specific manner with individual researchers and institutions. To facilitate data analysis, we created a GitHub repository where researchers can download, modify, or even create new versions of our template scripts for data preprocessing and certain analysis techniques through GitHub. ESM: experience sampling methodology; IRB: institutional review board.

---

**To request to run a study through the platform, researchers need to submit a research inquiry detailing the study purpose on the SOMA website [40]. Data for SOMAScience are acquired through the SOMA app, which is freely available on Google Play and the Apple App Store.**
Play [41] and the Apple App Store [42] and can be found by searching for “SOMA Pain Manager.” Anyone is able to download and use the app, regardless of whether they are participating in a research study. For associating people’s app data with a specific study, the researcher will be assigned a unique study ID (1 per study) and a list of individual app IDs (1 per expected participant), which need to be sent to the study participants. Participants can then install and register on the SOMA app and enroll in a specific study using the study ID and unique app ID (instruction videos on [43]; Figure 2). This use of study and app IDs allows the assignment of individual participants’ data to single and multicenter studies.

Figure 2. Schematic overview of study enrollment on the SOMA app. Study participants can download and register on the SOMA app directly using the links to Google Play or the Apple App Store or by searching for “SOMA Pain Manager.” Inside the main menu of the app, they can sign into a specific research study using a pre-sent study ID and App ID. They are shown a copy of their institutional review board–approved study-specific informed consent form for their records. Encrypted research data on SOMA will be sent to our database and transferred to the researchers of each individual study.

Upon enrollment, participants will be sent regular reminders through the app to fill out short pain surveys (details about the ESM and data content are in the following sections). At the time of publication, assessments are restricted to the features listed below. Future releases may offer the option to request additional features and questions. After each assessment, encrypted data are transferred to our application programming interface, stored on an actively managed secure database (see “Maximizing Privacy and Security” in Multimedia Appendix 1), and then shared with researchers from individual studies (Figures 1 and 2).

Implementation of APA Guidelines in SOMA

The APA’s app evaluation model stands as a notable benchmark for evaluating the suitability of health-related smartphone apps intended for patient populations [35]. It delineates 5 pivotal criteria to assess apps: accessibility, privacy and security, clinical foundation, engagement style, and therapeutic goal [44]. To keep pace with the rapidly advancing field of health apps, the system is regularly updated and refined [45]. Stemming from this APA model is a comprehensive database [46], which facilitates app evaluations across the 5 core domains [47]. Such initiatives are crucial in establishing public-facing, user-friendly standards for health apps and ensuring the development of safe and efficacious apps that benefit users [48-50].

In the development of SOMA Science, we have deeply integrated the cardinal principles of the APA model. Recognizing the emphasis that this model (and other akin evaluation frameworks) places on robust privacy, security, usability, and clinical foundations, we meticulously factored in specific technological elements during the app’s creation [50]. For readers interested in the technological nuances and our dedicated approaches to privacy and security aligned with the APA guidelines, we direct you to Multimedia Appendix 1 [51-60].

Development of SOMA Science

A multidisciplinary team at the Psychiatry, Embodiment, and Computation Lab at Brown University, comprising academic neuroscientists, psychologists, pain physicians, software engineers, and designers, collaboratively conceptualized and initialized the early design and features of the SOMA app. The SOMAScience platform is supported by the Brainstorm program at the Carney Institute for Brain Science at Brown University, a new program to accelerate the translation of computational brain science into real-world applications that benefit patients, the scientific community, and society as a whole.

Incorporation of Patient and User Feedback

Any app meant for long-term use must provide a simple, user-friendly interface tailored to its target audience. With this in mind, we actively sought initial feedback from individuals experiencing chronic pain to shape our app’s delivery structure. Through comprehensive one-on-one Zoom (Zoom Video Communications, Inc) discussions with a select patient group (n=4), representing a diverse pain spectrum including migraines, postsurgical musculoskeletal pain, cancer pain, and inflammatory bowel disease, we gleaned insights into our early app prototypes. These discussions involved a detailed
walkthrough of each interface, where patients aired their thoughts in a guided interview.

Drawing from this feedback, we refined the “Check-In” interface’s design and flow. A common sentiment among initial users was the importance of daily tracking of pain and mood fluctuations in tandem with daily activities. To accommodate this, we introduced the “Trends” screen, a visual tool designed not only to foster self-management and a deeper understanding of pain dynamics but also to facilitate effective communication with health care providers. To further enhance the app’s utility, we incorporated screens to monitor various treatment modalities and transitioned the mood and emotion tracking interfaces to use intuitive visual analog scales (VASs) in lieu of a 2D rating system.

Responding to the patient’s desire for a more personalized experience, we introduced an interactive chatbot during the onboarding phase. This chatbot briefly engages users, gathering foundational demographic details and an introductory snapshot of their pain experiences. Existing studies vouch for the efficacy of chatbots in extending support to people with chronic pain [61], making this an evolving component with forthcoming features in SOMA that focus on pain interventions.

Our iterative refinement strategy incorporated a beta-testing phase. Initially, 30 internal testers actively engaged with the app daily across a gamut of devices and operating platforms, enabling us to identify and correct technical glitches and enhance the user experience, especially regarding the “Trends” data visualization. To expand our feedback, SOMA was then shared with a larger patient interest group (over 250 people with chronic pain), leading to critical refinements and the inclusion of user-suggested enhancements. This ongoing feedback mechanism ensures the continuous improvement and evolution of SOMA.

Section 2: Pain Data Acquisition Through SOMA

Deep Data Acquisition Using ESMs

Overview

The SOMA app uses ESMs to gather multidimensional and longitudinal pain data for SOMA. ESMs, also known as ecological momentary assessments, provide real-time, self-report data about individuals’ thoughts, feelings, and experiences (eg, “How do you feel right now?”) in the context of people’s daily lives [33]. Previous ESM studies on pain have shown high completion rates (>85%) and demonstrated the feasibility of using these daily self-reports for pain [62], in line with findings about the high completion rates of mobile-delivered ESM studies in general [63].

ESMs have several benefits over traditional self-report measures. First, they offer real-time data that are less prone to recall bias, allowing for the capture of critical experiences that might be missed by retrospective long-term self-report measures [64-67]. Second, ESMs can capture contextual information about an individual’s thoughts and experiences, such as knowing what activities a person engaged in when they experienced pain [68]. Thus, they provide insight into the longitudinal dynamics of multidimensional aspects of pain in people’s natural ecological environments across time and context [69,70]. This enables the generation of rich data sets that could be used to identify candidate behavioral “biomarkers” or “assays” that predict transitions in disease states based on self-report alone [68,71]. For example, there is preliminary evidence that longitudinal measures of pain can predict acute pain state transitions after surgery [72] and identify treatment response time courses in patients with chronic pain [73].

Momentary, Situational, Retrospective, and Prospective Assessments in SOMA

One limitation of existing ESM studies is that they typically solicit several short, momentary reports throughout the day [74]. While this approach reduces bias in pain reports resulting from memory recall or pain beliefs, it may still miss important short-term pain dynamics, such as flare-ups, and fail to assess the role of expectations in the development and treatment of pain [75]. To address this limitation, SOMA uses a multifaceted approach, which includes 4 daily assessment types on the SOMA app: momentary assessments (called random check-ins), voluntary self-initiated entries (called quick check-ins), and both retrospective and prospective assessments or coverage assessments [75] (which are both part of an evening routine at the end of the day; Figure 3).
ESM assessment types

- **Situational**: Quick Check
- **Momentary**: Random Check
- **Retrospective**: Evening Routine
- **Prospective**: Morning Routine

Random and quick check-ins capture various aspects of mood, activities, pain, and pain location and can be completed in less than 30 seconds. Quick check-ins can be performed at any time, for example, during or shortly after a flare-up. Random check-ins reflect classical ESM assessments and only occur during randomly selected moments within a specific time window (e.g., 3 checks per day between 8 AM and 6 PM). Users receive notifications on their phone when the random check-in is available and have the option to snooze the notification for a predefined time window (e.g., a maximum of 60 minutes).

The evening routine assesses recall of pain, mood, activities, and any pain-related treatments over the past day (retrospective), as well as predictions of pain, mood, and activities for the next day (prospective). This routine is available during a prolonged, preset, but fixed time window at night (e.g., 6 PM-11 PM) to promote habit formation that increases the likelihood of long-term app use. It takes less than 3 minutes to complete.

Longitudinal Assessments with SOMA Science

Pain is inherently dynamic, fluctuating not just daily but hourly and even on a minute-to-minute scale, even without changes in physiological markers [18,47,62]. Traditional methods, which measure pain intensity sporadically during clinical trials or medical visits, might not capture a patient’s holistic pain experience due to their limited assessment windows.

While some studies aim for more granular pain assessments, they often focus on brief periods. A systematic review found the median duration for ESM studies in pain to be 14 days and a mere 7 days for general mHealth ESM studies [40,61]. Such short durations can overlook pivotal phases, like the transition from acute to chronic pain over 3-6 months. Consequently, the role of self-reported pain dynamics in acute pain recovery or its evolution to chronic pain remains ambiguous.

The SOMA app is designed to bridge this gap. Its check-ins are concise, using straightforward design principles for ease of use. The chosen metrics cater to diverse pain experiences, facilitating collaboration and data sharing among researchers. Moreover, the app’s “Trends” feature empowers users to track their pain, treatments, activities, and mood over diverse timeframes (i.e., weekly, monthly, and annually). This aids in providing users with a deeper understanding of their pain journey, ultimately supporting more effective self-management.

Wide Data Acquisition Using Smartphones

While large data sets on repeated multidimensional pain ratings beyond intensity alone are still few and far between, smartphones offer a unique opportunity to expand data acquisition beyond classical experimental settings [74]. Smartphone access has increased tremendously in the past decade (84% of US households reported owning at least 1 smartphone) [76]. Data acquired remotely through smartphone apps facilitate large-scale, real-world studies without the constraints of traditional laboratory studies. The results of such pragmatic studies are more generalizable than highly selective traditional randomized controlled trials [77-79]. SOMA Science was built to allow remote monitoring of longitudinal symptoms and treatments to maximize high-quality data in large-scale pragmatic studies. To further facilitate a much wider array of user input, the SOMA app is compatible with both Android and iOS devices, meaning anyone in the United States with a smartphone can use it.

Smartphone-based pain assessments offer a solution to the limited geographic, economic, and cultural diversity in traditional pain studies. Smartphones are pervasive, even in low-resource [80] or rural areas [72,81], where almost half of the world’s smartphone owners live [82]. They are also widely used by older adults [83], who are often left out of laboratory-based pain studies. Additionally, there is a need to consider how pain and its treatment vary across racial, ethnic, and cultural backgrounds for comprehensive care. Even within a specific culture, there are important differences in how pain is experienced and treated across different racial and ethnic backgrounds that need to be accounted for to deliver the best pain care [84-86]. Upcoming translations of the app into languages like Spanish and German, in collaboration with experts familiar with the culture, will further diversify data and
insights. Translation into other languages will follow, and collaboration to translate the app is welcomed.

**Section 3: Data Content**

**Overview**

With the rising number of health-focused smartphone apps, there is also a growing need for transparency in the selection of measures for the app. Here we briefly detail the process of selection for each measure in SOMA Science as assessed through the SOMA app, its basis in the scientific and clinical literature, and what gaps it was chosen to address. The goal is to provide transparency in the design and selection process to facilitate the development of research studies using SOMA Science. Specific measures may be refined over time with user feedback and as scientific studies using SOMA Science identify areas for improvement.

**Measuring Pain Intensity, Unpleasantness, and Interference**

Multidimensional pain assessment is a crucial aspect of clinical practice as it helps determine the effectiveness of treatment and recovery. The SOMA app assesses the pain intensity scale in addition to pain interference and pain unpleasantness to provide a more holistic measure of daily pain symptom load [87]. Participants are asked to rate each pain question on a sliding VAS, providing a simple and comparable assessment of daily pain experience (Figure 4). This operationalizes a practice advocated in the International Statistical Classification of Diseases and Related Health Problems for providers to use a composite rating of pain intensity, distress, and interference of pain to determine overall burden [88]. The composite pain score provides a multidimensional solution that balances the limitations of the unidimensional pain scale while still being easily quantifiable. The ability to analyze the 3 measures individually and as a composite score helps identify meaningful individual variability, enabling specific predictions between individuals. For example, pain unpleasantness could be more predictive of future pain in individual A, while pain interference may be more important for individual B’s long-term outcomes.

![Figure 4. Upper panel: example screens on the SOMA app for check-ins and trends. Lower panel: example screens on the SOMA app from Quick Check, Random Check, and Evening Routine. From left to right: assessing pain on a visual analog scale (VAS), indicating pain location, assessing mood on a VAS, emotion selection, and activity selection.](image)

Importantly, SOMA’s 3 pain questions were chosen because they are directly comparable to results from established pain questionnaires, such as the Brief Pain Inventory [89] or the McGill Pain Inventory [90]. They also satisfy the standards set by major scientific and regulatory bodies, such as the IMMPACT recommendations, the NIH Helping to End Addiction Long-term initiative, and the FDA guidelines for assessing multidimensional components of pain [13,28,29,87]. In this way, SOMA’s multidimensional pain assessment of intensity, unpleasantness, and interference can provide important supplemental measures that are directly comparable to established clinical benchmarks and standards of care. This is...
critical for researchers looking to establish and validate novel pain biomarkers or end points.

**Measuring Pain Locations**

Pain localization is an important aspect of pain assessment. Conventional methods of measuring pain location in medical appointments and research studies involve having individuals indicate it on a body map, such as the Brief Pain Inventory [89], the McGill Pain Questionnaire [91], or the Michigan Body Map [92]. This approach can pinpoint differences in peripheral and central pain pathology based on the localization and stability of pain representation over time. For instance, nociceptive or inflammatory pain is usually precisely localized somatically and does not change much over time, while neuropathic or chronic primary pain is often experienced in multiple bodily locations, radiates, or changes over time [93].

More recent methods of digital quantification, like the ones used on the SOMA app, have established the reliability and validity of body maps for pain assessments [94,95]. Interactive body maps delivered through digital or tablet apps are more effective than traditional paper or laptop assessments [92,96]. Yet a review of smartphone apps that use the body map for tracking pain found that few actually quantified the location ratings or provided any summary feedback [97]. The SOMA app’s interactive body map offers 46 different discrete location options on the front and back of the body that participants indicate in every daily check-in (Figure 5 and Multimedia Appendix 1). The use of discrete points ensures uniformity across devices and accounts for differences in participants’ finger size or dexterity. The “Trends” section of the SOMA app displays the body map with the percentage of times a location has been selected, enabling users to visualize the frequency of pain at a given location. For participants who experience nonspecific, difficult-to-localize, or widespread pain, such as fibromyalgia, there is an additional option to indicate “My pain is everywhere” on the body map.

**Figure 5.** Pain map included on the SOMA app covering 46 discrete pain locations.

---

**Measuring Interventions**

The treatment of pain has been incredibly difficult to get right. The newest clinical guidelines advocate the use of multimodal, multidisciplinary approaches [27,98,99]. Such approaches emphasize a combination of pain treatments that include medications, restorative therapies (eg, physical therapy), interventional procedures (eg, epidural injections), behavioral interventions (eg, cognitive behavioral interventions), and complementary and integrative medicine (eg, acupuncture). Combinations of these therapies have been associated with the best long-term pain outcomes [100,101] and satisfy a biopsychosocial approach to pain [102].

It can be challenging for individuals and providers to determine which treatments are most effective for them, as the effects of many treatments for pain may not become apparent for weeks or even months (eg, cognitive and physical interventions, certain medications, and surgery) [103,104]. In determining how to measure treatments through the SOMA app, we followed the recommendations of the 2019 Department of Health and Human Services Pain Strategy [27].

In the case of medications, many different pathophysiologic mechanisms are targeted with different classes of medications. The use of different medications often changes over time, so we designed the medication screen on the SOMA app to be able to capture such changes. We, therefore, included 20 options across the main classes of pain medications for both acute and chronic pain on the SOMA app, detailed further in Multimedia Appendix 1.

A second treatment screen includes the recommended nonpharmacologic approaches to pain. Combinations of these treatments are often used by a single person over time to target different pain mechanisms [105]. SOMAScience currently provides the ability for people to track up to 20 different nonpharmacologic therapies across these 4 major classes, detailed in Multimedia Appendix 1. SOMAScience’s broad treatment tracking capabilities therefore facilitate the type of
wide data needed to understand differences in treatment use across users.

**Measuring Emotions**

Emotion and pain are fundamentally related. Definitions of pain acknowledge that pain is partially an emotional experience [106,107], yet few researchers would reduce pain down to just another emotional state such as sadness or happiness. The complexity of the pain-emotion relationship is highlighted by reviews of the neural circuits of each construct, showing both shared and functionally dissociable brain regions [108,109]. Unfortunately, the theoretical and empirical understanding of how emotion and pain are connected is limited, as much previous research only focuses on cross-sectional correlations between the 2 constructs [110,111].

Despite the renaissance of emotion research since the 1960s [112], emotion researchers continue to disagree on what constitutes an appropriate emotion measure [113]. The discrete emotion perspectives suggest that specific emotions such as anger, fear, happiness, sadness, disgust, and surprise are special kinds of biologically distinct responses associated with unique behavioral, physiological, and experiential correlates [114,115]. The dimensional perspectives consider emotions to be organized along a set of common dimensions such as valence (unpleasantness) and arousal (intensity) that are combined with cognitive and cultural knowledge to form emotional states.

When designing an emotion list for the SOMA app, we wanted a set of emotions that are commonly experienced, had a balance of positive and negative emotions, and were relevant to pain experiences (eg, worry). We started by selecting emotions experienced more than 5% of the time from a large-scale ecological momentary assessment study [116], removing rare emotions such as contempt or awe. Next, we balanced the remaining emotion set by using positive and negative emotions from the clinically relevant scale (Positive and Negative Affect Schedule) [117]. Lastly, we allowed participants to indicate no emotion (neutral) or emotion not listed (other).

A complementary method from a dimensional perspective is to assess general mood through a single-dimensional VAS of valence ranging from very unpleasant to very pleasant. Although there are other dimensional scales that could also be assessed, such as arousal [118,119] or goal congruence [120], valence is known to capture the majority of variance between emotion states [121], as self-reports of emotions tend to be highly correlated within a positive or negative valence [122]. Accordingly, on the SOMA app, we ask users to evaluate their current, past, or future mood on this VAS, ranging from unpleasant (0) to very pleasant (100). One major benefit of this measure is that we can quantify daily emotional experiences even if the participant does not self-report any discrete emotions from our finite list, reducing the need to interpolate or remove missing data from our analysis of the dynamics of emotion and pain (Figure 4).

**Measuring Activities**

People’s experiences of both pain and emotion are intrinsically tied to the activities they are engaging in each day. There have been many large-scale studies investigating the relationship between emotions and behavior, showing that physical exercise is meaningfully associated with reduced mental health difficulties [123] and that people’s choice of activities is motivated by minimizing negative affect and maximizing positive affect [124,125]. In particular, people seem to engage in mood-elevating activities (eg, socializing) when they are feeling down and mood-depressing activities (eg, work and chores) when they are feeling up. At the same time, being in pain affects both mood and daily activities.

Most people intuitively reduce their activities when they experience new-onset pain. Acute pain generally functions as an alarm bell in the brain to signal tissue damage, with the urge to rest considered a protective mechanism to prevent further injury and promote healing [126]. However, outside of the initial acute phase, a lack of activity can hinder long-term recovery and may signal underlying changes in affective and motivational brain circuits that have been causally linked to the transition from subacute to chronic low back pain [126,127]. The synergistic impact of activity engagement on mood, pain, and physical function is why activity engagement is promoted by pain self-management and rehabilitation programs alike [128]. While many people assume that patients with chronic pain move less than those not in pain, objective actigraphy data do not differ between patients with chronic pain and no-pain controls [129,130]. What has been less studied is the types and range of daily activities and how they change between acute and chronic pain stages. While people may recognize changes in their daily activities as a result of pain (eg, no longer walking as much), they may not have insight into the relationship between pain, mood, emotion, and activities (ie, which activities increase or decrease pain in the short vs long term). It remains unclear whether certain patterns of activity engagement at different stages of pain experience are important for long-term outcomes.

For this reason, we added an activities-tracking feature alongside mood and pain tracking. The activities screen on the SOMA app contains 20 activities that are known to have a dynamic relationship with mood based on large-scale, longitudinal data sets [125] or to be common among patients with pain (eg, medical visits). Specific activities selected are detailed in Multimedia Appendix 1. In addition to having the person report either momentary or daily activities, we also have participants reflect on how each activity they completed made them feel that day and how much it affected their pain. Taken together, these measurements provide a full picture of participants’ daily activities and help understand the potential bidirectional relationships between emotion, pain, and behavior. Users are also able to visualize their ratings of how much a given activity affected their pain and mood over time in the dedicated trends screen, which may help people develop insight into how certain activities help or hinder their recovery.

**Measuring Predictions**

Expectations play an undeniably large role in pain perception. In artificial laboratory settings where healthy participants receive painful stimulation, a wealth of findings show that expectations about pain can increase the pain experience (nocebo effect) or conversely decrease it (placebo effect) [131,132]. This has important implications for the experience of pathological pain
[133,134], where expectations about pain treatment are reliable predictors of treatment response [135]. These studies suggest that an ongoing cognitive modulation of pain is an important determinant of ongoing pain perception [136]. Expectations for pain relief most likely shift over time, the longer someone has experienced pain the more difficult the pain is to control [137,138]. Subtle longitudinal changes in expectations are believed to occur as pain becomes chronic, but capturing these changes in research studies is challenging. For this reason, we included an assessment of daily predictions about pain, mood, and activities on the SOMA app to capture how predictions vary over time within the same person or between different types of users (eg, patients with acute vs chronic pain).

We decided to assess expectations on the SOMA app by asking users to predict their expected levels of pain intensity, unpleasantness, interference, mood, and activities for the next day using the same scales used to capture their actual rating for that day. This allows the assessment of the bilateral influence of pain and mood expectations on actual experiences of pain, mood, and activities that are entered the following day.

**Discussion**

There is a great need for easy-to-use tools that help those in pain, their medical providers, and the larger health care system identify risk factors and predict the onset of chronic pain. Pain management is a rapidly evolving field that increasingly relies on assessments and treatments that are multimodal and multidisciplinary. Traditional, unidimensional assessments of self-reported pain fail to capture the nuances of pain experience and multimodal pain management. Therefore, there is an urgent need for research tools that have been specifically designed to capture this complexity.

To address this gap, we built the SOMAScience platform. Briefly, the platform uses the smartphone app SOMA to collect longitudinal, multidimensional, ESM-based pain data that capture daily pain intensity, unpleasantness, inference, mood, activities, and predictions. The SOMA app provides a simple and pleasing user interface that can promote pain self-management through visualization of pain trends over time, helping encourage individual insight into factors that exacerbate or alleviate pain. The visualizations provided can also be used to improve communication of multidimensional pain burdens to health care providers. At the same time, the larger SOMAScience platform enables user data to be included in registered single and multicenter studies.

In this paper, we have detailed the clinical and technological considerations taken into account in developing SOMAScience and the scientific rationale behind its measurements. We believe this platform is capable of meeting the requirement for tools to acquire deep and wide-ranging pain data over time, which has been largely absent from existing pain data sets. As such, SOMAScience can be used to answer a broad range of research questions, such as the correlation between initial pain dynamics and the eventual development of chronic pain (ie, predicting the transition from acute to chronic pain), evaluating both short-term and long-term effects of various treatments on pain experiences, or identifying distinct symptom clusters (ie, pain phenotypes). Moreover, the data available are sufficient to calculate more detailed multidimensional and longitudinal clinical trial or observational study end points.

The primary focus of this paper was to introduce SOMAScience as a platform for scientific studies. In the long term, we also plan to build platforms with a more clinical focus that facilitate integration with medical care (SOMAClinic) and the support of treatments (SOMAtherapeutics). This will involve connecting the SOMA app to electronic medical records and including interfaces to health trackers (eg, daily actigraphy, heart rate, or sleep data from health kits or wearables). The intention is to have a significant positive impact, both in terms of advancing research on pain and improving the lives of people with pain.

**Acknowledgments**

This work was supported by the Brainstorm program at the Robert J & Nancy D Carney Institute for Brain Science (FHP), the COBRE Center for Nervous System Function (grant NIGMS 5P20GM103645; FHP), and conducted with the help of research staff at the Center for Computation and Visualization at Brown University and volunteers from the Rhode Island School of Design. CZG has been awarded an individual research grant that covered her effort on this paper from the National Center for Complementary and Integrative Health (grant F30AT012306). We want to thank everyone who has supported the development of SOMAScience and the SOMA app over the past years, including but not limited to Benjamin Andrew, Rashi Dhar, Louis Rakovich, and Isabel Restrepo.

**Conflicts of Interest**

None declared.

Multimedia Appendix 1

The technological and design details regarding the building of the SOMA app, as well as specific measures selected for SOMA app screens, that may be of interest to some readers. [DOCX File, 22 KB - mhealth_v12i1e47177_app1.docx ]

**References**


39. SOMA science: analysis repository for SOMA science. GitHub. URL: https://github.com/SOMAAnalysis/ [accessed 2023-12-08]

40. Research inquiry. SOMA. URL: https://somatheapp.com/researchinquiry/ [accessed 2023-12-08]


43. SOMA study registration tutorial. YouTube. URL: https://www.youtube.com/watch?v=z1hpomiyXFm [accessed 2023-12-08]


46. M-Health Index and Navigation Database. MindApps. URL: https://mindapps.org [accessed 2023-12-08]


<table>
<thead>
<tr>
<th>Source</th>
<th>URL</th>
<th>Access Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>mindLAMP digital app. Division of Digital Psychiatry at BIDMC</td>
<td><a href="https://www.digitalpsychiatry.org/lamp.html">https://www.digitalpsychiatry.org/lamp.html</a></td>
<td>2023-12-12</td>
</tr>
<tr>
<td>Lamp Platform.</td>
<td><a href="https://docs.lamp.digital/">https://docs.lamp.digital/</a></td>
<td>2023-12-12</td>
</tr>
<tr>
<td>MongoDB</td>
<td><a href="https://www.mongodb.com/">https://www.mongodb.com/</a></td>
<td>2023-12-12</td>
</tr>
<tr>
<td>App designs and interactive features to increase mHealth adoption: user expectation survey and experiment. J Mheal Health 2021;9(11):e29815 [FREE Full text]</td>
<td><a href="https://doi.org/10.2196/29815">10.2196/29815</a> [Medline: 34734829]</td>
<td>2023-12-12</td>
</tr>
</tbody>
</table>

---

**Note:** The table above lists the sources and URLs provided in the document. The URLs are accessed on December 12, 2023, and the Medline IDs are included for references. If the Medline ID is not provided, the URL is referenced. The sources are cited in the context of their relevance to the topic at hand.
77. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 2005;365(9453):82-93. [doi: 10.1016/S0140-6736(04)17670-8] [Medline: 15639683]


Abbreviations

APA: American Psychiatric Association
ESM: experience sampling methodology
FDA: Food and Drug Administration
IMMPACT: Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials
mHealth: mobile health
NIH: National Institutes of Health
VAS: visual analog scale
Accuracy of Fitbit Charge 4, Garmin Vivosmart 4, and WHOOP Versus Polysomnography: Systematic Review

An-Marie Schyvens1,2, MSc; Nina Catharina Van Oost3, MSc; Jean-Marie Aerts3, MSc, PhD; Federica Masci3, MSc, PhD; Brent Peters4, MSc; An Neven5, MSc, PhD; Hélène Dirix4, MSc, PhD; Geert Wets4, MSc, PhD; Veerle Ross4,5, MSc, PhD; Johan Verbraecken1,2, MD, PhD

Corresponding Author:
An-Marie Schyvens, MSc

Abstract

Background: Despite being the gold-standard method for objectively assessing sleep, polysomnography (PSG) faces several limitations as it is expensive, time-consuming, and labor-intensive; requires various equipment and technical expertise; and is impractical for long-term or in-home use. Consumer wrist-worn wearables are able to monitor sleep parameters and thus could be used as an alternative for PSG. Consequently, wearables gained immense popularity over the past few years, but their accuracy has been a major concern.

Objective: A systematic review of the literature was conducted to appraise the performance of 3 recent-generation wearable devices (Fitbit Charge 4, Garmin Vivosmart 4, and WHOOP) in determining sleep parameters and sleep stages.

Methods: Per the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement, a comprehensive search was conducted using the PubMed, Web of Science, Google Scholar, Scopus, and Embase databases. Eligible publications were those that (1) involved the validity of sleep data of any marketed model of the candidate wearables and (2) used PSG or an ambulatory electroencephalogram monitor as a reference sleep monitoring device. Exclusion criteria were as follows: (1) incorporated a sleep diary or survey method as a reference, (2) review paper, (3) children as participants, and (4) duplicate publication of the same data and findings.

Results: The search yielded 504 candidate articles. After eliminating duplicates and applying the eligibility criteria, 8 articles were included. WHOOP showed the least disagreement relative to PSG and Sleep Profiler for total sleep time (−1.4 min), light sleep (−9.6 min), and deep sleep (−9.3 min) but showed the largest disagreement for rapid eye movement (REM) sleep (21.0 min). Fitbit Charge 4 and Garmin Vivosmart 4 both showed moderate accuracy in assessing sleep stages and total sleep time compared to PSG. Fitbit Charge 4 showed the least disagreement for REM sleep (4.0 min) relative to PSG. Additionally, Fitbit Charge 4 showed higher sensitivities to deep sleep (75%) and REM sleep (86.5%) compared to Garmin Vivosmart 4 and WHOOP.

Conclusions: The findings of this systematic literature review indicate that the devices with higher relative agreement and sensitivities to multistate sleep (ie, Fitbit Charge 4 and WHOOP) seem appropriate for deriving suitable estimates of sleep parameters. However, analyses regarding the multistate categorization of sleep indicate that all devices can benefit from further improvement in the assessment of specific sleep stages. Although providers are continuously developing new versions and variants of wearables, the scientific research on these wearables remains considerably limited. This scarcity in literature not only reduces our ability to draw definitive conclusions but also highlights the need for more targeted research in this domain. Additionally, future research endeavors should strive for standardized protocols including larger sample sizes to enhance the comparability and power of the results across studies.

(JMIR Mhealth Uhealth 2024;12:e52192) doi:10.2196/52192

KEYWORDS

sleep; wearable device; validation; polysomnography; assessing sleep; PRISMA; Preferred Reporting Items for Systematic Reviews and Meta-Analyses
Introduction

Sleep problems have emerged as a widespread concern with implications on health and quality of life for many people worldwide [1]. It has been suggested that 67% of adults worldwide have sleep problems [2,3]. The amount and quality of sleep that someone enjoys have a lasting impact during wakefulness. It affects mental health; physical well-being; and even the risk of developing lifestyle diseases such as cardiovascular diseases, obesity, depression, and type 2 diabetes [4-7]. Considering that sleep is vital to our health and quality of life, it is reasonable to wonder how long someone actually sleeps each night and if someone is getting enough restful and restorative sleep to keep the body and mind in optimal condition. Hence, this is why sleep tracking has gained immense popularity over the past few years. The majority of sleep trackers provide data on sleep architecture and hypnograms through their associated apps, offering insights into sleep stages and patterns [8]. In addition, these wearables can notify you about specific factors that might be affecting your sleep patterns such as drinking water, exercise, meditation, and regular bedtimes. As such, they can be a useful tool to obtain more insights into sleeping habits and patterns and to help optimize your sleep hygiene and quality. However, to improve sleep, an accurate, objective measurement is mandatory.

Polysomnography (PSG) is the gold-standard method for objectively assessing sleep. PSG records signals of brain activity, eye movements, and muscle tone, as well as audio and video, enabling it to classify sleep stages [9]. However, PSG may not be ideal for monitoring sleep in particular settings, as it is expensive, labor-intensive, and time-consuming; requires various equipment and technical expertise; and is impractical for long-term use or in-home environment settings [10,11]. PSG involves real-time monitoring of various parameters. In addition, applying and removing the sensors, organizing the patient administration, and thoroughly analyzing the data that PSG add is quite labor-intensive for sleep technicians. In addition to applying and removing the sensors and possibly completing questionnaires and other administrative tasks, PSG requires an overnight stay at a sleep clinic or laboratory, which makes PSG time-consuming for both the sleep technicians and patients. Due to the inherent limitations of PSG, several alternatives have been proposed. First, the use of sleep diaries is inexpensive and straightforward for consumers, but the subjective self-ratings they require result in frequent inaccuracies and incompleteness [12]. Additionally, they fail to measure sleep architecture and stages. Second, electroencephalogram (EEG) wearables can provide a home evaluation of sleep architecture and stages; however, they come with a high cost and can be technologically complex [12]. Therefore, wearables using an accelerometer and photoplethysmography (PPG) are being explored as a feasible alternative, largely due to their lower cost, convenience, and ability to measure sleep in clinical and personal settings [13].

Accelerometers and PPG sensors monitor different physiological and movement patterns throughout the night. Accelerometers are small, electromechanical devices that measure acceleration along multiple axes (usually 3: x, y, and z) to detect position changes, turning over, or significant body movements during the night [14]. Due to the body movement variations specific to each sleep stage, accelerometers can provide information about wakefulness and general sleep stages. However, they may tend to overestimate sleep due to poorly distinguishing between sleep and sedentary supine wake periods (eg, lying down while reading or watching television), or they could underestimate sleep due to potential body movements during sleep being categorized as awakenings [12,15-18]. In addition, they may not be as accurate in distinguishing between different non–rapid eye movement (REM) stages and detecting subtle changes in sleep architecture [14]. By combining an accelerometer with a PPG sensor in wearables, a more comprehensive and accurate assessment of sleep could be provided. PPG sensors are a noninvasive technology that uses a light source and a photodetector at the surface of the skin to measure the volumetric variations of blood circulation and thus can be used to monitor heart rate, heart rate variability, blood flow, and blood oxygen levels [19-21]. Due to the specific cardiovascular features of each sleep stage, PPG can provide more information about the sleep stages in addition to the accelerometer [21,22]. The benefits of these sensors used in wearables are their low-cost, noninvasive nature and their ability to provide continuous monitoring and real-time data. However, the readings of PPG can be affected by motion artifacts, skin pigmentation, or tissue thickness. In addition, they could be susceptible to environmental factors such as ambient light and temperature [23-25].

Although many have doubts about their accuracy in monitoring sleep, wearable sleep-tracking devices are widely used and becoming more technologically advanced, creating strong interest from researchers and clinicians for their possible use as alternatives to PSG. This was demonstrated by Nguyen et al [26], who used wearables that provide inactivity alerts and personal feedback to increase physical activity and improve sleep for survivors of breast cancer [26].

Given the disadvantages of PSG and the corresponding growing popularity of wearable devices for sleep tracking among consumers and medical organizations, the objective of this paper was to appraise the performance of recent-generation wearable devices in determining sleep parameters and sleep stages through a review of relevant publications. To limit the overwhelming amount of wearables and their corresponding research papers, we performed a search to select a limited number of recent, frequently used wearables, using the following criteria: recent generation; good ease of use (affordable, unobtrusive, and sufficient battery life); and assessment of variables that could also be used for monitoring sleep, stress, fatigue, and sleepiness—namely, heart rate, heart rate variability, stress indicator, and activity. The candidate wearables selected up-front were Fitbit Charge 4, Garmin Vivosmart 4, and WHOOP.

Methods

Search Strategies

In adherence with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement (Checklist 1), a comprehensive search using the PubMed, Web of Science, Google Scholar, Scopus, and Embase databases was conducted
[27]. Relevant keywords such as “validity,” “accuracy,” “assessment,” “performance,” “wearable,” “sleep tracker,” “sleep-tracking,” “polysomnography,” “wristband,” “Whoop,” “Fitbit Charge 4,” and “Garmin Vivosmart 4” were used (see Multimedia Appendix 1). This search was initially completed by May 16, 2023, and was repeated by November 23, 2023 (see Figure 1). This systematic literature review was not registered and a protocol was not predefined.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram describing the search strategy of databases to retrieve and qualify publications of relevance for this systematic literature review.

Eligibility Criteria
Retrieved publications qualified for the review if they (1) involved the validity of sleep data of any marketed model of the candidate wearables and (2) incorporated PSG or an ambulatory EEG monitor as a reference sleep monitoring device. Exclusion criteria were as follows: (1) incorporated a sleep diary or survey method as a reference, (2) review paper, (3) children as participants, and (4) duplicate publication of the same data and findings.

Data Extraction
The following items were extracted: type of sleep tracker; number, gender, and age of participants; number of nights of sleep assessment; reference sleep monitoring device; and study outcomes relative to the denoted reference standard—the precision of measuring the parameters of total sleep time (TST),
light sleep (LS), deep sleep (DS), and REM sleep, as well as the sensitivity to sleep (the proportion of correctly classified sleep epochs by the wearable); specificity for sleep (the proportion of correctly classified wake epochs by the wearable); agreement; and if applicable, Cohen $\kappa$ for multistate categorization of sleep periods. These Cohen $\kappa$ values are a measure of interdevice reliability, often used in the context of validation studies where 2 or more methods of devices are used to assess a particular characteristic or condition [28]. In the case of wearable validation studies for sleep, $\kappa$ values are often used to assess the agreement between the wearable device’s sleep detection algorithm and a reference standard, such as PSG. In addition to the observed agreement between the devices, $\kappa$ values also take into account the possibility that the agreement comes by chance. The interpretation of $\kappa$ values is often categorized as follows: values $\leq 0$ indicate no agreement, 0.01-0.20 indicate none to slight agreement, 0.21-0.40 indicate fair agreement, 0.41-0.60 indicate moderate agreement, 0.61-0.80 indicate substantial agreement, and 0.81-1.00 indicate almost perfect agreement [28].

**Results**

**Overview of Included Studies**

Figure 1 presents a visual summary of the selection and qualification of articles for this review. A total of 8 publications were retrieved through a search of databases performed by May 2023 and again by November 2023. Table 1 presents an overview of all included studies and the extracted details of each qualifying study involving the different wearable models. Few papers that met the eligibility criteria for the candidate wearables were found; we found 3 papers that satisfy the criteria for Fitbit Charge 4 [28-30], 2 for Garmin Vivosmart 4 [27,28], and 4 for WHOOP [22,30-32].

**Table 1.** Summary table of included papers in this systematic literature review.

<table>
<thead>
<tr>
<th>Author (year)</th>
<th>Wearable</th>
<th>Reference</th>
<th>Participants, n</th>
<th>Sex, n</th>
<th>Age (y), mean (SD)</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doheny et al [33] (2021)</td>
<td>Fitbit Charge 4</td>
<td>PSG$^a$</td>
<td>2</td>
<td>• N/A$^b$</td>
<td>N/A</td>
<td>1 night, laboratory based; followed by 7 nights at home</td>
</tr>
<tr>
<td>Renerts et al [34] (2022)</td>
<td>Fitbit Charge 4</td>
<td>PSG</td>
<td>8</td>
<td>• N/A</td>
<td>N/A</td>
<td>1 night, laboratory based</td>
</tr>
<tr>
<td>Dong et al [18] (2022)</td>
<td>Fitbit Charge 4</td>
<td>PSG</td>
<td>37</td>
<td>• Female: 20</td>
<td>48.8 (2.1)</td>
<td>1 night, laboratory based</td>
</tr>
<tr>
<td>• Male: 17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouritzen et al [29] (2020)</td>
<td>Garmin Vivosmart 4</td>
<td>PSG</td>
<td>18</td>
<td>• Female: 13</td>
<td>56.1 (12.0)</td>
<td>1 night, laboratory based</td>
</tr>
<tr>
<td>• Male: 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stone et al [30] (2020)</td>
<td>Garmin Vivosmart 4 and WHOOP Sleep Profiler (ambulatory EEG$^c$ monitor)</td>
<td>PSG</td>
<td>5</td>
<td>• Female: 3</td>
<td>27.8 (7.6)</td>
<td>Home environment, 98 nights for all study devices</td>
</tr>
<tr>
<td>• Male: 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miller et al [31] (2020)</td>
<td>WHOOP</td>
<td>PSG</td>
<td>12</td>
<td>• Female: 6</td>
<td>22.9 (3.4)</td>
<td>10 days, laboratory based</td>
</tr>
<tr>
<td>• Male: 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miller et al [32] (2021)</td>
<td>WHOOP</td>
<td>PSG</td>
<td>6</td>
<td>• Female: 3</td>
<td>23.0 (2.2)</td>
<td>9 nights, laboratory based</td>
</tr>
<tr>
<td>• Male: 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miller et al [22] (2022)</td>
<td>WHOOP</td>
<td>PSG</td>
<td>53</td>
<td>• Male: 27</td>
<td>25.4 (5.9)</td>
<td>1 night, laboratory based</td>
</tr>
<tr>
<td>• Female: 26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$^a$PSG: polysomnography.
$^b$N/A: not applicable.
$^c$EEG: electroencephalogram.

Participants were diverse: healthy adults as well as participants diagnosed with Huntington disease [33,34] and chronic insomnia [18]. Sample size varied substantially between investigations, from 2 to 53 participants. The average age of the participants was 34 years old. Out of the 8 studies, 6 (75%) were fully conducted in a sleep laboratory [18,22,29,31,32,34], 1 (12%) was conducted in the home environment [30], and 1 (12%) conducted an overnight laboratory-based PSG followed by 7 nights in the home environment [33]. In all, 3 (38%) laboratory-based studies had a duration of more than 1 night [30-32].

**Comparison of Sleep Parameters Assessed by Wearables Versus PSG**

**Fitbit Charge 4**

As shown in Table 2, of the 3 Fitbit Charge 4 versus PSG comparisons, 2 (67%) reported an overestimation of TST (5 and 23 minutes), whereas 1 (33%) reported a nonsignificant underestimation of 11 minutes. For LS, 2 (67%) of the 3 comparisons reported a similar significant overestimation (37.5 and 37.7 minutes). For DS, all 3 papers reported an underestimation (4.1, 12.5, and 41.4 minutes). REM sleep was overestimated in 2 (67%) out of 3 papers by 5.2 and 11.5
minutes and underestimated in 1 (33%) paper by 4.7 minutes. The sensitivity to sleep was quite high for all 3 papers (89.9%-93.6%). In contrast, the results for specificity for sleep were quite divergent (48.8%, 62.2%, and 73%). The sensitivities to DS and REM sleep could only be extracted from 2 papers: 54% and 96% for DS and 76% and 97% for REM sleep.

Table 3 shows the results for the sensitivity to and specificity for sleep. The sensitivity to sleep was quite high for all 3 papers (89.9%-93.6%). In contrast, the results for specificity for sleep were quite divergent (48.8%, 62.2%, and 73%). The sensitivities to DS and REM sleep could only be extracted from 2 papers: 54% and 96% for DS and 76% and 97% for REM sleep.

<table>
<thead>
<tr>
<th>Wearable and paper</th>
<th>Fitbit Charge 4</th>
<th>Garmin Vivosmart 4</th>
<th>WHOOP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TST (min)</td>
<td>LS (min)</td>
<td>DS (min)</td>
</tr>
<tr>
<td></td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
</tr>
<tr>
<td></td>
<td>P value</td>
<td>P value</td>
<td>P value</td>
</tr>
<tr>
<td>Doheny et al [33]</td>
<td>23 (N/A)</td>
<td>37.5 (N/A)</td>
<td>-12.5 (N/A)</td>
</tr>
<tr>
<td>Renerts et al [34]</td>
<td>5 (26.8)</td>
<td>N/A</td>
<td>-4.1 (21.8)</td>
</tr>
<tr>
<td>Dong et al [18]</td>
<td>-11.0 (N/A)</td>
<td>37.69 (N/A)</td>
<td>-41.38 (N/A)</td>
</tr>
<tr>
<td>Mean</td>
<td>5.67 (N/A)</td>
<td>37.6 (N/A)</td>
<td>-19.33 (N/A)</td>
</tr>
<tr>
<td>Mouritzen et al [29]</td>
<td>27.8 (29.5)</td>
<td>36.5 (71.7)</td>
<td>13.4 (98.1)</td>
</tr>
<tr>
<td>Stone et al [30]</td>
<td>66 (N/A)</td>
<td>19.8 (N/A)</td>
<td>33.6 (N/A)</td>
</tr>
<tr>
<td>Mean</td>
<td>46.9 (N/A)</td>
<td>28 (N/A)</td>
<td>23.5 (N/A)</td>
</tr>
<tr>
<td>Stone et al [30]</td>
<td>16.2</td>
<td>-10.2</td>
<td>1.8</td>
</tr>
<tr>
<td>Miller et al [31]</td>
<td>8.2 (32.9)</td>
<td>-3.7 (44.4)</td>
<td>-3.7 (26.4)</td>
</tr>
<tr>
<td>Miller et al [32]</td>
<td>-17.8 (61.1)</td>
<td>-8.9 (55.9)</td>
<td>-15.5 (30.1)</td>
</tr>
<tr>
<td>Miller et al [32]</td>
<td>-12.2 (36.3)</td>
<td>-15.6 (50.7)</td>
<td>-19.6 (34.3)</td>
</tr>
<tr>
<td>Mean</td>
<td>-1.4 (N/A)</td>
<td>-9.6 (N/A)</td>
<td>-9.25 (N/A)</td>
</tr>
</tbody>
</table>

aN/A: not applicable.
Table. The sensitivity to sleep, specificity for sleep, agreement, and Cohen κ coefficient for multistate categorization of Fitbit Charge 4, Garmin Vivosmart 4, and WHOOP versus polysomnography.

<table>
<thead>
<tr>
<th>Wearable and paper</th>
<th>Sensitivity to sleep (%), mean (SD)</th>
<th>Specificity for sleep (%), mean (SD)</th>
<th>Sensitivity to light sleep (%), mean (SD)</th>
<th>Sensitivity to deep sleep (%), mean (SD)</th>
<th>Sensitivity to REM sleep (%), mean (SD)</th>
<th>Agreement (%), mean (SD)</th>
<th>Cohen κ for multistate categorization, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fitbit Charge 4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doheny et al [33]</td>
<td>93.6 (2.6)</td>
<td>48.8 (17.7)</td>
<td>N/A</td>
<td>96 (N/A)</td>
<td>97 (N/A)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Renerts et al [34]</td>
<td>90 (N/A)</td>
<td>73 (N/A)</td>
<td>N/A</td>
<td>54 (N/A)</td>
<td>76 (N/A)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Dong et al [18]</td>
<td>89.9 (4.0)</td>
<td>62.2 (26.2)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Mean</td>
<td>91.2 (N/A)</td>
<td>61.3 (N/A)</td>
<td>N/A</td>
<td>75 (N/A)</td>
<td>86.5 (N/A)</td>
<td>NA</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Garmin Vivosmart 4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouritzen et al [29]</td>
<td>98 (3)</td>
<td>30 (17)</td>
<td>60 (17)</td>
<td>45 (26)</td>
<td>34 (26)</td>
<td>48 (10)</td>
<td>0.20 (0.11)</td>
</tr>
<tr>
<td>Stone et al [30]</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Mean</td>
<td>98 (3)</td>
<td>30 (17)</td>
<td>60 (17)</td>
<td>45 (26)</td>
<td>34 (26)</td>
<td>48 (10)</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>WHOOP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stone et al [30]</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Miller et al [31]</td>
<td>95 (N/A)</td>
<td>51 (N/A)</td>
<td>62 (N/A)</td>
<td>68 (N/A)</td>
<td>70 (N/A)</td>
<td>64 (N/A)</td>
<td>0.47 (N/A)</td>
</tr>
<tr>
<td>Miller et al [32]</td>
<td>90 (N/A)</td>
<td>60 (N/A)</td>
<td>61 (N/A)</td>
<td>64 (N/A)</td>
<td>66 (N/A)</td>
<td>63 (N/A)</td>
<td>0.47 (N/A)</td>
</tr>
<tr>
<td>Miller et al [22]</td>
<td>90 (N/A)</td>
<td>56 (N/A)</td>
<td>58 (N/A)</td>
<td>62 (N/A)</td>
<td>66 (N/A)</td>
<td>60 (N/A)</td>
<td>0.44 (N/A)</td>
</tr>
<tr>
<td>Mean</td>
<td>91.7 (N/A)</td>
<td>56 (N/A)</td>
<td>60 (N/A)</td>
<td>65 (N/A)</td>
<td>67 (N/A)</td>
<td>62 (N/A)</td>
<td>0.46 (N/A)</td>
</tr>
</tbody>
</table>

aREM: rapid eye movement.
bN/A: not applicable.

**Garmin Vivosmart 4**

Both papers comparing Garmin Vivosmart 4 to PSG or SleepProfiler reported an overestimation of TST (27.8 and 66 minutes), as shown in Table 2. For LS, both papers again reported an overestimation (19.8 and 36.5 minutes). For DS, both papers reported an overestimation (13.4 and 33.6 minutes). On the contrary, REM sleep was underestimated by both papers by 3 and 22.1 minutes. The study from Mouritzen et al [29] reported sensitivities to LS, DS, and REM sleep of 60%, 65%, and 34%, respectively; a sensitivity to sleep of 98%; a specificity for sleep of 98%; and an agreement of 48%; and a Cohen κ for multistate categorization of sleep periods of 0.20 (see Table 3).

**WHOOP**

As shown in Table 2, of the 4 WHOOP versus PSG comparisons, 2 (50%) reported an overestimation of TST (8.2 and 16.2 minutes), whereas the other 2 (50%) reported an underestimation (12.2 and 17.8 minutes). All 4 comparisons reported an underestimation of LS, from 3.7 to 15.6 minutes. For DS, 3 (75%) papers reported an underestimation (3.7, 15.5, and 19.6 minutes). Stone et al [30] reported a small overestimation of DS of 1.8 minutes. REM sleep was overestimated in all 4 studies, from 6 to 22.9 minutes. Table 3 shows that the studies from Miller et al [22,31,32] reported sensitivities to LS, DS, and REM sleep ranging from 58% to 62%, from 62% to 68%, and from 66% to 70%, respectively. The studies also reported a sensitivity to sleep ranging from 90% to 95%, a specificity for sleep ranging from 51% to 60%, an agreement ranging from 60% to 64%, and a Cohen κ for multistate categorization of sleep periods ranging from 0.44 to 0.47.

**Comparison of the Mean Values of the Different Wearables**

From Tables 2 and 3, the means of the different values extracted from the papers (ie, TST, LS, DS, REM sleep, sensitivity to sleep, specificity for sleep, sensitivity to LS, sensitivity to DS, sensitivity to REM sleep, agreement, and Cohen κ coefficient) were calculated, which are summarized in Table 4. Altogether, WHOOP deviated the least compared to the gold-standard PSG for TST, LS, and DP but showed the highest difference from PSG for REM sleep (ie, a mean overestimation of 21 min).
Table. Mean differences of the sleep parameters in minutes and means of sensitivities to sleep, specificity for sleep, and agreement assessed by the wearables compared to polysomnography.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Fitbit Charge 4</th>
<th>Garmin Vivosmart 4</th>
<th>WHOOP</th>
</tr>
</thead>
<tbody>
<tr>
<td>TST* (min)</td>
<td>5.7</td>
<td>46.9</td>
<td>−1.4</td>
</tr>
<tr>
<td>LS† (min)</td>
<td>37.6</td>
<td>27.9</td>
<td>−9.6</td>
</tr>
<tr>
<td>DS‡ (min)</td>
<td>−19.2</td>
<td>23.5</td>
<td>−9.3</td>
</tr>
<tr>
<td>REM§ sleep (min)</td>
<td>4.0</td>
<td>−12.5</td>
<td>21.0</td>
</tr>
<tr>
<td>Sensitivity to sleep (%)</td>
<td>91.2</td>
<td>98.0</td>
<td>91.7</td>
</tr>
<tr>
<td>Sensitivity to LS (%)</td>
<td>N/A</td>
<td>60.0</td>
<td>60.0</td>
</tr>
<tr>
<td>Sensitivity to DS (%)</td>
<td>75.0</td>
<td>45.0</td>
<td>65.0</td>
</tr>
<tr>
<td>Specificity for sleep (%)</td>
<td>86.5</td>
<td>34.0</td>
<td>67.0</td>
</tr>
<tr>
<td>Agreement (%)</td>
<td>N/A</td>
<td>48</td>
<td>62</td>
</tr>
<tr>
<td>Cohen κ</td>
<td>N/A</td>
<td>0.20</td>
<td>0.46</td>
</tr>
<tr>
<td>Papers, n</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

*TST: total sleep time.
†LS: light sleep.
‡DS: deep sleep.
§REM: rapid eye movement.
N/A: not applicable.

Garmin Vivosmart 4 showed the largest deviations out of the 3 different wearables compared to PSG (ie, a mean overestimation of 46.9 minutes for TST, a mean overestimation of 27.9 minutes for LS, a mean overestimation of 23.5 minutes for DS, and a mean underestimation of 12.5 minutes for REM sleep). Additionally, the sensitivities to LS, DS, and REM sleep were lower compared to those of Fitbit Charge 4 and WHOOP.

The mean values of Fitbit Charge 4 deviated the least from PSG for REM sleep, with a mean overestimation of only 4 minutes. On the contrary, it showed the largest deviation to PSG for LS, with a mean overestimation of 37.6 minutes. For TST and DS, Fitbit Charge 4 showed on average better results than Garmin Vivosmart 4 but worse results than WHOOP, namely, a mean overestimation of 5.7 minutes and a mean underestimation of 19.2 minutes, respectively. The sensitivities to LS, DS, and REM sleep were higher compared to those of Garmin Vivosmart 4 and WHOOP.

Discussion

PSG is still the gold-standard method to objectively assess sleep. However, PSG is not ideal for monitoring sleep in particular settings and for long-term follow-up. To overcome these limitations, consumer sleep-tracking devices are becoming more widely used and technologically advanced, creating strong interest from researchers and clinicians for their possible application as alternatives to PSG. Since limited research has been performed to validate the different consumer sleep-tracking wearables, we aimed to review the available literature on the selected wearables to determine the most accurate, commercially available wrist-worn device that can be used in a clinical setting for long-term sleep monitoring.

Thus far, at most, only 7 relevant studies investigated the performance of Fitbit Charge 4, Garmin Vivosmart 4, and WHOOP against PSG. The study of Stone et al [30] investigated the performance of both Garmin Vivosmart 4 and WHOOP relative to Sleep Profiler. After reviewing these studies, the results suggest that WHOOP presented the least amount of disagreement relative to PSG and Sleep Profiler for TST, LS, and DS but showed the largest amount of disagreement for REM sleep. Fitbit Charge 4 and Garmin Vivosmart 4 both showed moderate accuracy in assessing sleep stages and TST compared to PSG. Fitbit Charge 4 showed the least amount of disagreement for REM sleep relative to PSG. In addition, Fitbit Charge 4 showed higher sensitivities for LS, DS, and REM sleep compared to Garmin Vivosmart 4 and WHOOP. Garmin Vivosmart 4 showed the lowest sensitivities to LS, DS, and REM sleep compared to Fitbit Charge 4 and WHOOP.

Some of the studies performed evaluations of the accuracy of wearables in detecting sleep by using an epoch-by-epoch analysis. It involves breaking down the continuous stream of sleep data of PSG into discrete time intervals called “epochs” [35]. Afterward, each epoch is compared individually to the corresponding epoch generated by the wearables. The results of the epoch-by-epoch analysis in this review showed high sensitivity to sleep, ranging from 91.2% to 98.0%, but lower specificity for sleep, ranging from 30% to 61%. The low specificity for sleep, or variability in specificity for sleep, is a commonly observed phenomenon in the validation of devices that primarily rely on actigraphy to estimate sleep [15,36,37].
The challenge of accurately separating wake episodes during sleep stems from the similarities in movement between restful wakefulness and sleep. Hence, it can be inferred that devices that have improved their ability to detect wake epochs during sleep have refined their proprietary algorithms to include metrics other than movement, such as heart rate and heart rate variability, in the detection of wakefulness [22].

Furthermore, it is crucial to contextualize the comparison of the wearables’ agreement with that of PSG (see Multimedia Appendix 2), taking into consideration that the scoring of PSG is subject to variability among technicians [38]. As reported by Danker-Hopfe et al [39], the intrarater reliability ranges from 86.5% to 97.5% depending on the sleep stage, with an overall accuracy of 81% and a Cohen κ coefficient of 0.7505 (see Table 5). Given this benchmark, Fitbit Charge 4 seems to provide reasonable estimations of multistate sleep. However, it is imperative to acknowledge that achieving the same level of accuracy as PSG may pose a significant challenge for wearables, particularly considering the observed low Cohen κ coefficients of the wearables. The finding of a low κ value indicates poor sleep stage differentiation despite including PPG signals. However, the beneficial role of PPG signals in accelerometer-based sleep tracking remains unilluminated, since we do not know how these signals are processed and applied [29].

### Table 5. Sleep stage–specific degree of agreement according to American Academy of Sleep Medicine standards [38].

<table>
<thead>
<tr>
<th>Stage</th>
<th>Agreement (%)</th>
<th>Cohen κ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>81</td>
<td>0.7505</td>
</tr>
<tr>
<td>Wake</td>
<td>95.6</td>
<td>0.4608</td>
</tr>
<tr>
<td>REM&lt;sup&gt;a&lt;/sup&gt;</td>
<td>97.5</td>
<td>0.9054</td>
</tr>
<tr>
<td>N3&lt;sup&gt;b&lt;/sup&gt;</td>
<td>93.8</td>
<td>0.7285</td>
</tr>
<tr>
<td>N2&lt;sup&gt;c&lt;/sup&gt;</td>
<td>86.5</td>
<td>0.7188</td>
</tr>
<tr>
<td>N1&lt;sup&gt;d&lt;/sup&gt;</td>
<td>90.1</td>
<td>0.4608</td>
</tr>
</tbody>
</table>

<sup>a</sup>REM: rapid eye movement.
<sup>b</sup>N3: stage 3 non-REM sleep.
<sup>c</sup>N2: stage 2 non-REM sleep.
<sup>d</sup>N1: stage 1 non-REM sleep.

On a similar note, companies usually do not share the methodology they use to score the sleep data from the wearables, nor do they publish the kind of rigorous research sleep experts need to establish the credibility of the sleep reports they produce.

The findings of this systematic literature review about Fitbit Charge 4, Garmin Vivosmart 4, and WHOOP are based on our recent comprehensive search of databases for relevant published articles. The included research studies have certain limitations. For example, several investigations evaluating the sleep-tracking capabilities of these wearables involved a relatively limited number of participants, potentially impacting the generalizability of the results. Efforts were made to get additional information about the sample sizes and participants directly from the authors of some included studies. Unfortunately, despite our attempts to contact the authors, we did not receive a response. It is crucial to address these issues transparently, as they may impact the generalizability of our findings. Small sample sizes can introduce variability and limit the statistical power of the findings, underscoring the need for larger and more diverse cohorts to validate the devices’ accuracy across different demographics and sleep conditions. Additionally, it is essential to consider that most of the studies have been conducted in controlled environments, namely sleep laboratories. Although this controlled setting allows for precise data collection and monitoring, it may not fully reflect the real-world sleep experiences of individuals in their natural environments [40].

Sleep laboratory conditions may differ substantially from home environments, where factors such as ambient light, noise, and personal sleep habits can vary widely [41]. Therefore, the generalizability of the findings from laboratory studies to everyday scenarios should be approached with caution. In addition to acknowledging the potential limitations associated with conducting studies in controlled sleep laboratory environments, it is important to recognize the presence of the “first-night effect” in both PSG and wearable sleep-tracking technologies. The first-night effect refers to the phenomenon where an individual’s sleep patterns and quality may be altered due to the unfamiliarity with the sleep-monitoring setup, regardless of whether it occurs in a sleep laboratory or at home [42,43]. This phenomenon is not exclusive to sleep laboratories as it extends to home environments where individuals may experience similar disruptions during the initial adaptation to sleep-tracking devices [44,45]. Although some wearable users may initially find it uncomfortable or unfamiliar to wear a device on their wrist while sleeping, which could potentially impact their sleep quality on the first night of use, there is not a widely recognized first-night effect associated with the use of wrist-worn wearables [46]. In addition, the data of PSG and wearables were collected each time under the same circumstances and environmental factors either in a sleep laboratory or in the home environment, making the first-night effect less relevant. Another notable aspect of the studies is the variations in the duration of the validation of the wearables. The diverse durations across studies are likely due to practical constraints or differences in research protocols. Our decision to include studies with varying protocols was motivated by the aim to include as many relevant papers as possible in our...
comprehensive review, considering the scarcity of literature on this topic. Future research endeavors should strive for standardized protocols including larger sample sizes to enhance the comparability and power of the results across studies.

Despite these limitations, the findings of this review indicate that the devices with higher relative agreement and sensitivities for multistate sleep (ie, Fitbit Charge 4 and WHOOP) seem appropriate for deriving suitable estimates of sleep parameters and could be used to monitor sustained, meaningful changes in sleep architecture (ie, time spent in different stages of sleep). However, analyses regarding the multistate categorization of sleep (as a specific sleep stage or wake) indicate that all devices can benefit from further improvement. Providers are continuously developing new versions and variants of wearables, which present difficulties for those undertaking independent validation studies. Nevertheless, it can be reasonably assumed that newer models from the same provider will perform at least as well, if not better, than older models when compared against the relevant gold standards [22]. However, although the wearable technology market keeps developing wearable devices, the scientific research on these wearables against PSG remains considerably limited. This scarcity in literature not only reduces our ability to draw definitive conclusions but also highlights the need for more targeted research in this domain. Therefore, the data presented here should not be considered obsolete when the models analyzed are superseded by newer models. Instead, these data should serve as the best approximation of the expected performance of any subsequent models that may be released.

Acknowledgments
This work was supported by Flanders Innovation & Entrepreneurship - VLAIO (Vlaams Agentschap Innoveren & Ondernemen; HBC.2021.0387).

Authors' Contributions
AMS and JV conceptualized and designed the structure of this systematic literature review, outlining the key themes and objectives. AMS conducted an extensive literature review, critically analyzed relevant studies, and synthesized the information to provide a comprehensive overview. AMS, NCVO, JMA, FM, BP, AN, HD, VR, GW, and JV actively participated in the review's writing and revision processes. The collaborative effort of the consortium resulted in a comprehensive and well-balanced review. Each author has approved the final version of the manuscript.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Search string. In adherence with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement, a comprehensive search of the PubMed, Web of Science, Google Scholar, Scopus, and Embase databases was conducted as shown in the search string.

[PDF File, 49 KB - mhealth_v12i1e52192_app1.pdf ]

Multimedia Appendix 2
Normative values of sleep parameters. These normative values in healthy male and female individuals of different age groups (mean and SD), derived from the widely recognized Rechtschaffen and Kales scoring system of polysomnography, can provide an additional benchmark when delving into the comparative analysis of sleep parameters measured by wearables [47].

[PDF File, 21 KB - mhealth_v12i1e52192_app2.pdf ]

Checklist 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.

[PDF File, 64 KB - mhealth_v12i1e52192_app3.pdf ]

References


Abbreviations

- **DS**: deep sleep
- **EEG**: electroencephalogram
- **LS**: light sleep
- **PPG**: photoplethysmography
- **PRISMA**: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
- **PSG**: polysomnography
- **REM**: rapid eye movement
- **TST**: total sleep time
Accuracy of Fitbit Charge 4, Garmin Vivosmart 4, and WHOOP Versus Polysomnography: Systematic Review

Schyvens AM, Van Oost NC, Aerts JM, Masci F, Peters B, Neven A, Dirix H, Wets G, Ross V, Verbraecken J

Please cite as:
Schyvens AM, Van Oost NC, Aerts JM, Masci F, Peters B, Neven A, Dirix H, Wets G, Ross V, Verbraecken J
Accuracy of Fitbit Charge 4, Garmin Vivosmart 4, and WHOOP Versus Polysomnography: Systematic Review
JMIR Mhealth Uhealth 2024;12:e52192
URL: https://mhealth.jmir.org/2024/1/e52192
doi:10.2196/52192

© An-Marie Schyvens, Nina Catharina Van Oost, Jean-Marie Aerts, Federica Masci, Brent Peters, An Neven, Hélène Dirix, Geert Wets, Veerle Ross, Johan Verbraecken. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 27.3.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Clinical Efficacy of Mobile App–Based, Self-Directed Pulmonary Rehabilitation for Patients With Chronic Obstructive Pulmonary Disease: Systematic Review and Meta-Analysis

Chiwook Chung1,2, MD; Jong Won Lee3, MD, PhD; Sei Won Lee1,*, MD, PhD; Min-Woo Jo4,*, MD, PhD

1
2
3
4
* these authors contributed equally

Corresponding Author:
Sei Won Lee, MD, PhD

Abstract

Background: Pulmonary rehabilitation is well known to improve clinical symptoms (including dyspnea), quality of life, and exercise capacity in patients with chronic obstructive pulmonary disease (COPD). However, researchers have reported difficulties in practicing center-based pulmonary rehabilitation. Recently, mobile app–based pulmonary rehabilitation has become available in clinical practice. We investigated the clinical outcomes of mobile app–based pulmonary rehabilitation in patients with COPD.

Objective: The objective of our study was to evaluate the clinical efficacy of mobile app–based pulmonary rehabilitation versus conventional center-based pulmonary rehabilitation for patients with COPD, using a systematic review and meta-analysis.

Methods: A systematic search of the literature published between January 2007 and June 2023 was performed, using the PubMed, Embase, Cochrane, and CINAHL databases to identify relevant randomized controlled trials involving patients with COPD. Pulmonary rehabilitation programs needed to provide an exercise program on a smartphone app. Study outcomes, including exercise capacity, symptom scores, quality of life, and hospitalization, were evaluated. The meta-analysis evaluated mean differences in 6-minute walk test distances (6MWDs), COPD Assessment Test (CAT) scores, modified Medical Research Council (mMRC) dyspnea scale scores, St. George Respiratory Questionnaire (SGRQ) scores, and risk ratios for hospitalization resulting from disease exacerbation.

Results: Of the 1173 screened studies, 10 were included in the systematic review and 9 were included in the meta-analysis. Further, 6 studies were multicenter studies. There were a total of 1050 participants, and most were aged ≥65 years. There were discrepancies in the baseline participant characteristics, smartphone apps, interventions, and study outcomes among the included studies. In the meta-analysis, 5 studies assessed 6MWDs (mean difference 9.52, 95% CI −3.05 to 22.08 m), 6 studies assessed CAT scores (mean difference −1.29, 95% CI −2.39 to −0.20), 3 studies assessed mMRC dyspnea scale scores (mean difference −0.08, 95% CI −0.29 to 0.13), 2 studies assessed SGRQ scores (mean difference −3.62, 95% CI −9.62 to 2.38), and 3 studies assessed hospitalization resulting from disease exacerbation (risk ratio 0.65, 95% CI 0.27-1.53). These clinical parameters generally favored mobile app–based pulmonary rehabilitation; however, a statistically significant difference was noted only for the CAT scores (P=.02).

Conclusions: Despite some discrepancies in the baseline participant characteristics and interventions among studies, mobile app–based pulmonary rehabilitation resulted in favorable exercise capacity, symptom score, quality of life, and hospitalization outcomes when compared with conventional pulmonary rehabilitation. In the meta-analysis, the CAT scores of the mobile app–based pulmonary rehabilitation group were significantly lower than those of the control group (P=.02). In real-world practice, mobile app–based pulmonary rehabilitation can be a useful treatment option when conventional center-based pulmonary rehabilitation is not feasible.

Trial Registration: PROSPERO CRD42023466965; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=466965

(JMIR Mhealth Uhealth 2024;12:e41753) doi:10.2196/41753

KEYWORDS
pulmonary rehabilitation; COPD; chronic obstructive pulmonary disease; mobile application; mobile app; 6MWD; 6-minute walk test distance; CAT; COPD Assessment Test; mMRC; modified Medical Research Council; SGRQ; St. George Respiratory Questionnaire; exacerbation; rehabilitation; mHealth; mobile health; clinical efficacy; PRISMA; mobile phone
Introduction

Chronic obstructive pulmonary disease (COPD) is characterized by persistent respiratory symptoms and airflow limitation, which are usually caused by exposure to noxious gases or particles [1]. Recently, the prevalence of COPD has increased, making it a leading cause of morbidity and mortality worldwide [2,3]. Approximately 3,500,000 people experience COPD, and it is the third leading cause of disability-adjusted life years (1305 disability-adjusted life years per 100,000 population, 6.21% of total noncommunicable diseases disability-adjusted life years) in South Korea [4,5]. COPD has various extrapulmonary features and might be a systemic disease rather than a disease that only affects the airway [6]. Various clinical information is relevant to the mortality of patients with COPD, including information on physical activity, disability, lung function, long-term oxygen therapy, BMI, quality of life, depressive symptoms, marital status, comorbidity, and hospitalization [7-9]. Additionally, the BODE (BMI, Airflow Obstruction, Dyspnea, and Exercise Capacity) index, which includes BMI, airflow obstruction as assessed by the forced expiratory volume in 1 second (FEV₁), dyspnea as assessed by the modified Medical Research Council (mMRC) dyspnea scale, and exercise capacity as assessed by the 6-minute walk test distance (6MWD), is well known to predict mortality in patients with COPD [10,11].

Pulmonary rehabilitation is a comprehensive intervention for improving the physical and psychological conditions of people with chronic respiratory diseases through exercise training, education, and behavior modification [12]. Pulmonary rehabilitation has been shown to improve dyspnea, quality of life, and exercise capacity in patients with COPD [1,12-14]. Furthermore, patients with chronic respiratory diseases have decreased respiratory muscle mass and strength, which are accompanied by decreased respiratory function. In this population, pulmonary rehabilitation with exercise training is the only way to improve respiratory function [15]. The pulmonary rehabilitation programs used in previous landmark studies were composed of exercise training that was performed 30 to 45 minutes per day, 3 to 5 days per week, for at least 8 to 12 weeks [16,17]. However, researchers reported difficulties in practicing center-based pulmonary rehabilitation, including a lack of facilities; low health insurance coverage; a lack of awareness among physicians; and a lack of motivation, transport barriers, and low levels of social support among patients [4,18,19]. Thus, alternatives to center-based pulmonary rehabilitation are desperately needed [20]. Recently, the demand for telerehabilitation in pulmonary rehabilitation is increasing, owing to advances in telemedicine and challenges with face-to-face rehabilitation during the COVID-19 pandemic [20-22]. Among telerehabilitation modalities, mobile app–based pulmonary rehabilitation has been used in clinical trials; however, the clinical evidence for mobile app–based pulmonary rehabilitation from these studies has been inconclusive due to the heterogeneity in participants, study designs, and formats of apps [23-32]. Furthermore, previous systematic reviews focused on telerehabilitation [20], home telemonitoring [33], or patient support apps [34]. Therefore, we aimed to compare the clinical outcomes of mobile app–based, self-directed pulmonary rehabilitation programs (ie, those without telemonitoring but with exercise programs) in patients with COPD to those of conventional pulmonary rehabilitation because exercise programs are key components of pulmonary rehabilitation that improve chronic respiratory diseases and health-enhancing behaviors [12].

Methods

Data Sources and Literature Search

Literature searches were performed by using the PubMed, Embase, Cochrane, and CINAHL databases. The searches were conducted for literature published since 2007 because the iPhone (Apple Inc) and Android (Google LLC) smartphones were released in June 2007 and September 2008, respectively. The databases were searched for literature published up to June 30, 2023. Only full-text studies written in English were included. The search strategy was based on a PICOTS-SD (population, intervention, comparison, outcomes, time, setting, and study design) list (Multimedia Appendix 1). Briefly, the search algorithm focused on keywords related to “chronic pulmonary disease,” “mobile application,” and various clinical outcomes. If needed, authors were contacted for further information.

Eligibility Criteria and Study Selection

Each study was reviewed by 2 authors (CC and MWJ) independently according to the inclusion and exclusion criteria. The inclusion and exclusion criteria are presented in Table 1. The screening of titles and abstracts and the subsequent full-text review were performed by 2 authors (CC and MWJ) independently. Disagreements during the selection process were resolved through a discussion between 3 authors (CC, MWJ, and SWL).
### Inclusion and exclusion criteria.

<table>
<thead>
<tr>
<th>Inclusion criteria</th>
<th>Exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article type</td>
<td>Full-text articles</td>
</tr>
<tr>
<td>Language</td>
<td>English</td>
</tr>
<tr>
<td>Study design</td>
<td>Randomized controlled trials</td>
</tr>
<tr>
<td>Participants’ age</td>
<td>Adults</td>
</tr>
<tr>
<td>Disease</td>
<td>COPD</td>
</tr>
<tr>
<td>Smartphone app</td>
<td>Conventional or newly developed smartphone apps</td>
</tr>
<tr>
<td>Intervention</td>
<td>Pulmonary rehabilitation, including exercise programs, provided by a smartphone app</td>
</tr>
<tr>
<td>Control</td>
<td>Conventional pulmonary rehabilitation, including exercise programs (center-based rehabilitation or education)</td>
</tr>
<tr>
<td>Study outcome</td>
<td>At least 1 of the following outcomes: 6-minute walk test distance, COPD Assessment Test score, modified Medical Research Council dyspnea scale score, St. George Respiratory Questionnaire score, and hospitalization resulting from disease exacerbation</td>
</tr>
</tbody>
</table>

**Note:**
- COPD: chronic obstructive pulmonary disease.
- N/A: not applicable.

### Data Collection and Risk of Bias Assessment

Two authors (CC and MWJ) independently collected data regarding (1) general information about the study (authors, year, country, and study setting), (2) descriptions of study arms (number, sex, and age of participants), (3) characteristics of interventions, (4) inclusion and exclusion criteria, and (5) results for outcomes; they also double-checked these data. Two authors (CC and MWJ) independently assessed the risk of bias in the included studies. Discrepancies were resolved in discussions with the third author (SWL).

### Study Outcomes

In the meta-analysis, study outcomes, including exercise capacity, symptom scores, quality of life, and hospitalization, were assessed. Exercise capacity was measured by using 6MWDs. The symptom scores were measured by using the COPD Assessment Test (CAT) and the mMRC dyspnea scale. Quality of life was measured by using the St. George Respiratory Questionnaire (SGRQ). Hospitalization was defined as hospitalizations resulting from disease exacerbation. The primary time points for the analysis were baseline and the end of the intervention.

### Statistical Analysis

The continuous variables included the 6MWD, CAT score, and SGRQ score. The mMRC dyspnea scale score was a categorical variable, and it was calculated as a continuous value. Hospitalization resulting from disease exacerbation was a dichotomous variable. The variables at the time of follow-up were compared between groups. The mean differences and risk ratios between the intervention group and the control group were calculated, along with 95% CIs. The chi-square test and the $I^2$ statistic were used to assess statistical heterogeneity. If $I^2$ was <50%, the fixed effect model was used. Publication bias was visually assessed by using a funnel plot analysis because the limited number of studies with results for each outcome prevented us from performing the Egger test. The meta-analysis was performed by using Review Manager (RevMan) version 5.4 (The Cochrane Collaboration).

### Ethical Considerations

This study complied with the Declaration of Helsinki, and all methods were performed in accordance with the relevant guidelines.

### Results

#### Study Selection

An initial literature search identified a total of 1851 articles from the PubMed, Embase, Cochrane, and CINAHL databases; thereafter, 1173 articles remained after duplicates were removed. After evaluating titles and abstracts, 299 articles remained eligible for a full-text review. The full-text review was performed according to the criteria mentioned in the Eligibility Criteria and Study Selection section, and 10 articles were finally included in the systematic review [23-32]. Notably, 1 study was excluded from the meta-analysis because exercise capacity was evaluated by using the incremental shuttle walk test (ISWT) instead of 6MWDs [32]. Therefore, 9 studies were included in the meta-analysis [23-31] (Figure 1).
Characteristics of Included Studies

Characteristics of studies are described in Table 2. Studies were published after 2014, with almost half of them (4/10, 40%) published in 2020 [24,26,28,29]. Further, 6 studies were multicenter studies [24,25,27,29-31], and 3 studies enrolled fewer than 50 participants; the largest number of participants was 343 [25,28,29,32]. There were 1050 total participants, who were generally aged ≥65 years. More male participants were enrolled than female participants, and Wang et al [32] enrolled only male participants. In the study by North et al [28], participants were recruited after hospital admission with an acute exacerbation. In the studies by Vorrink et al [31] and Wang et al [32], participants were recruited after pulmonary rehabilitation. Kwon et al [27] recruited 2 groups of participants in the intervention arm, comprising the fixed regimen group and the fixed-interactive regimen group, according to exercise programs. Various formats of mobile apps were used for the studies; 2 studies in the United Kingdom used myCOPD, a digital health care app approved by the National Health Service [24,28], and 1 study in China used WeChat (Tencent Holdings Ltd), a popular mobile messenger app in China [26]. The follow-up duration ranged between 3 weeks and 12 months [23,31].
### Characteristics of included studies.

<table>
<thead>
<tr>
<th>Study (author, year)</th>
<th>Setting</th>
<th>Country</th>
<th>Sample size, n (%)</th>
<th>Age (y), mean (SD)</th>
<th>Mobile app</th>
<th>Follow-up duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barata et al [23], 2022</td>
<td>Single center</td>
<td>Romania</td>
<td>Male: 42 (72.4); female: 16 (27.6)</td>
<td>64.9 (5.7)</td>
<td>Control</td>
<td>21 d</td>
</tr>
<tr>
<td>Crooks et al [24], 2020</td>
<td>Multicenter</td>
<td>United Kingdom</td>
<td>Male: 11 (37.9); female: 18 (62.1)</td>
<td>65.9 (7.3)</td>
<td>Intervention</td>
<td>90 d</td>
</tr>
<tr>
<td>Demeyer et al [25], 2017</td>
<td>Multicenter</td>
<td>Belgium</td>
<td>Male: 111 (64.9); female: 60 (35.1)</td>
<td>64 (8)</td>
<td>Control</td>
<td>12 wk</td>
</tr>
<tr>
<td>Jiang et al [26], 2020</td>
<td>Single center</td>
<td>China</td>
<td>Male: 44 (83); female: 9 (17)</td>
<td>70.9 (6.4)</td>
<td>Intervention</td>
<td>6 mo</td>
</tr>
<tr>
<td>Kwon et al [27], 2018</td>
<td>Multicenter</td>
<td>Republic of Korea</td>
<td>Male(^a): 23 (85.2); female(^a): 4 (14.8); male(^b): 26 (86.7); female(^b): 4 (13.3)</td>
<td>64 (8)(^b); 65 (7)(^b)</td>
<td>Intervention</td>
<td>12 wk</td>
</tr>
<tr>
<td>North et al(^c) [28], 2020</td>
<td>Single center</td>
<td>United Kingdom</td>
<td>Male: 13 (65); female: 7 (35)</td>
<td>65.1 (6.3)</td>
<td>Control</td>
<td>90 d</td>
</tr>
<tr>
<td>Park et al [29], 2020</td>
<td>Multicenter</td>
<td>Republic of Korea</td>
<td>Male: 19 (86.4); female: 3 (13.6)</td>
<td>70.5 (9.4)</td>
<td>Intervention</td>
<td>6 mo</td>
</tr>
<tr>
<td>Spielmanns et al [30], 2023</td>
<td>Multicenter</td>
<td>Switzerland</td>
<td>Male: 17 (51.5); female: 16 (48.5)</td>
<td>66.1 (6.8)</td>
<td>Control</td>
<td>6 mo</td>
</tr>
<tr>
<td>Vorrink et al(^e) [31], 2016</td>
<td>Multicenter</td>
<td>Netherlands</td>
<td>Male: 42 (50); female: 42 (50)</td>
<td>62 (9)</td>
<td>Intervention</td>
<td>12 mo</td>
</tr>
<tr>
<td>Wang et al(^e) [32], 2014</td>
<td>Single center</td>
<td>Taiwan</td>
<td>Male: 12 (100)</td>
<td>71.4 (1.9)</td>
<td>Control</td>
<td>6 mo</td>
</tr>
</tbody>
</table>

\(^a\)The fixed regimen group.  
\(^b\)The fixed-interactive regimen group.  
\(^c\)Participants were recruited after hospital admission with an acute exacerbation.  
\(^d\)COPD: chronic obstructive pulmonary disease.  
\(^e\)Participants were recruited after pulmonary rehabilitation.

The interventions in the studies are described in Table 3. Disease education and monitoring were provided in 5 studies [24,26,28-30,35], and the other 5 studies provided only exercise programs [23,25,27,31,32]. The level of exercise could be adjusted according to the participants’ exercise capacity in 5 studies [23,25,27,31,32]. In particular, Kwon et al [27] provided 2 kinds of exercise regimens, and walking distances were adjustable in both regimens. In cases of COPD exacerbation or
poor compliance to pulmonary rehabilitation, participants could contact health care professionals in 7 studies [24-26,28-30,32,35]. Jiang et al [26] gave incentives to participants, that is, participants could obtain gifts at a mall by using acquired points.

Table. Interventions of included studies.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Exercise adjustment</th>
<th>Exercise monitoring</th>
<th>Disease education</th>
<th>Disease monitoring</th>
<th>Social support</th>
<th>Contact with health care professionals</th>
<th>Incentive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barata et al [23]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Crooks et al [24]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Demeyer et al [25]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Jiang et al [26]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Kwon et al [27]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>North et al [28]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Park et al [29]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Spielmanns et al [30]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Vorrink et al [31]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Wang et al [32]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

Most studies (7/10, 70%) included adult participants with physician-diagnosed COPD; diagnoses were made according to the GOLD (Global Initiative for Chronic Obstructive Lung Disease) criteria [1]. Some studies did not include participants with severe COPD as defined by the GOLD criteria [29,31], and others did not set limitations for disease severity. Generally, participants with recent acute exacerbations, participants undergoing long-term home oxygen therapy, or participants with other medical conditions that did not allow for physical exercise were excluded. In the study by North et al [28], participants were included after hospitalization with an acute exacerbation (Table S1 in Multimedia Appendix 2).

Participants were evaluated on various dimensions of outcomes, including exercise capacity, disease severity, quality of life questionnaires, and acute exacerbation. Wang et al [32] reported favorable exercise capacity and serum inflammatory biomarker outcomes; however, this study was excluded from the meta-analysis because exercise capacity was reported based on the ISWT and limb muscle strength. Crooks et al [24] and North et al [28] reported that inhaler technique improved in the intervention group, which was beneficial to disease control. Demeyer et al [25] reported that lung function did not improve during pulmonary rehabilitation in the intervention and control groups, and musculoskeletal events occurred more often in the intervention group. Barata et al [23] reported that the maximal inspiratory and expiratory pressures improved in the intervention group (Table S2 in Multimedia Appendix 2).

Risk of Bias in Studies

The overall risk of bias in studies was considered low. However, the risk of performance bias was inevitably considered high in all studies because participant blinding was impossible, owing to the nature of the intervention (Figure 2). Funnel plots of comparisons showed fairly symmetrical distributions, which might mean less publication bias (Figure S1 in Multimedia Appendix 3).
Figure 2. Risk of bias in the included studies [23-31]. A: Risk of bias graph. B: Risk of bias summary; “Kwon H 2018 (1)” denotes the fixed regimen group, and “Kwon H 2018 (2)” denotes the fixed-interactive regimen group.

Meta-Analysis of Clinical Outcomes

Figure 3 shows the meta-analysis of study outcomes. In terms of statistical heterogeneity, the chi-square test and $I^2$ statistic for each meta-analysis showed no important heterogeneity. Exercise capacity was reported in various forms, including 6MWDs, ISWT results, the number of steps per day, and metabolic equivalents, in 8 studies [23,25,27,29-32]. Wang et al [32] reported on the ISWT only, and Crooks et al [24] and Spielmanns et al [30] reported the number of steps per day only. Thus, the 6MWD, which was used in 5 studies, was included in the meta-analysis [23,25,27,29,31]; there was no statistically significant difference between groups (mean difference 9.52, 95% CI −3.05 to 22.08 m; $P=.14$).

CAT scores were reported in 7 studies [23-28,30]; however, Demeyer et al [25] reported the CAT scores as medians and IQRs. Thus, the CAT scores from 6 studies were analyzed [23,24,26-28,30]. The CAT scores of the intervention group were significantly lower than those of the control group (mean difference −1.29, 95% CI −2.39 to −0.20; P=.02). Dyspnea was measured by using the mMRC dyspnea scale in 3 studies [26-28], and the scores did not significantly differ between groups (mean difference −0.08, 95% CI −0.29 to 0.13; P=.45). The quality of life was assessed in 6 studies, using various questionnaires [24-26,28,29,31], and SGRQ scores were reported in 2 trials [26,28]; there was no statistical difference in these scores between groups (mean difference −3.62, 95% CI −9.62 to 2.38; P=.24).

The exacerbation of COPD was reported as outpatient clinic visits, emergency room visits, or hospitalizations in 4 studies [24,25,28,29]. Among them, hospitalizations were reported in 3 studies [24,28,29]. The frequency of hospitalization was not statistically different between groups (risk ratio 0.65, 95% CI 0.27 to 1.53; P=.32).

We also performed a subgroup analysis for the 6MWDs and CAT scores based on the baseline study results (6MWDs: ≥400 m vs <400 m; CAT scores: ≥20 vs <20) [23-31]. The subgroup
analysis did not show statistically significant differences (all $P$ values were >0.05). Furthermore, we performed a subgroup analysis for the CAT scores based on the rehabilitation programs (exercise program only vs exercise and self-management programs) [23,24,26-28,30]. Among studies offering both exercise and self-management programs, the CAT scores of the intervention group were significantly lower than those of the control group (mean difference $-2.16$, 95% CI $-3.93$ to $-0.39$; $P=0.02$; Figure S2 in Multimedia Appendix 3 [23-31]).

**Discussion**

**Principal Results and Implications**

We reviewed and described the clinical outcomes of mobile app–based pulmonary rehabilitation in patients with COPD. Participants and interventions were heterogeneous in their characteristics; however, participants who underwent mobile app–based pulmonary rehabilitation showed favorable exercise capacity, symptom score, quality of life, and hospitalization outcomes when compared to participants who underwent conventional pulmonary rehabilitation. In the meta-analysis, the 6MWDs, mMRC dyspnea scale scores, SGRQ scores, and exacerbations in the mobile app–based pulmonary rehabilitation group were not inferior to those in the control group, and the CAT scores were superior to those in the control group. Considering the difficulties in practicing conventional center-based pulmonary rehabilitation, mobile app–based pulmonary rehabilitation may be a useful treatment option when conventional pulmonary rehabilitation is not feasible.

**Mobile App–Based Pulmonary Rehabilitation**

Pulmonary rehabilitation has been traditionally delivered in outpatient, inpatient, and community settings, comprising ≥2 sessions per week for at least 4 weeks [14]. In 2015, the American Thoracic Society/European Respiratory Society policy statement requested researchers to adopt alternative formats for pulmonary rehabilitation, demonstrate clinical outcomes that are at least comparable to those of traditional pulmonary rehabilitation programs, and evaluate cost-effectiveness and safety [36]. Since then, clinical trials have reported data on the clinical outcomes and safety of pulmonary rehabilitation program models, including home-based rehabilitation, telerehabilitation, web-based rehabilitation, community rehabilitation, primary care rehabilitation, rehabilitation requiring minimal resources, and combined heart failure/pulmonary rehabilitation models [22]. Mobile app–based pulmonary rehabilitation can be regarded as a type of telehealth intervention [20] that provides health care at a distance through telecommunications or web-based technologies [37]. It may improve the accessibility of pulmonary rehabilitation for patients with chronic respiratory diseases by providing health care access and services for patients who are geographically or socially isolated, are engaged with full-time work, or are hard to transport due to the disease or comorbidities [20].

**Further Development of Apps**

Various types of apps were used in the studies. Some authors used newly developed apps, and others used myCOPD or the social messenger app WeChat [24,26,28]. Some apps, such as myCOPD, provided self-management programs for COPD, including education and symptom management programs [24,28]; however, other apps provided only exercise programs [25,27]. Although this study focused on clinical improvements in participants who underwent pulmonary rehabilitation, it should also be considered that overall self-management programs, such as disease education and symptom management programs, have affected clinical outcomes. However, pulmonary rehabilitation is defined as a comprehensive intervention that includes exercise training, education, and behavior change [12]. Recently, Holland et al [22] suggested that desirable components of pulmonary rehabilitation should include education, self-management training, smoking cessation, and an action plan for exacerbation, as well as a home exercise program. Therefore, apps that provide both exercise programs and self-management programs should be included in mobile app–based pulmonary rehabilitation.

Considering the challenges in center-based pulmonary rehabilitation and the shortage of health care resources, home-based pulmonary rehabilitation has been studied as an alternative to center-based pulmonary rehabilitation [38-43]. However, compliance to pulmonary rehabilitation is an important issue in home-based pulmonary rehabilitation, and a lack of motivation is an important reason for poor compliance [44]. In a study of home-based pulmonary rehabilitation without supervision, patients with good compliance showed significant improvements in CAT scores, BODE index scores, and FEV1 values when compared to patients with poor compliance [45]. Similarly, Crooks et al [24] described that there was an estimated $-0.22$ (95% CI $-0.74$ to $-0.31$) decrease in the CAT score for every 7-day increase in app use (adjusted for baseline CAT score, COPD severity, and study site). However, North et al [28] reported that as time passed, the number of app users decreased in mobile app–based pulmonary rehabilitation. Therefore, patients are required to steadily run the app and perform pulmonary rehabilitation to achieve clinical improvement. Various methods were used in studies to enhance compliance, such as sending text messages with activity proposals to participants, contacting participants via telephone, providing incentives, and having participants communicate with other participants [25,26,29,31,32]. Additionally, activity level (step counts) was monitored by using a pedometer, and feedback was provided to participants [25,29,31,32]. In real-world practice, health care interventions and action plans should be considered in cases of poor compliance because poor compliance might reflect deconditioning or acute exacerbation among patients [22,26,44].

**Further Development of Rehabilitation Programs**

In clinical practice, exercise levels in pulmonary rehabilitation should be individualized according to patients’ exercise capacity [12,13]. Therefore, in mobile app–based pulmonary rehabilitation, maintaining appropriate exercise levels is a matter of concern. Some apps provided adjustable exercise regimens according to the changes in participants’ exercise capacity [25,27,31,32]. Kwon et al [27] designed exercise regimens in which the exercise levels were adjusted according to the maximum walking speed in the 6-minute walk test and the degree of breathing difficulty after exercise. Vorrink et al [31]...
designed physical activity goals that were set according to average steps per day. To maintain appropriate exercise levels, apps should provide adjustable and individualized exercise programs based on patients’ exercise capacity and activity level data that are collected via wearable devices or smartphone-mounted sensors.

Considering the study designs included in this review, it is important to develop strategies for improving compliance to rehabilitation and design individualized exercise programs to achieve significant improvements in clinical outcomes in future studies. Moreover, most studies (6/10, 60%) had rather small sample sizes (<100 individuals) for demonstrating the efficacy of pulmonary rehabilitation programs [24,27-30,32]. In addition, most studies (8/10, 80%) did not provide data regarding app usage, which could have been used in the subgroup analysis related to compliance [23,25-27,29-32]. Therefore, further studies with larger sample sizes and data on app usage are needed.

Nutrition support is also an important part of pulmonary rehabilitation [13,22]. In this review, some of the included apps provided disease education; however, a nutrition support program was not provided [24,26,28-30,35]. Nutrition support may help patients with COPD to maintain an adequate BMI and increase their muscle mass [13,22]. Exercise training that is accompanied by nutrition support might improve respiratory sarcopenia and enhance clinical benefits [15]; thus, further studies are needed in this area.

Clinical Outcomes and Prognosis

Exercise capacity and physical activity data can be used to predict the prognosis of patients with COPD. Exercise capacity inversely correlates with mortality in patients with COPD [46]. Physical activity also inversely correlates with exacerbation and mortality in patients with COPD [47]. Some of the reviewed studies reported physical activity as daily step counts, and these data were too widely distributed to be synthesized in the meta-analysis [24,30]. Moreover, the 6MWDs were not significantly different in the meta-analysis (P=14), and Wang et al [32] reported improvements in the ISWT and limb muscle mass in the intervention group. Thus, further studies are required to ascertain whether mobile app–based pulmonary rehabilitation can improve exercise capacity and physical activity in patients with COPD.

In some studies, we noticed that mobile app–based pulmonary rehabilitation improved quality of life, including the SGRQ, Clinical COPD Questionnaire, and Chronic Respiratory Disease Questionnaire scores [24-26,28,29,31]. Among these, the CAT scores significantly improved in the intervention group, as per the meta-analysis (P=0.02) [23,24,26-28,30]. The CAT scores correlated with the severity of airflow limitation and disease exacerbation in patients with COPD [48,49]. Taken together, mobile app–based pulmonary rehabilitation programs might improve clinical outcomes, such as acute exacerbation and mortality. Unfortunately, in the meta-analysis, there was no statistically significant difference in acute exacerbations between groups (P=.32) because the study periods (range 3-6 mo) might have been too short to observe acute exacerbations [24,28,29]. Therefore, further studies with long-term follow-ups are required to evaluate the effect of mobile app–based pulmonary rehabilitation on acute exacerbations and mortality.

Limitations

First, discrepancies in the baseline status of participants were one of the main obstacles in synthesizing clinical outcomes. In the study by North et al [28], participants were evaluated after hospitalization with an acute exacerbation. In the studies by Vorrink et al [31] and Wang et al [32], physical activity in participants with COPD was evaluated after pulmonary rehabilitation. Despite this heterogeneity, participants who underwent mobile app–based pulmonary rehabilitation showed consistently favorable results for clinical parameters. Second, discrepancies in the clinical parameters were also an obstacle in synthesizing clinical outcomes. Among the various parameters for exercise capacity, a meta-analysis could be performed on the 6MWD, as it was used in half of the reviewed studies (5/10, 50%) [23,25,27,29,31], and the 6MWD is a well-established surrogate marker in patients with COPD [1,50]. Questionnaires about quality of life, including the SGRQ, EQ-5D-5L, Clinical COPD Questionnaire, and Chronic Respiratory Disease Questionnaire, also showed generally favorable results in patients who underwent mobile app–based pulmonary rehabilitation [24-26,28,29,31]. Although clinical outcomes did not reflect statistically significant improvement in participants who underwent mobile app–based pulmonary rehabilitation and decisive evidence was hard to derive, this study showed that clinical outcomes generally favored mobile app–based pulmonary rehabilitation. Considering the difficulties with center-based pulmonary rehabilitation in real-world practice, mobile app–based pulmonary rehabilitation could be a reasonable alternative to conventional pulmonary rehabilitation.

Conclusion

In conclusion, this review shows that many mobile apps have been applied to pulmonary rehabilitation for patients with COPD. There were discrepancies in the baseline participant characteristics and interventions among studies. Nevertheless, in some studies, patients who participated in mobile app–based pulmonary rehabilitation showed favorable exercise capacity, symptom score, quality of life, and hospitalization outcomes when compared with those who underwent conventional pulmonary rehabilitation. In the meta-analysis, the 6MWDs, mMRC dyspnea scale scores, SGRQ scores, and exacerbations in the mobile app–based pulmonary rehabilitation group were not inferior to those in the control group, and the CAT scores were superior to those in the control group. Therefore, in real-world practice, mobile app–based pulmonary rehabilitation can be a useful treatment option when conventional center-based pulmonary rehabilitation is not feasible.
Acknowledgments
This research was supported by a grant of the Korea Health Promotion R&D Project, funded by the Ministry of Health & Welfare, Republic of Korea (grant HS21C0096), and a grant from the Korea Health Technology R&D Project, provided through the Korea Health Industry Development Institute (KHIDI) and funded by the Ministry of Health & Welfare, Republic of Korea (grant HI20C1058). We thank Asan Medical Library at University of Ulsan College of Medicine for helping with the literature search.

Data Availability
The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions
CC, JWL, SWL, and MWJ were responsible for the concept and design of the study. CC, SWL, and MWJ were responsible for data acquisition, data curation, and formal analysis. CC, JWL, and SWL drafted the original manuscript. All authors have revised and approved the final manuscript. All authors take responsibility for the accuracy of the content of the final manuscript. JWL and SWL obtained funding. Authors SWL and MWJ are co-corresponding authors. MWJ can be contacted at mdjominwoo@gmail.com. Generative artificial intelligence was not used in any portion of manuscript writing.

Conflicts of Interest
None declared.

Multimedia Appendix 1
PICOTS-SD (population, intervention, comparison, outcomes, time, setting, and study design) search strategy for mobile apps for patients with chronic pulmonary disease.
[DOCX File, 20 KB - mhealth_v12i1e41753_app1.docx ]

Multimedia Appendix 2
Supplementary tables for the inclusion criteria, exclusion criteria, and clinical outcomes of the included studies.
[DOCX File, 54 KB - mhealth_v12i1e41753_app2.docx ]

Multimedia Appendix 3
Supplementary figures for the outcomes of included studies.
[DOCX File, 107 KB - mhealth_v12i1e41753_app3.docx ]

Checklist 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File, 32 KB - mhealth_v12i1e41753_app4.docx ]

References


Application of eHealth Tools in Anticoagulation Management After Cardiac Valve Replacement: Scoping Review Coupled With Bibliometric Analysis

Ying Wu1,2*, MSc; Xiaohui Wang2*, MSc; Mengyao Zhou2, BSc; Zhuoer Huang2, MSc; Lijuan Liu3, MSc; Li Cong2, PhD

1Center for Moral Culture, Hunan Normal University, Changsha, China
2School of Medicine, Hunan Normal University, Changsha, China
3Teaching and Research Section of Clinical Nursing, Xiangya Hospital of Central South University, Changsha, China
*these authors contributed equally

Corresponding Author:
Li Cong, PhD
School of Medicine
Hunan Normal University
371 Tongzipo Road
Changsha, 410013 China
Phone: 86 0731 889124
Email: congli@hunnu.edu.cn

Abstract

Background: Anticoagulation management can effectively prevent complications in patients undergoing cardiac valve replacement (CVR). The emergence of eHealth tools provides new prospects for the management of long-term anticoagulants. However, there is no comprehensive summary of the application of eHealth tools in anticoagulation management after CVR.

Objective: Our objective is to clarify the current state, trends, benefits, and challenges of using eHealth tools in the anticoagulation management of patients after CVR and provide future directions and recommendations for development in this field.

Methods: This scoping review follows the 5-step framework developed by Arksey and O’Malley. We searched 5 databases such as PubMed, MEDLINE, Web of Science, CINAHL, and Embase using keywords such as “eHealth,” “anticoagulation,” and “valve replacement.” We included papers on the practical application of eHealth tools and excluded papers describing the underlying mechanisms for developing eHealth tools. The search time ranged from the database inception to March 1, 2023. The study findings were reported according to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews). Additionally, VOSviewer (version 1.6.18) was used to construct visualization maps of countries, institutions, authors, and keywords to investigate the internal relations of included literature and to explore research hotspots and frontiers.

Results: This study included 25 studies that fulfilled the criteria. There were 27,050 participants in total, with the sample size of the included studies ranging from 49 to 13,219. The eHealth tools mainly include computer-based support systems, electronic health records, telemedicine platforms, and mobile apps. Compared to traditional anticoagulation management, eHealth tools can improve time in therapeutic range and life satisfaction. However, there is no significant impact observed in terms of economic benefits and anticoagulation-related complications. Bibliometric analysis suggests the potential for increased collaboration and opportunities among countries and academic institutions. Italy had the widest cooperative relationships. Machine learning and artificial intelligence are the popular research directions in anticoagulation management.

Conclusions: eHealth tools exhibit promise for clinical applications in anticoagulation management after CVR, with the potential to enhance postoperative rehabilitation. Further high-quality research is needed to explore the economic benefits of eHealth tools in long-term anticoagulant therapy and the potential to reduce the occurrence of adverse events.

(JMIR Mhealth Uhealth 2024;12:e48716) doi:10.2196/48716
KEYWORDS

eHealth tool; cardiac valve replacement; anticoagulation management; scoping review; bibliometrics analysis; rehabilitation

Introduction

Valvular heart disease involves damage to the cardiac valves caused by various factors such as valve degeneration and rheumatic heart disease [1]. Currently, an estimated 209 million people suffer from valvular heart disease worldwide. With the aging population, it is expected that the prevalence of valvular heart disease among those aged 65 years and older will increase by 50% globally by 2030 [2,3]. Furthermore, valvular heart disease caused approximately 300,000 deaths in 2019, accounting for nearly 2% of cardiovascular disease deaths worldwide. Every year, 8.7 million life years and 10.7 million disability-adjusted life years are lost to valvular heart disease [2]. The fundamental treatment for valvular heart disease is the repair or replacement of damaged heart valves through surgery or interventional therapy. Patients need long-term oral anticoagulants after cardiac valve replacement (CVR) surgery; however, improper use of anticoagulants can lead to bleeding, embolism, eventual valve failure, repeat surgeries, and even death [4,5]. Efficient anticoagulation management (AM) plays a critical role in promoting safe medication therapy after CVR, reducing adverse events, and improving the long-term prognosis of patients.

The traditional AM model has many obstacles and challenges. On the one hand, patients need to travel between their home and hospital regularly for follow-up and laboratory examinations and long-term anticoagulant medication, which costs them a lot of time, energy, and economy after CVR and also reduces patients’ confidence in rehabilitation and treatment compliance [6]. On the other hand, it is difficult for doctors to monitor and manage patients and adjust treatment plans promptly, which may affect the quality and effect of anticoagulant therapy, and increase the risk of adverse events [7]. Therefore, the exploration of personalized, intelligent, and efficient AM models is crucial in promoting cardiac recovery among patients undergoing CVR. Several guidelines highlight the potential of eHealth tools to facilitate AM and improve medical outcomes [8] and recommend the use of eHealth tools to assist physicians with anticoagulation decision-making support [9].

eHealth tools generally refer to digital devices based on information and communication technology, which can be used for disease prevention, diagnosis, treatment, monitoring, and management [10]. eHealth tools can provide patients with convenient, efficient, and accurate medical services through digital communication and remote disease management. Some studies have found that eHealth tools may become a potentially cost-effective and effective alternative to traditional “face-to-face” anticoagulation therapy. A meta-analysis of 12 randomized controlled trials (RCTs) showed that telemedicine combined with portable coagulometers significantly improved the time in the therapeutic range (TTR) and reduced the incidence of thrombotic events in patients with cardiovascular disease [11]. Inpatients undergoing anticoagulation therapy receiving the recommended dosage of anticoagulation software can effectively reduce readmissions and 30-day mortality as well as hospitalization costs [12]. A comprehensive understanding of the application status and effectiveness of eHealth tools is essential to establish a scientific, traceable, and integrated AM model.

Although several narrative reviews and meta-analyses have summarized the usage of mobile health technology in patients with cardiovascular diseases [13-15], uncertainties remain regarding the effectiveness and current status of using eHealth tools in AM of patients after CVR. A comprehensive summary of the development, application, and interrelationships among different research papers or groups is lacking. Scoping reviews entail extensive searches and a rigorous selection of research literature on a specific topic, which can include a comprehensive overview of the current research status and trends of the topic [16]. Scientometric analysis can quantitatively evaluate cooperative relationships by statistically analyzing publications and graphically presenting the social and intellectual connections of relevant literature [17]. Through the joint analysis of scoping review and scientometric analysis, the research status and trends of a certain field can be comprehensively evaluated from different perspectives, and the blind spots and unsolved problems can be determined to provide guidance and enlightenment for further research.

By conducting a scoping review and scientometric analysis, this study aimed to (1) summarize the application status of eHealth tools in AM of patients after CVR, (2) identify the hotspots and provide guidance for future research and practice, and (3) to provide a reference for promoting the wider application and sustainable development of eHealth tools in AM of patients after CVR.

Methods

Design

We used Arksey and O’Malley’s [18] 5-step framework for the scoping review. This review also followed the recommended items in the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist (Multimedia Appendix 1). Five commonly used databases were searched, and the papers were screened based on the title, abstract, and full text. We exported the complete records of filtered papers as plain text files and imported them into VOSviewer software (version 1.6.18; Centre for Science and Technology Studies) to build visualization maps. We chose coauthorship analysis to construct network visualization maps of countries, institutions, and authors to understand the status of research collaboration in the field of eHealth tools. By cluster analysis and keyword overlay visualization, keywords are divided into different clusters and stacked over time, which can identify different themes and current research hotspots.

Scoping Review

In contrast to systematic and narrative reviews, scoping reviews focus on an initial appraisal of the current extent, scope, and
nature of the research literature and take the dissemination process further by summarizing the relevant existing research activities. It is also an appropriate way to map the key concepts and identify knowledge gaps [19]. We aimed to provide an overview of the use of eHealth tools to assist anticoagulation therapy in patients after CVR and highlight the current status, trends, and challenges in this field. Therefore, the scoping review was appropriate for this study.

Identifying the Research Question for Scoping Review
The first question that guided our scoping review was what are the range and effectiveness of eHealth tools services in AM of patients after CVR? The second one was what are the benefits and barriers of applying eHealth tools in AM?

Identifying Databases and Studies
PubMed, MEDLINE, Web of Science, CINAHL, and Embase were searched from inception to March 1, 2023. Searches were not limited to a specific geographic region, and any literature published in non-English languages was excluded. Searches included combinations of free text words and index terms using Boolean operators. Moreover, a manual retrospective search of the references was conducted as a supplement. Detailed search strategies for each database are described in Multimedia Appendix 2.

Study Selection
The Population, Concept, and Context framework is recommended by the Joanna Briggs Institute to identify the main concepts in primary review questions, guide the search strategy, and ensure application of the inclusion and exclusion criteria [20]. Therefore, we used the Population, Concept, and Context framework to regulate the scoping review process. Population was defined as patients who had undergone CVR surgery and received anticoagulant therapy postoperatively. Concept referred to the practical application and effectiveness of eHealth tools in the AM of patients after CVR. Context had no special restrictions, as eHealth tools can be applied in the patient’s home, hospital, anticoagulant therapy clinic, primary care center, and so forth. The inclusion criteria were as follows: (1) publication types including cross-sectional studies, longitudinal studies, cohort studies, case-control studies, pilot studies, and RCTs; (2) published in full text; and (3) published from the inception of each database to March 1, 2023. The exclusion criteria were as follows: (1) non-English publications; (2) publication types including empirical research, reviews, editorials, reports, case reports, letters, and conference proceedings or papers or abstracts; (3) qualitative studies reporting user experiences about eHealth tools in AM of patients after CVR; and (4) studies describing only the potential mechanisms or development process of eHealth tools, rather than their practical applications. The literature records retrieved were imported into Note Express software to screen for duplicate papers. Two researchers (YW and XW) independently conducted the initial screening of the titles and abstracts based on the inclusion and exclusion criteria. Full texts were then examined for secondary screening of potentially eligible papers. Any disagreement during the screening process was resolved through discussion with the third researcher (LC), and the final selection of papers was determined based on the established criteria.

Data Extraction and Analysis
Two researchers (YW and XW) independently extracted data from the included studies using standardized tables and cross-checking their findings. Any discrepancies were resolved through discussion with the third researcher (LC). The information extracted included authors, year, country, study design, study objective, participants’ characteristics, study location, content elements, outcome measures, and study conclusions. We summarized the literature on the services and effects of eHealth tools in the AM of patients after CVR and presented the main concepts and findings of the literature using data charts and tables.

Bibliometric Analysis
VOSviewer is a widely used tool by researchers for bibliometric analysis, providing effective visualization and revealing connections between research materials [21]. The process of scientometric analysis in this study included 2 parts. First, coauthorship analysis was performed to explore the relationship among researchers, research institutions, and countries, which contributed to understanding the trends of scientific cooperation. We chose “coauthorship” as the analysis type, selected “countries,” “organizations,” and “authors” as the units of analysis and “full counting” as the counting method. To achieve clearer effects, the minimum number for each project was set as 1. Second, keywords network visualization and overlay visualization were analyzed to explore the hotspots and frontiers. We chose the “author keywords” in the “co-occurrence” type for analysis, the counting method was “full counting,” and the minimum cluster size to 6. We merged keywords with the same meaning and deleted redundant keywords. The minimum number was set to 1, and we obtained 40 keywords in total.

Results
Characteristics of the Included Studies
We conducted a systematic search of 5 databases, which yielded 534 studies that may be relevant to the topic. After eliminating duplicates, screening 384 titles and abstracts, and reviewing 70 full texts, we found that 25 papers met the inclusion and exclusion criteria. Figure 1 shows the selection process flowchart based on PRISMA-ScR [22]. The literature included in this study comprised 12 RCTs [23-34], 8 cohort studies [35-42], 2 pilot studies [43,44], 1 longitudinal study [45], 1 cross-sectional study [46], and 1 cross-over study [47]. The 25 papers comprised 27,050 participants in total, with the sample size for each study ranging from 49 to 13,219, the duration of intervention was 1-24 months, and the follow-up period was 1-514 months. Most of the participants were older than 40 years of age and had undergone mechanical valve replacement surgery. A table in Multimedia Appendix 3 provides a summary of the study characteristics and participant demographics.
Types of Anticoagulation Services Offered by eHealth Tools

Through the comprehensive understanding and interpretation of the included literature, we found that eHealth tools can be broadly classified into 4: computer-based support systems, electronic health records (EHRs), telemedicine platforms, and mobile apps. The application scenarios of eHealth tools are gradually expanding from large general hospitals to primary health care centers and finally to patients’ homes. Most studies have reported the beneficial effects of eHealth tools on the rehabilitation outcome and anticoagulation quality in patients undergoing CVR, reflecting the potential and application prospect of eHealth tools in this field. Table 1 shows the specific types of eHealth tools and their application scenarios. Multimedia Appendix 3 shows the specific services provided by eHealth tools and their impacts on anticoagulation outcomes.
Table 1. Types and application scenarios of eHealth tools (N=25).

<table>
<thead>
<tr>
<th>Author, Year, Country</th>
<th>eHealth tools</th>
<th>Places</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fitzmaurice et al (1996) [23], UK</td>
<td>Decision support systems (Anticoagulation Management Support System, Warwick)</td>
<td>Primary care clinic</td>
</tr>
<tr>
<td>Fitzmaurice et al (2000) [25], UK</td>
<td>Softop Information (Warwick, UK)</td>
<td>Primary care practices</td>
</tr>
<tr>
<td>Testa et al. (2006) [44], Italy</td>
<td>Electronic patient records (TaoNet, EDP-Progetti, Bolzano, Italy)</td>
<td>Peripheral health units</td>
</tr>
<tr>
<td>Poller et al (2008) [27], UK</td>
<td>PARMA5 (Instrumentation Laboratory, Milan, Italy) and DAWN AC (program 4S Dawn Clinical Software, Milntheorpe, UK)</td>
<td>Hospital</td>
</tr>
<tr>
<td>Ryan et al (2009) [28], Ireland</td>
<td>CoaCare (ZyCare Inc., Chapel Hill, NC, United States)</td>
<td>Home</td>
</tr>
<tr>
<td>Cafozza et al (2011) [39], Italy</td>
<td>Automated computer systems coagulometer (HemoSense, San Jose, CA, United States)</td>
<td>Hospital</td>
</tr>
<tr>
<td>Bussey et al (2013) [43], United States</td>
<td>ClotFree system (ClotFree; Genesis Advanced Technologies, Lakehills, TX, United States)</td>
<td>Clinical research center</td>
</tr>
<tr>
<td>Ferrando et al (2015) [40], Spain</td>
<td>SintromacWebInternet-based system (Grifols, Barcelona, Spain)</td>
<td>Home</td>
</tr>
<tr>
<td>Cao et al (2018) [37], China</td>
<td>Digital anticoagulation clinic</td>
<td>Home and hospital</td>
</tr>
<tr>
<td>Zhu et al (2021) [32], China</td>
<td>A mobile user interface medical network follow-up platform</td>
<td>Home</td>
</tr>
<tr>
<td>Cao et al (2021) [42], China</td>
<td>Alfalfa</td>
<td>Home</td>
</tr>
<tr>
<td>Alnazi et al (2022) [47], Saudi Arabia</td>
<td>WhatsApp</td>
<td>Home</td>
</tr>
<tr>
<td>Amruthlal et al (2022) [46], India</td>
<td>Smartphone app</td>
<td>Home</td>
</tr>
<tr>
<td>Ageno and Turpie (1998) [24], Canada</td>
<td>DAWN AC (Business Technology, Milntheorpe, Cumbria, the UK)</td>
<td>Hospital</td>
</tr>
<tr>
<td>Manotti et al (2001) [26], Italy</td>
<td>PARMA system (release 3.2, Parma, Italy)</td>
<td>Manotti et al (2001), Italy</td>
</tr>
<tr>
<td>O'Shea et al (2008) [35], United States</td>
<td>Management program (CoaCare, ZyCare Inc, Chapel Hill, NC, United States)</td>
<td>Home</td>
</tr>
<tr>
<td>Soliman Hamad et al (2009) [29], the Netherlands</td>
<td>Anticoagulant aid website</td>
<td>Home</td>
</tr>
<tr>
<td>Christensen et al (2011) [30], Denmark</td>
<td>Computer system (CSO/AC; IntraMed A/S, Værløse, Denmark)</td>
<td>Home</td>
</tr>
<tr>
<td>Thompson et al (2013) [33], United States</td>
<td>Automated computer systems coagulometer (HemoSense, San Jose, CA, United States)</td>
<td>Hospital</td>
</tr>
<tr>
<td>Koertke et al (2015) [34], Germany</td>
<td>The Institute for Applied Telemedicine</td>
<td>Heart centers</td>
</tr>
<tr>
<td>Talboom et al (2017) [38], the Netherlands</td>
<td>Portavita eHealth platform</td>
<td>Home</td>
</tr>
<tr>
<td>Jiang et al (2021) [41], China</td>
<td>Yixing App</td>
<td>Home</td>
</tr>
<tr>
<td>Jiang et al (2022) [36], China</td>
<td>Alfalfa</td>
<td>Home</td>
</tr>
<tr>
<td>Erba et al (2022) [45], Italy</td>
<td>PARMA GTS (Werfen, Milan, Italy) and WhatsApp</td>
<td>Anticoagulant clinic and home</td>
</tr>
</tbody>
</table>

*aPARMA: Program for Archive, Refertation, and Monitoring of Anticoagulated patients.

Of the 25 included studies, 6 used computer-based support systems [23-27,39], 4 used EHR [33,38,43,44], 9 applied telemedicine platforms [28-32,34,35,37,40], and 6 used mobile apps [36,41,45-47]. Computer-based support systems are computer applications designed to aid clinicians in making diagnostic and therapeutic decisions in patient care [48]. Such support systems are typically installed in hospitals, large clinics for anticoagulant therapy, and primary health care centers. They can help doctors predict the appropriate dose of anticoagulants and the timing of the next international normalized ratio (INR) test based on the patient’s latest INR value [23-27,39]. EHRs are generated by 1 or more interactions between medical institutions and patients [49]. Doctors use these tools to record the patients’ medication information, laboratory measurement data, clinical history, or symptoms of discomfort for reference in the next visit. They use this information to make the best decision possible and summarize their clinical experience, which provides an important basis for improving the overall medical
level. EHR can be shared between primary health care clinics and hospitals through a central database to support CVR surgical follow-up [33].

A telemedicine platform is a means of receiving medical services remotely through various communication technologies [50]. It enables remote communication and data sharing between doctors and patients, thus providing digital medical services for patients after CVR, such as web-based consultation, real-time communication, and disease management [34,37,40,43]. Patients can use portable devices to measure INR at home and transmit the results to the telemedicine platforms. Doctors can then adjust the dosage of anticoagulants and the treatment plan remotely and determine the changes in anticoagulants and conditions of patients after CVR on time [28-30,35]. Moreover, telemedicine platforms can help to supervise the patient’s self-management regime at home for a long time, helping patients correctly understand and implement treatment plans and improving treatment compliance and outcomes [28,29,31,37].

Mobile apps are services that run on smartphones, tablets, or other mobile devices. Some have been developed specifically for AM [36,41,42,46], while others are smartphone-based messaging apps [45,47]. In the research we included, mobile apps mainly mentioned the following four services: (1) Medication assistant: the system automatically generated an oral warfarin regimen based on the patient’s latest INR value and previous warfarin dose, which were reviewed by specialists and sent to patients [42,46]. At the same time, the apps set intelligent reminders every day to urge the patient to take medicines regularly [36,41]. (2) Digital consultation: patients could communicate with doctors on the internet about their condition and anticoagulation treatment [47]. (3) Push health information: apps could send videos or pictures to help patients understand disease-related information [41,42]. (4) Patients’ interactive community: the patients could communicate with other postoperative patients, share their own rehabilitation experiences, and gain support and encouragement [41].

Application Effects of eHealth Tools in AM

Comprehensive considering the evaluation indexes of anticoagulation effects on patients after CVR can provide better scientific guidance for clinicians and patients. In the literature included, the evaluation of eHealth tools in AM mainly involved three aspects. (1) Clinical outcome: This indicator was reported in 24 papers, including TTR (n=19) [23-28,30,32,33,35-40,42-44,47], rate of achieving target INR (n=5) [27,29,31,42,43], bleeding or thrombotic events (n=18) [23,25,27,28,30,32,42,44,45], and mortality (n=1) [34]. Seven RCTs have demonstrated that using eHealth tools for AM in patients after CVR is a secure and efficacious approach that significantly enhances TTR (P<.05) [23,25,26,28,30,32,33]. Computer-based support systems dosing was found to be more effective in improving TTR than medical staff–monitored dosage at the majority of centers (P<.001) [24,27,42]. Meanwhile, eHealth tools had a significantly higher number of INR within the target range compared to the conventional administration group (P<.05) [27,29]. However, several studies showed no statistical difference (P>.05) in the occurrence of bleeding or thrombotic events between the eHealth tools group and the conventional group [23,25,27,30,33]. (2) Economic benefits: Five papers focused on the health economic benefits of eHealth tools, including the frequency of INR tests (n=3) [24,33,43] and cost-effectiveness (n=2) [23,25]. Research showed that the cost of using computerized decision support software was offset by not going to the hospital [23]. However, when considering the costs associated with establishing a nurse-led clinic, the total cost was higher than traditional treatment [25]. In addition, patient self-testing at home could accelerate INR results, but it is not statistically significant in reducing INR test frequency and generating economic benefits [33,34]. (3) Patient satisfaction: The conclusion of 2 studies revealed that most patients were satisfied with the use of eHealth tools (P<.001) [23,44], which also improved their quality of life [29,44]. These findings demonstrate the potential of eHealth tools in enhancing AM. However, further studies are warranted to explore the economic benefits of long-term anticoagulant therapy and determine whether it can reduce the incidence of clinical adverse events in patients after CVR.

Scientometric Analysis

Publication Trends

The earliest literature on using eHealth tools for AM in patients after CVR was published in 1996. Figure 2 shows the trend of publications, with the number of published papers gradually increasing after 2019. The purple bar above the year label represents the publication volume, and the blue line shows cumulative publication trends. China (n=5) and Italy (n=4) stand out as the 2 countries with the most published papers, the United States and the United Kingdom have each published 3 papers, and other nations have published only 1 or 2 papers. Developed countries had research published around 2000, while transitional countries, such as China and India, only paid attention to this topic around 2019, which may be influenced by cultural differences and medical and economic levels. The majority of the 25 papers in this review were published in thrombus-related journals. However, the application of eHealth tools in the AM of patients after CVR involves not only cardiology and antithrombotic therapy but also the intersection and interdisciplinary cooperation of medical informatics, electronic technology, and health sciences. Some of the papers included in this review were also published in internet medicine journals. Therefore, when seeking the latest progress, readers should not focus only on traditional thrombosis-related journals.
Collaborative Analysis of Countries, Research Institutions, and Authors

The 25 papers selected for this review originated from 12 countries and were associated with 46 institutions. Figure 3A depicts the academic collaboration among 5 of these countries, with China being the only transitional country represented. Each node represents a country, and the connection between nodes represents the cooperative relationship. Figure 3B shows that only 8 of 46 participating institutions show close cooperative relationships, and most of them are from Italy. Each node represents an institution, and the connection between nodes represents the cooperative relationship. It should be noted that there are deficiencies in the exchanges and cooperation between transitional and high-income countries, which poses challenges in enhancing the sharing and complementarity among international academic resources. Furthermore, 89 authors participated in the publication of 25 papers. Each node represents an author, the node size depends on the number of authorial publications, and the different colors of the line represent different cooperation networks (Figure 3C). Two collaborative networks have been formed centered on Palareti G, who is affiliated with the Orsola-Malpighi Hospital in Italy and plays a crucial role in the development of eHealth tools in the AM of patients after CVR.
Figure 3. Collaborative network among countries, research institutions, and authors in publications related to the use of eHealth tools for anticoagulation management (AM) in patients after cardiac valve replacement (CVR). (A) Cooperation map among countries. (B) Collaboration network among research institutions. (C) Cooperation network of authors.

Keyword Co-Occurrence Analysis and Cluster Analysis

To present the keyword map more clearly, a cluster analysis was conducted for keywords. Figure 4A indicates that current research on the application of eHealth tools in AM after CVR is mainly focused on 4 distinct areas. The node size indicates the frequency of keyword occurrence, and different colors represent different clusters. We found that 16 keywords in the red cluster were closely related to thrombus formation and bleeding, 10 keywords in the green cluster related to the innovation of INR test techniques (home testing, self-testing, etc.), keywords in the blue cluster referred to artificial intelligence (AI) and machine learning, and yellow cluster included telemedicine, health care delivery, and remote consultation.
To further explore the hotspots and future directions, we superimposed time on keyword clustering to create keyword overlay visualization (Figure 4B). The larger the node size, the higher the frequency of keywords and the color of nodes corresponded to the average appearance time of keywords. According to high-frequency keyword analysis, anticoagulant drugs for patients undergoing CVR mainly included vitamin K antagonists led by warfarin. Additionally, the research hotspots of eHealth tools in AM have focused on AI, machine learning, and smartphone apps in recent years. Amruthal et al [46] constructed warfarin dosage prediction models by linear regression, support vector regression, logistic regression, and multilayer perceptron and found that the prediction model based on support vector regression showed the best predictive performance. Support vector regression was installed in a simple user-friendly Android mobile app. The introduction of advanced...
machine learning algorithms provides more choice and support for AM, which may help to improve the efficiency and accuracy of anticoagulant treatment.

**Discussion**

**Summary**

This review provides a full evaluation and scientometric analysis of the use of eHealth tools in the AM of patients undergoing CVR surgery. We identified and included 25 studies investigating eHealth tools, such as computer-based support systems, EHRs, telemedicine platforms, and mobile apps. The development of eHealth tools is becoming increasingly important in promoting AM, and the emergence of machine learning and AI has introduced new opportunities for innovation and promotion. Over the past decade, there has been a lack of large-sample and long-term follow-up RCT studies on AM based on eHealth tools, highlighting the need for further research to verify their roles and effects. There remain many areas for improvement to promote the application of eHealth tools in the AM of patients after CVR.

**Key Developments and Benefits Provided by eHealth Tools for AM**

The AM of patients after CVR involves a large workload, complex patient information, high technical content, and a long follow-up period. The establishment and application of eHealth tools can improve the efficiency of clinical and management work [36,41,48,49]. The computer-based support system facilitates the standardization of anticoagulant recommendations and avoids differences in the dosage prescribed by medical staff, especially in primary health care institutions lacking anticoagulation therapy experience [51]. However, previously, due to technological limitations, these systems could not closely approximate doctors’ judgment and decision-making skills, and manual verification and intervention were needed to ensure accuracy [24]. With the development of machine learning, the use of deep learning, reinforcement learning, and ensemble learning are increasing gradually in warfarin management after CVR, which can further improve the accuracy of prediction [52,53]. Additionally, doctors can adjust anticoagulant doses of patients more accurately according to their EHR, thus avoiding excessive or insufficient therapy and reducing the occurrence of complications. Although these 2 eHealth tools are highly targeted, easy to operate, and improve work efficiency, their functions are relatively simple and not enough for more comprehensive management of patients undergoing anticoagulant therapy after CVR.

Telemedicine platforms and mobile apps offer a promising solution for expanding access to high-quality medical resources by providing a digital communication channel between clinicians and patients who have undergone CVR [32-36,38,39,41,44]. Through these 2 ways, doctors can provide electronic prescriptions, digital consultations, health education, and self-management support to not only help patients better understand postoperative rehabilitation but also develop their knowledge and skills and improve the compliance, safety, and effectiveness of AM. However, telemedicine platforms depend on high-quality communication devices to be compatible with other medical devices and systems to perform remote processing of medical data and real-time monitoring of patients. In contrast, mobile apps are more portable and can be used anytime and anywhere, with a more intuitive interface and personalized services for anticoagulant therapy [39,44-47]. However, the older population, who comprise the majority of patients undergoing CVR surgery, may have difficulty learning and adapting to mobile apps and data on the internet. This may affect the accuracy of communication and information transfer between doctors and patients and limit postoperative follow-up. Additionally, practical limitations, such as network instability, data security, low degree of automation, and the need for medical institution support, could affect the effectiveness of telemedicine platforms and mobile apps.

**Research on Hotspots of eHealth Tools Based on Scientometric Analysis**

Research on the use of eHealth tools for AM after CVR surgery is lacking on a global scale. Developed nations have conducted more studies in this area, which may be attributed to their advanced economic status, superior medical quality, technical proficiency, and greater patient acceptance [54,55]. To promote the development of eHealth tools, it is suggested that countries provide a platform for exchanging and sharing the latest research results and invite scholars from different fields to participate in medical exchange programs, visiting scholar programs, medical conferences, and seminars. Moreover, universities can encourage students to effectively establish a cooperative culture, organize interdisciplinary research groups, conduct interdisciplinary training and education, and cultivate medical talents with an interdisciplinary background and international vision.

Based on the results of keyword clustering and overlay visualization, AI and machine learning have been adopted as research methods for developing eHealth tools and become a hotspot in recent years. These methods have provided valuable insights for future research. Genetic algorithms, backpropagation neural networks, and adapted neural-fuzzy inference system models have the potential to establish more accurate and stable prediction models of warfarin individual maintenance dose for patients after CVR [52,56]. Medical professionals can use biometric technology to visually identify the patient, the medication, and the confirmed ingestion, thereby reducing the risk of noncompliance in patients’ anticoagulant therapy [57]. To promote the wide application and sustainable development of eHealth tools, it is necessary to research intelligent auxiliary tools, combining sophisticated machine learning techniques and AI models to enhance the precision and dependability of prediction and innovate anticoagulant therapy tools.

**Future Directions for eHealth Tools in AM of Patients After CVR**

In 25 studies included, although the main outcome indicators of eHealth tools in AM covered most of the clinical indicators, they could not fully reflect the overall situation of patients from the perspectives of economy, society, psychology, and user experience. Therefore, it is suggested that improvements can be made in the following aspects: (1) Patient interaction indicators: Researchers can monitor the frequency of patients...
logging in or using eHealth tools in the background, learn about their experience and needs, and then adjust the tool design and functions accordingly. For example, a user interface with voice support, large font, and novice guidance can help patients overcome the potential obstacles in using eHealth tools [58].

(2) Indicators of acceptability: Researchers should evaluate the acceptance of patients and health care providers using patient-physician satisfaction, use rate, and patient-medications compliance. (3) Anticoagulant dosage index: This can be used to evaluate the accuracy and consistency of manual administration and eHealth tool–assisted administration to ensure the correctness of AM decisions. (4) Sustainability indicators: Patients can evaluate the user experience of eHealth tools from engagement, function, esthetics, and information quality to ensure long-term use and promotion [59].

In addition to the aforementioned optimization metrics, several other important factors should be considered to facilitate eHealth tools. First, it is critical to ensure that the design of eHealth tools is compatible with other medical devices and systems that the patients may use (eg, blood clotting machines and electrocardiographs). To minimize errors and improve data accuracy, these tools should connect seamlessly, allowing data to be shared and integrated across different platforms. Lubitz et al [60] used compatible, wearable devices and Android or iOS smartphones–assisted electrocardiogram patch monitoring, which effectively identified undiagnosed atrial fibrillation at an early stage. Second, providing effective training and support to older or less educated patients is critical for increasing the acceptance and use of eHealth tools. Previous studies have shown that tailored coaching and education programs can help patients operate these tools effectively [61]. Third, the use of advanced machine learning algorithms and AI models helps further improve the accuracy and reliability of eHealth tools. Zeng et al [62] used various machine learning to construct a dynamic anticoagulant treatment scheme for hospitalized patients after CVR surgery. The results showed that the performance of reinforcement learning was significantly better than the other algorithms, and the quality of anticoagulation was significantly optimized [62]. Finally, implementing effective evaluation mechanisms for patients lost to follow-up is critical to ensure they receive appropriate anticoagulant treatment over the long term. Porter et al [63] conducted a 2-year follow-up of patients receiving oral anticoagulants and found that 12-week INR test intervals were feasible, which saved patient’s time and improved their compliance [63]. By addressing these factors and optimizing the application of eHealth tools in the AM of patients after CVR, it is possible to significantly improve the overall therapeutic effect and reduce the risk of complications.

Limitations
This review has some limitations. First, we did not assess the quality of the included studies because the study was to provide a comprehensive overview of the existing research landscape related to the application of eHealth tools in AM for patients after CVR. However, we believe that evaluating the quality of the literature will increase the significance of the study. Second, we did not include studies published in the form of other sources, such as qualitative studies, because we were unable to identify eligible studies. Third, concluding trends based on 25 papers are limited, and the conclusions of scientometric analysis should be interpreted with caution. The scientometric analysis in our research is influenced by several limitations, such as the omission of qualitative data, language barriers, and the potential for sample bias. To address the constraints associated with scientometric analysis, researchers can broaden their data sources, integrate qualitative data, and encompass papers published in various languages.

Conclusions
The application of AM based on eHealth tools is expected to truly reflect the social and economic benefits of digital intelligence, thus benefiting patients undergoing CVR. This study provides an overview of the scope, benefits, and future development of eHealth tools in AM for researchers, health care professionals, and post-CVR patients. However, the economic benefits and long-term impact of adverse events need further explored. It is suggested that future eHealth tools in AM should concentrate on enhancing patient contact, acceptance, cost-effectiveness, and sustainability while combining sophisticated algorithms to enhance the precision and dependability of eHealth tools.

Acknowledgments
The authors specially thank the National Natural Science Foundation of China, Hunan Provincial Natural Science Foundation, and Hunan Normal University. This work was supported by the National Natural Science Foundation of China (82100490 to LC), Hunan Provincial Natural Science Foundation (2023JJ30425 to LC), Hunan Normal University Students’ Innovation and Entrepreneurship Training Program (2022263 to MZ), and Hunan Normal University School of Medicine Open Foundation Project (KF2022043 to XW).

Data Availability
The data sets used and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest
None declared.

Multimedia Appendix 1
References


Abbreviations

AI: artificial intelligence
AM: anticoagulation management
CVR: cardiac valve replacement
EHR: electronic health record
INR: international normalized ratio
RCT: randomized controlled trial
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Review
TTR: time in therapy range

©Ying Wu, Xiaohui Wang, Mengyao Zhou, Zhuoer Huang, Lijuan Liu, Li Cong. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 05.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Dissemination Strategies for mHealth Apps: Systematic Review

Henri Claude Moungui1, MPH, MSc; Hugues Clotaire Nana-Djeunega2, MSc, PhD; Che Frankline Anyiang3, MSc; Mireia Cano4, MSc; Jose Antonio Ruiz Postigo5, MD, PhD; Carme Carrion4, MSc, PhD

1Universitat Oberta de Catalunya, Barcelona, Spain
2Higher Institute for Scientific and Medical Research, Yaounde, Cameroon
3Texila American University, Georgetown, Guiana
4eHealth Lab Research Group, eHealth Center & School of Health Sciences, Universitat Oberta de Catalunya, Barcelona, Spain
5Prevention, Treatment and Care Unit, Department of Control of Neglected Tropical Diseases, World Health Organization, Geneva, Switzerland

Corresponding Author:
Henri Claude Moungui, MPH, MSc
Universitat Oberta de Catalunya
Rambla del Poblenou, 156
Barcelona, 08018
Spain
Phone: 34 672192283
Email: henrimoungui@yahoo.fr

Abstract

Background: Among the millions of mobile apps in existence, thousands fall under the category of mobile health (mHealth). Although the utility of mHealth apps has been demonstrated for disease diagnosis, treatment data management, and health promotion strategies, to be effective they must reach and be used by their target audience. An appropriate marketing strategy can ensure that apps reach potential users and potentially convert them to actual users. Such a strategy requires definitions of target end users, communication channels, and advertising content, as well as a timeline for effectively reaching and motivating end users to adopt and maintain engagement with the mHealth app.

Objective: The aim of this study was to identify strategies and elements that ensure that end users adopt and remain engaged with mHealth apps.

Methods: A systematic search of the PubMed, PsycINFO, Scopus, and CINAHL databases was conducted for suitable studies published between January 1, 2018, and September 30, 2022. Two researchers independently screened studies for inclusion, extracted data, and assessed the risk of bias. The main outcome was dissemination strategies for mHealth apps.

Results: Of the 648 papers retrieved from the selected databases, only 10 (1.5%) met the inclusion criteria. The marketing strategies used in these studies to inform potential users of the existence of mHealth apps and motivate download included both paid and unpaid strategies and used various channels, including social media, emails, printed posters, and face-to-face communication. Most of the studies reported a combination of marketing concepts used to advertise their mHealth apps. Advertising messages included instructions on where and how to download and install the apps. In most of the studies (6/10, 60%), instructions were oriented toward how to use the apps and maintain engagement with a health intervention. The most frequently used paid marketing platform was Facebook Ads Manager (2/10, 20%). Advertising performance was influenced by many factors, including but not limited to advertising content. In 1 (10%) of the 10 studies, animated graphics generated the greatest number of clicks compared with other image types. The metrics used to assess marketing strategy effectiveness were number of downloads; nonuse rate; dropout rate; adherence rate; duration of app use; and app usability over days, weeks, or months. Additional indicators such as cost per click, cost per install, and clickthrough rate were mainly used to assess the cost-effectiveness of paid marketing campaigns.

Conclusions: mHealth apps can be disseminated via paid and unpaid marketing strategies using various communication channels. The effects of these strategies are reflected in download numbers and user engagement with mHealth apps. Further research could provide guidance on a framework for disseminating mHealth apps and encouraging their routine use.

(JMIR MHealth UHealth 2024;12:e50293) doi:10.2196/50293
KEYWORDS

mobile health; mHealth; mobile health apps; mHealth apps; dissemination; marketing strategies; digital marketing; engagement; onboarding; systematic review; systematic; market; marketing; app; apps; adoption; consumer; mobile phone

Introduction

Mobile Health Apps

Among the vast array of mobile apps currently available, health care apps serve various purposes, including disease diagnosis [1], health promotion, and disease prevention [2,3]. Such uses of mobile technology to provide patients with health care support or health service providers with technical support in a direct, low-cost, and engaging manner fall under the category of mobile health (mHealth) [4].

With approximately 200 new mHealth apps released every day, the number available now exceeds 300,000 [5]. One factor in this proliferation is the increasing use of mHealth technology by health service providers who not only seek advice from apps but also prescribe them to their patients [6]. In October 2020, Germany became the first country to cover the prescription costs of certain mHealth apps through statutory health insurance [7]. Moreover, because mHealth apps have the potential to replace a number of health provider tasks, it has been suggested that expertise in the use of mHealth-related technologies should be recognized as an essential competency for providers [1].

Dissemination of mHealth Apps to Users

The amount of academic research on mHealth apps has also increased, particularly in the areas of usability, effectiveness, adoption, and assessment. However, the highly important aspects of dissemination and marketing are as yet underexamined. App marketing refers to measures aimed at making a mobile app better known and acquiring users (ie, generating app downloads) and, moreover, contacting users and encouraging them to reach a specified goal [8].

An mHealth app is not a stand-alone product that can work effectively without human interaction, which cannot take place if users are unaware that a certain app exists and is accessible. Such “human touches,” although extraneous to the app itself, can be crucial for promoting use [9].

As users are key actors in mHealth adoption, it is critical to understand how they navigate the various stages from app discovery to frequent use. Google has created such a model [10]. It consists of four key stages: (1) discover (users come across an app and download it to their device), (2) onboard (the process of first use and registration), (3) engage (users start using the app regularly), and (4) embed (the desired outcome as users view the app as “vital” to their lives). Only a small proportion of users currently reach the embedment stage with any app [10]; for instance, the literature is sparse regarding the long-term integration and penetration of mobile interventions within mental health and other support service settings [11].

Increasing the chances of an app achieving embedment requires understanding users and placing them at the core of mHealth services. This process would start with persona definitions: fictional archetypes of actual product users. A persona enables program designers to create high-quality programs that effectively meet user needs [2]. In the marketing world, this also means segmenting users and locating them on the marketing funnel, which is a visual representation of the different phases in a customer’s journey toward conversion and their relationship with a product. By segmenting customers based on where they are located in the funnel, marketers target these groups much more effectively [12].

Positive customer experiences and journeys rely on ensuring that the consumer sees value in an app as a channel for accessing products and services and as a 2-way platform for seamless interactions. Although marketing strategies play a crucial role during the early stages of the customer journey, they have been subjected to very little analysis [13].

Marketing of Health Apps

A successful marketing strategy can ensure that an mHealth app reaches potential users and ease the adoption process. Such a strategy would clearly define target end users; determine the appropriate communication channels, content, and timelines to effectively reach users; and market the app as an attractive product, encouraging people to download it and become regular users. The strategy would include a mix of activities, depending upon the type of app and upon the stage of the launch period (from before the launch to after the launch), including email marketing, targeted advertising, and social media promotion [14].

Marketing services have evolved alongside information and communication technologies. In turn, digital marketing has provided a series of customized platforms for communicating with specific stakeholders using computers, smartphones, and tablet computers [15]. These channels enable information to be gathered and include websites as well as various social media platforms such as Facebook, YouTube, X (the platform formerly known as Twitter), Pinterest, TikTok, and LinkedIn. Traditional marketing also remains an option, with products being promoted on radio and television channels, as well as via printed posters in public spaces, flyers, and face-to-face conversations [16].

The cost of promoting an app will depend upon where money is spent, and those promotional activities that do not cost money will demand time. Although potential customers may be offered incentives such as money or supplemental products, the marketer or marketing firm involved in digital marketing can be offered a fee per click, download, or install. A small pilot trial of activities that cost money is recommended to assess results [14].

Specialized services are available for driving digital campaigns; for example, Facebook Ads Manager is a paid service that oversees paid digital marketing campaigns across the Facebook platform. Google Universal App Campaigns (UAC) is another paid service that promotes mobile apps by distributing marketing messages across Google formats and networks, such as the first page of applicable Google search results and small banner advertisements on relevant YouTube channels [17]. As Google
shares information among platforms, including Google Display Network, YouTube, and Google Play Store, Google UAC can capture the number of Google-driven impressions, clicks, and installs on Android devices.

The effectiveness of a marketing strategy can be observed not only through the number of app downloads but also by the effects of users interacting with the app. Referred to as mobile app engagement, this is defined by a variety of operational metrics, such as the number of log-ins, the number of days of use, the number of pages visited, and the number of tasks or modules completed [18]. Another gauge of marketing effectiveness is user onboarding. In the context of mobile apps, user onboarding is the process of providing instructions and highlighting key benefits and features via a set of example screens when the user first launches the app [19].

**Objectives**

The aim of this study was to review existing evidence on strategies and elements relevant to how the dissemination of mHealth apps is currently carried out and how these elements contribute to encourage user onboarding and engagement with mHealth apps.

**Methods**

**Overview**

This study was carried out following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [20] (Multimedia Appendix 1). The protocol of this review was registered with PROSPERO (CRD42022352369) [21].

**Eligibility Criteria**

Eligible sources were academic papers. All cohort studies, cross-sectional studies, and randomized controlled trials reporting on the dissemination and marketing of an mHealth app were included. The outcome expected from included studies was strategies and elements relevant to how the dissemination of mHealth apps is currently carried out and how these elements contribute to encourage user onboarding and engagement with mHealth apps.

Editorials, letters to the editor, scoping reviews, systematic reviews, meta-analyses, conference abstracts, commentaries, essays, book chapters, and study protocols were excluded, as were studies with any other study design (eg, bibliometric analysis, modeling study, systematic or web search or review of apps, landscape analysis, and scorecard analysis). We also excluded studies with participants aged <18 years and those not reporting the expected outcome. Search languages were limited to French and English.

**Information Sources and Search Strategy**

Searches were conducted on PubMed, PsycINFO, Scopus, and CINAHL databases over the 5-year period from January 1, 2018, to September 30, 2022 (refer to Multimedia Appendix 2 for the search strategy). The following search terms were used individually or combined according to Medical Subject Headings terms: “apps,” “mHealth,” “marketing,” “promotion,” and “dissemination.” Moreover, we conducted searches on JMIR and mHealth journals and cross-checked the reference lists of the selected studies to locate additional studies that met the inclusion criteria. The main outcome was dissemination strategies for mHealth apps.

**Study Selection and Data Collection Process**

All retrieved studies were imported into Rayyan (Rayyan Systems Inc) [22] and duplicate records eliminated. Screening consisted of blind peer review by 2 independent investigators. Any conflict was resolved through discussion or the adjudication of a third investigator.

**Results**

**Selection of Studies**

We identified 638 records through database searches (PubMed: n=215, 33.7%; PsycINFO: n=60, 9.4%; Scopus: n=283, 44.4%; and CINAHL: n=80, 12.5%) and 10 records through JMIR and mHealth journal searches. Of the total 648 records, 127 (19.6%) duplicates were removed. Next, of the remaining 521 articles, 502 (96.4%) were removed after title and abstract screening. The reasons for exclusion were unrelated outcome or outcome other than the subject of our review (309/502, 61.6%), study protocol (136/502, 27.1%), published review (48/502, 9.6%), study participants aged <18 years (6/502, 1.2%), and other study design (3/502, 0.6%). We then assessed the remaining 19 full texts for eligibility and excluded 9 (47%) for unrelated outcomes; thus, 10 (53%) papers were included in this review. A PRISMA-compliant flow diagram [20] of the paper selection process is shown in Figure 1. The characteristics of the studies selected are summarized in Table 1.
Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the paper selection procedure.
<table>
<thead>
<tr>
<th>Authors, year; country</th>
<th>Population</th>
<th>Intervention</th>
<th>Outcomes</th>
<th>Design</th>
<th>Quality</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kvedarienė et al [23], 2019; Lithuania</td>
<td>N=149</td>
<td>Monitoring of allergic rhinitis and asthma in real life in Lithuanian users</td>
<td>High app engagement</td>
<td>Longitudinal study</td>
<td>Low</td>
<td>Selection bias: patients recruited from an allergy clinic</td>
</tr>
<tr>
<td>Buss et al [24], 2022; Australia</td>
<td>N=46</td>
<td>Test the feasibility of an app-based intervention for cardiovascular and diabetes risk awareness and prevention</td>
<td>The app scored the highest for the information section and the lowest for the engagement section of the scale</td>
<td>Nonrandomized controlled trial (cohort study)</td>
<td>Low</td>
<td>Small sample size</td>
</tr>
<tr>
<td>Arshanapally et al [17], 2022; United States</td>
<td>N=NI</td>
<td>Investigate the outcomes of a paid digital marketing campaign to promote a mHealth app about parent-engaged developmental monitoring</td>
<td>Paid digital marketing can be an effective strategy to promote mHealth apps targeting parents of young children</td>
<td>Implementation study</td>
<td>Low</td>
<td>No relevant limitations</td>
</tr>
<tr>
<td>Resnick et al [25], 2021; United States</td>
<td>N=41</td>
<td>Assess the usability, acceptability, and user engagement of the Healthier Together mobile app</td>
<td>The app strongly engaged participants, with promising results on participants' knowledge of cancer prevention behaviors and success in achieving their cancer prevention behavioral goals</td>
<td>Mixed methods intervention</td>
<td>Low</td>
<td>Small sample size</td>
</tr>
</tbody>
</table>

Note: NI indicates not indicated.
<table>
<thead>
<tr>
<th>Authors, year; country</th>
<th>Population</th>
<th>Intervention</th>
<th>Outcomes</th>
<th>Design</th>
<th>Quality</th>
<th>Limitations</th>
</tr>
</thead>
</table>
| Zlotorzynska et al [26], 2021; United States | • N=NI  
• Sex: 0% female  
• Age: 18-24 years  
• Other details: YMSM | Paid web-based recruitment campaign to recruit HIV-negative or unknown status YMSM for 4 randomized controlled trials of mHealth HIV prevention interventions | • Instagram advertisements yielded the highest proportions of eligible contacts who were racial or ethnic minority individuals and aged <18 years | Randomized controlled trials | Moderate | • Participants offered incentive to enroll in study |
| Rajani et al [27], 2021; United Kingdom | • N=154  
• Sex: 38.8% female  
• Age: 18-65 years  
• Other details: smokers | Examine the association among perceived usefulness, perceived ease of use, and frequency of use of gamification features embedded in smoking cessation apps on self-efficacy and motivation to quit smoking | • The use of the apps was associated with increased self-efficacy and motivation-to-quit levels 4 weeks after app use compared with baseline  
• Perceived frequency of use of gamification features was associated with an increase in self-efficacy and motivation to quit  
• Higher baseline self-efficacy and motivation to quit were both associated with smaller increases in self-efficacy and motivation-to-quit levels 4 weeks after using the mobile apps compared with preapp use | Observation-al study | Low | • Participants incentivized |
| Roberts et al [28], 2019; United Kingdom | • N=32  
• Sex: 68.8% female  
• Age: ≥18 (mean 60, SD 11; range 37-78) years  
• Other details: diagnosed with breast, prostate, or colorectal cancer | To seek opinions of survivors of breast, prostate, and colorectal cancer regarding using apps to promote PA | • Multiple factors affect engagement with PA apps, and this is highly personalized  
• Apps that promote walking are the most appealing for survivors of cancer  
• PA apps should be integrated into cancer care | Cross-sectional study | Low | • Small sample  
• Participants were offered a £10 (£12.7) voucher as an incentive for completion of study, and costs were reimbursed if asked to install an app that was not free to download |
| Bidargaddi et al [29], 2018; United States and Australia | • N=1255  
• Sex: NI  
• Age: NI  
• Other details: NI | To study the effect of time-varying push notifications on engagement in self-monitoring activity | • Pushing a notification with a tailored health message affects near-time proximal engagement with the self-monitoring activity in the app | Microrandomized trial design | High | • No relevant limitations |
| Hui et al [30], 2018; United States | • N=101  
• Sex: 87.1% female  
• Age: ≥16 years  
• Other details: patients with active asthma | The impact of different recruitment strategies and app features on adoption and continued use | • Adherence was dependent upon motivation derived from a sense that the health care professional and researcher were interested in the results and that using an app to support their self-management could improve their asthma control | Cohort study | Low | • Selection bias: participants recruited at clinics |

Market tests of the CycleBeads app in 7 countries  
In-app microsurveys  
No relevant limitations
Of the 10 selected studies, 6 (60%) were observational studies (longitudinal, cohort, implementation study, mixed methods, and cross-sectional), 1 (10%) reported on 4 randomized controlled trials [26], 1 (10%) reported on in-app microsurveys [31], and 1 (10%) reported on a microrandomized trial [29].

According to the Scottish Intercollegiate Guidelines Network (SIGN) [32] criteria, 8 (80%) of the 10 studies were of low quality, 1 (10%) of moderate quality, and 1 (10%) of high quality. A low-quality rating resulted from a small sample size, the study design (mostly observational), or possible selection bias. A summary of the design, quality, and limitations of the included studies can be found in Table 1.

The selected 10 studies covered 8 countries—Lithuania, the United Kingdom, Australia, the United States, Egypt, Ghana, India, and Jordan—with 50% (5/10) of the studies conducted in the United States and 20% (2/10) in the United Kingdom.

The number of participants in the selected studies ranged from 32 to 1255. Of the 10 studies, 3 (30%) that recruited participants through social media and used impressions and clicks as a proxy measure of their number did not state the number of participants. In 6 (60%) of the 10 studies, 50% to 100% of the participants were female; sex information was not given in 2 (20%) of the 10 papers, 1 (10%) study targeted adolescent and young male individuals, and 1 (10%) targeted female individuals only.

**Objectives of the Marketing Strategy**

In each of the studies reviewed, we looked for the objective of the marketing strategy implemented through the lens of the Google model of mobile app user journey (discover, onboard, engage, and embed) [10] (Table 2). We found that 60% (6/10) of the studies aimed to drive people to discover their mHealth apps (come across the apps and download them), get onboard (start using the apps), and stay engaged by using the apps for a couple of weeks or months. Marketing strategy in 30% (3/10) of the studies targeted the discover and onboard stages, whereas 1 (10%) of the 10 studies aimed at the engage stage only.
### Table 2. Summarized results of reviewed studies.

<table>
<thead>
<tr>
<th>Authors, year</th>
<th>Channels</th>
<th>Procedures of each dissemination strategy: How was it done?</th>
<th>Effects and results of each dissemination strategy</th>
<th>Objective with regard to the Google model for users’ journey with apps</th>
</tr>
</thead>
</table>
| Kvedarienė et al [23], 2019 | Face-to-face | • Health personnel trained patients how to use the app at a clinic | • The duration of app use in patients ranged from 1 to 680 (median 54, IQR 23-151) days  
• There was an increased frequency when the reported days were >200 (18%)  
• Adherence to mobile apps was higher when the app was promoted by physicians and when the users were taught how to use it | Discover, onboard, and engage |
| Buss et al [24], 2022 | Email | • Participants received an email and a user guide that included instructions to download an app from the app store on their mobile phones and then use it for 3 months  
• Researchers encouraged regular use  
• If they had questions or technical issues, participants could get in touch with researchers via email | • Of the 46 participants, 20 (43%) never used the app, and 15 (33%) dropped out  
• Adherence to app use (using the app at least once/week over 3 months) was 17% (8/46)  
• The median time between the first and last app use was 54 days  
• The research team did not actively engage with participants early in the study to verify that all participants could install the app  
• The intervention did not involve direct contact with health care professionals  
• The app did not contain enough interactive features | Discover, onboard, and engage |
| Arshanapally et al [17], 2022 | Social media | • Google- or Facebook-driven marketing message: when a user clicked on a marketing message, they were directed to either the Google Play Store or the Apple App Store, depending upon device type  
• Google Universal App Campaigns distributed marketing messages across several Google formats and networks  
• Facebook Ads Manager distributed marketing messages across the Facebook platform | • The Google-driven marketing messages garnered a total of 4,879,722 impressions and 73,956 clicks (clickthrough rate: 1.52%); from these clicks, there were 13,707 installs of the app on Google Play Store (18.53% install rate); the overall cost/install was US $0.93  
• The Facebook-driven marketing messages garnered a total of 2,434,320 impressions and 44,698 clicks (clickthrough rate: 1.84%); the average cost/install for Facebook could not be calculated because of the limitation of collecting Facebook-driven installs data  
• Animated graphics generated the greatest number of clicks among both English and Spanish audiences on Facebook when compared with other types of images | Discover, onboard, and engage |
| Resnick et al [25], 2021 | Face-to-face | • In-person information about the app, followed by installation of the mobile app on the participant’s mobile phone and instructions on how to select a goal, choose share settings, and invite other social ties  
• Weekly SMS text messages reminding participants to check in, share goal progress, and invite relatives | • Of the 171 participants, 41 (24%) downloaded the app  
• Engagement with the app (mean number of check-ins/participant out of 8 possible weekly check-ins during the study period) was 5.7  
• Of the 41 participants, 31 (76%) checked in during at least 4 of the 8 weeks  
• The mean System Usability Scale score was 87 (SD 12; median 90, IQR 78-95) | Discover, onboard, and engage |
<table>
<thead>
<tr>
<th>Authors, year</th>
<th>Channels</th>
<th>Procedures of each dissemination strategy: How was it done?</th>
<th>Effects and results of each dissemination strategy</th>
<th>Objective with regard to the Google model for users’ journey with apps</th>
</tr>
</thead>
</table>
| Zlotorzynska et al [26], 2021 | Advertising on Grindr, Snapchat, Instagram, and Facebook | • Advertising was purchased on Facebook, Instagram, Snapchat, Twitter, and Grindr; users who clicked on banner advertisements were taken directly to a study-specific eligibility screener and, if eligible, were asked to provide contact information for follow-up by respective study site staff.
• The Facebook Ads Manager proprietary algorithm allocated the distribution of advertisement placements (news feed, right-hand column, and Instagram feed and stories) that would provide the best performance.
• Advertising copy was written to give a very brief description of the study, identify the organization conducting study recruitment, mention the study incentive, and provide a call to action for potential participants seeking to learn more.
• Snapchat advertisements were created using Snap Publisher and were in the form of short videos up to 10 seconds long; these videos rotated through various stock photo images and superimposed text.
• Instagram placements were used to test advertisement performance.
• Images used in advertisements included stock photos, which were either purchased from web-based stock photo vendors (eg, Shutterstock) or accessed through Facebook’s free stock photo catalog integrated within the Facebook Ads Manager.
• Recruitment through various in-person and community outreach efforts included posting flyers, tabling at community events, reaching out to past study participants, and recruiting through clinics serving youth. | • Grindr and Snapchat advertisements produced the highest clickthrough rate compared with Facebook advertisements; these advertisements had the lowest proportions of users who initiated eligibility screeners.
• Facebook advertisements yielded the lowest cost/eligible contact, whereas advertisements on Twitter had the highest.
• Facebook or Instagram advertisements had much higher rates of screening and ultimately yielded much lower costs/eligible participant.
• The cost/eligible participant was markedly higher for the Instagram advertisements campaigns. | Discover and onboard |
| Rajani et al [27], 2021 | Social media and paper posters | • Participants were recruited via social media, and posters were displayed in public places in London, United Kingdom.
• Participants were provided instructions on the internet on how to download and start using the app. | • Of the 138 participants who installed the app, 116 (84.1%) completed all 4 weeks of the study.
• A 1-point increase in the average perceived frequency of gamification features was statistically significantly associated with a 3.35-point increase in self-efficacy from baseline to study end ($\beta=3.35, 95\% \text{ CI} 0.31-6.40$). | Discover and onboard |
<p>| Roberts et al [28], 2019 | Face-to-face, paper posters, email, and social media | | | Discover, onboard, and engage |</p>
<table>
<thead>
<tr>
<th>Authors, year</th>
<th>Channels</th>
<th>Procedures of each dissemination strategy: How was it done?</th>
<th>Effects and results of each dissemination strategy</th>
<th>Objective with regard to the Google model for users’ journey with apps</th>
</tr>
</thead>
</table>
| Bidargaddi et al [29], 2018 | Push notifications | - Participants recruited via advertisements within community-based cancer support groups (either by verbal descriptions from group leaders at meetings or via posters, flyers, and email mailing lists), Facebook cancer support groups, and charitable organizations  
- Participant randomly assigned to download 2 of 4 apps (Human, The Walk, The Johnson & Johnson Official 7 Minute Workout, and Gorilla Workout)  
- Guidance in downloading and installing each app was provided, if required; participants were asked to spend approximately 2 consecutive weeks using the apps | - Of the 40 participants, 32 (80%) completed the study (dropout rate: 20%)  
- Reasons for dropping out were lack of time, family circumstances (eg, bereavement), and not wanting to update their smartphone’s operating system or register credit card details with Google Play  
- Factors affecting engagement included participants’ perceptions of the advantages and disadvantages of using apps to support physical activity, the relevance of the app to the user, the quality of the app, and the behavior change techniques used to promote physical activity | Engage |
| Hui et al [30], 2018 | Social media and face-to-face | - Push notifications were sent at 1 of 6 chosen time points throughout the day, and a user could either receive or not receive a push notification at a chosen time point  
- At each considered time point, users were randomized to either receive or not receive a push notification containing a tailored health message with a 50% probability; once a time point was considered, the user was then considered unavailable for the remainder of the day  
- To mitigate the risk of users either turning off notifications or deleting the app owing to receiving too many push notifications, users were classified as either “available” or “unavailable” at each time point, and only those time points when users were “available” were considered for the push notification decision; several rules were applied to determine availability | - Sending a push notification containing a tailored health message was associated with greater engagement in a mobile health app  
- The effect of the pushed notifications is sustained over time; push notifications containing tailored health messages can attenuate the rate at which users disengage  
- Users who used the app less frequently were “unavailable” to receive push notifications on a greater number of days than those who used the app more often | Discover, onboard, and engage |
| Haile et al [31], 2018 | Social media and face-to-face | - Researchers sent letters inviting people to use the app for up to 3 months  
- Practice nurses approached adults with active asthma to try out a prototype app  
- The app was also advertised via social media (Facebook and Twitter [now known as X])  
- Nurses were not expected to teach patients to download and use the app  
- Technological support was provided by the research team and the app developer  
- The messages on social media included brief information and a link to information for patients on how to download the app | - A total of 300 patients received information on the app and study from Facebook, Twitter, or organic searches  
- Only 135 patients downloaded the app, with 111 (82.2%) registering an account on the app  
- Social media attracted 87 users, but only 15 (17%) used the app for 30 days  
- A total of 24 patients were recruited, and 13 (54%) continued for 30 days  
- Successful adoption was dependent upon the ease of downloading and sufficient motivation  
- Some patients needed technological assistance with downloading the app and starting to use the features | Discover and onboard |
Channels and Procedures of Each mHealth App Dissemination Strategy

As shown in Table 2, mHealth app promotion was conducted through various channels, including paper posters, face-to-face communications, email, and social media. The most reported channels were face-to-face [23,25,28,30,31] and social media [17,27,28,30,31], both of which appeared in 50% (5/10) of the studies. Messages were intended to inform people about the existence of the app, what it does and the benefits of using it, where to find the downloading link, downloading instructions, and information on how to use the app. Messages through these channels used text, voice-overs, and short videos. Messages on social media and email included a shortened URL to download the app. Face-to-face interactions [23,25,28,30,31] provided the opportunity for training support, including demonstrating app features, installing and using the app, and solving technical issues.

Google Play Store and Apple App Store were the main web repositories to which potential users were redirected for download. In addition, the social media pages of various organizations, such as the Facebook pages of cancer support groups [28], were also used to promote apps and provide download links.

Effects of Dissemination Strategies

Advertising performance was influenced by many factors, including but not limited to advertising content and creativity, competing advertisements within platforms, and emerging platforms that attracted users to new digital spaces [26]. The reviewed papers revealed that social media advertisements attracted substantial downloads over a short period with a large number of impressions [13] and prompted patients to respond to invitations to use the app. However, without the motivation provided by a trusted professional, users quickly ceased their use of the app; a dual promotion strategy was then used to increase app adoption: using social media for its reach and ease of response as well as promoting personal invitations from trusted health care professionals [25].

In 1 (10%) of the 10 studies, embedded push notifications were sent to users to keep them engaged with the app [24]. This effect was sustained over time, and push notifications containing tailored health messages reduced the rate at which users disengaged. The timing of notifications was a key factor, with the results suggesting that users were more likely to engage with an app within 24 hours when push notifications were sent at midday on a weekend. Animated graphics generated a greater number of clicks than other image types [17].

The success of mHealth app dissemination strategies in these studies was measured in terms of adherence, app use, the duration of app use, impressions, clicks, app downloads, and installs. Marketing messages with simple direct calls to action, such as “Track your child’s development,” generated high clickthrough rates (CTRs) [17].

Marketing Messages and Content of Communications

In some cases, promotional messages included questions prompting people to download and test the app, such as “Would you like to test an app that helps you manage your asthma better? [shortened URL] Need a helping hand to manage your asthma? Try our new app [shortened URL]” [30]. One study included the inquiring headline “Ever think of testing at home?” and explanatory body text: “Fenway Health is looking for young men to help test out a mobile app to support sexual health” [26].

In a study promoting an mHealth app about parent-engaged developmental monitoring [17], marketing messages with animated graphic interchange format (GIF) images as well as images of younger children performed higher than messages without GIF images and images of older children.

Paid or Nonpaid Marketing Campaigns

The paid marketing campaign platforms that were used included Facebook Ads Manager and Google UAC; in addition, advertisements were placed on Grindr, Snapchat, and Instagram [17,26,31]. In 50% (5/10) of the studies, participants were offered incentives to participate [24-28]. Of these 5 studies, 4 (80%) were conducted in the United States and 1 (20%) in Australia. No study reported that paid marketing resulted in higher app adoption than unpaid campaigns.

Key Indicators of mHealth App Marketing Strategies

Various metrics were used in the reviewed studies to assess mHealth app marketing strategies. Key indicators included nonuse rate, dropout rate, and adherence rate, as well as median time between first and last app use, which was generally measured in days or weeks. Nonuse rate was defined as the proportion of participants who never used the app [24], and the duration of use was assessed by determining the reported number...
of days of use [23]. The longest duration of app use was 680 days [23].

Dropout rate was defined as the proportion of participants who completely stopped using the app at least 14 days before they received an invitation to complete the end-of-study survey. Adherence rate was defined as the proportion of participants who used the app at least once every week over 3 months of the study [24].

Specific indicators for paid marketing strategies were cost per click (CPC), defined as the amount of money spent per click secured in each advertisement campaign or advertisement set; CTR, the number of clicks divided by impressions; and cost per impression [26]. Impression is the number of times an advertisement is viewed by a user on an advertising platform. For advertisements seen multiple times by the same user, each view is counted as an impression. Additional indicators were the number and percentage of clicks that took people through to the eligibility screener and of those who were eligible for the study [26].

**People Implementing the Marketing Processes**

In the studies reviewed, we found that promotional messages were mainly disseminated by researchers, trained nurses, and other health practitioners. Technological support was only provided by the research team and app developers. In some cases, recently enrolled patients contributed to promoting the app by sending download invitations to their family and friends [25].

**Marketing Concepts Reported in the Reviewed Studies**

Inspired by a book on how to successfully distribute apps [12], we screened the included studies to identify the marketing concepts that were used to promote mHealth apps in these studies (Table 3).
Table 3. Marketing concepts [12] reported in the reviewed studies.

<table>
<thead>
<tr>
<th>Authors, year</th>
<th>Incentivization(^a)</th>
<th>Personalization(^b)</th>
<th>Mobile app attribution(^c)</th>
<th>Loyalty marketing(^d)</th>
<th>Remarketing or re-targeting(^e)</th>
<th>A/B testing(^f)</th>
<th>Programmatic marketing(^g)</th>
<th>Predictive marketing(^h)</th>
<th>Thought-leadership marketing(^i)</th>
<th>Content marketing(^j)</th>
<th>Behavioral marketing(^k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kvedarienė et al [23], 2019(^l)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Buss et al [24], 2022(^l)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Arshanapally et al [17], 2022(^l)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Resnick et al [25], 2021(^l)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Zlotorzynska et al [26], 2021(^l)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Rajani et al [27], 2021(^m)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Roberts et al [28], 2019(^m)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Bidargaddi et al [29], 2018(^l)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Hui et al [30], 2018(^m)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Haile et al [31], 2018(^m)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

\(^a\)The incentivized model is the strategy of making a product, program, or other offering more attractive to customers by offering an incentive in exchange for buying or participating. In the app business, incentivization is normally used to quickly amass app installs. Of the 10 studies, half of the studies (n=5, 50%) used incentivization.

\(^b\)Personalization involves customizing the timing and content of marketing messages to the target user based on their preferences, habits, and behavior patterns. Personalized messages refer to every user by name and entice the user to become more engaged with an app with the right kind of incentive based on their characteristics, such as age, gender, location, profession, and financial segment. Half of the studies (5/10, 50%) used personalization.

\(^c\)Mobile app attribution is the process of recording and measuring the actions of app users, such as installs, level completions, in-app purchases, and other milestones. Mobile app attribution is essential to app marketing because it helps produce the data that are gathered and analyzed to measure how well marketing campaigns are working. The majority of the studies (7/10, 70%) used mobile app attribution.

\(^d\)Loyalty marketing is a marketing strategy that focuses on nurturing existing customers rather than acquiring new ones. Only 1 (10%) of the 10 studies used loyalty marketing.

\(^e\)Remarketing or re-targeting targets every individual who has come into contact with the product but has not converted or who converted but later abandoned the app. It allows marketers to reconnect with these categories of users and “bring them back” or increase the time they spend engaging with the app. None of the studies used remarketing or re-targeting.

\(^f\)A/B testing involves the use of several versions of the same advertisement distributed to different groups with different designs, color coding, calls to action, and message content to determine which version produces the highest conversion rate. Of the 10 studies, only 2 (20%) used A/B testing.

\(^g\)Programmatic marketing is the automated algorithm-based real-time buying and selling of advertising space through a bidding system, with the aim of reaching the right customers at the right time. None of the studies used programmatic marketing.

\(^h\)Predictive marketing involves using data science based on customer behavior and habits to make smarter marketing decisions. By gathering and analyzing data about user behavior and identifying patterns, marketers can make forecasts about user behavior and make informed decisions about the likelihood of the success of their marketing content and offerings. Of the 10 studies, only 1 (10%) used predictive marketing.

\(^i\)Thought-leadership marketing is the process of positioning a company as a leader in a specific domain by supplying customers with top-quality information. Only 1 (10%) of the 10 studies used thought-leadership marketing.
Content marketing is a marketing strategy that involves producing content that potential customers find useful, valuable, and relevant. Content marketing is highly effective at building a loyal user base and converting leads into customers. More than one-third of the studies (4/10, 40%) used content marketing.

Behavioral marketing involves segmenting the app’s user base based on user behavior with the aim of refining the marketing strategy and more effectively targeting users. More than one-third of the studies (4/10, 40%) used behavioral marketing.

These studies (n=6, 60%) used single-channel marketing, which involves reaching users through a single channel, eg, Facebook advertisements.

These studies (n=4, 40%) used multichannel marketing, which involves >1 channel, as opposed to an omni-channel marketing campaign, which attempts to reach users through all available channels.

In summary, most of the studies (7/10, 70%) reported using a combination of marketing concepts to advertise their mHealth apps. The most used concept was mobile app attribution (7/10, 70%). Mobile app attribution is essential to app marketing because it helps produce the data that are gathered and analyzed to measure how well marketing campaigns are working [12]. Furthermore, 40% (4/10) of the studies reported using at least 2 channels to market their mHealth apps.

**Discussion**

**Principal Findings**

We found that the marketing strategies used in almost all included studies (9/10, 90%) were aiming to drive at least app discovery and onboarding, with more than half (6/10, 60%) also targeting user engagement to mHealth apps. Social media, emails, television or radio, posters or flyers, and face-to-face communications were all used in the reviewed studies to inform people about the existence of mHealth apps, invite them to download, encourage them to use the apps, and maintain engagement. This is consistent with an integrative review of methods used to promote mobile apps, which also cited app store optimization via keywords and the inclusion of screenshots and videos for greater conversion rate, the use of push notifications, the promotion of apps via influencers, and the leveraging of user review and ratings [13].

Social media attracted many downloads over a short period, whereas emails were most often used for sharing instructions on how to download apps and interact with research teams for technical support.

The strategies used to promote mHealth apps included paid and unpaid marketing, and metrics such as CPC and CTR were used to measure effectiveness. Offering incentives to people to download and use the app did increase app downloads and use. However, it has been shown that people are less likely to keep using an app after incentivization, although the monetary value of the incentive could have a significant effect on the adherence [3]. Furthermore, a focused strategy is required to maintain a low app churn rate [12].

**Factors Influencing Engagement, Onboarding, or Adherence to mHealth Apps**

**Reasons for User Engagement, Onboarding, or Adherence to the App**

Onboarding was often reliant on the ease of downloading and sufficient motivation. In the study by Hui et al [30], adherence stemmed from awareness that a health care professional and a researcher were interested in the results and that using an app to support self-management could improve a participant’s control of their health condition. Additional factors in engagement included the perceptions of the advantages and disadvantages of using apps to support self-management specific to an individual’s health needs, the relevance of the app to the user, the quality of the app, and the behavior change techniques used to promote health [28]. All these factors are among those reported by a literature review that identified retention factors related to apps, such as feedback, appropriate reminders, and in-app support from peers or coaches [33]. Our findings also align with those of another systematic review that listed individualized reminders, user friendliness and technical stability of the apps, and personal support from health care professionals as intervention-related factors influencing adherence [3].

Push notifications and weekly SMS text messages inviting check-in were also used to influence user engagement [26], herein defined by number of check-ins. This is also consistent with the findings from a systematic review [34] that assessed 15 commercial apps for diabetes prevention and found that the app that included the notification features for activity tips, goal progress tips, goals adjustment, and completed goals had the highest engagement mean score (4.5 points out of 5). However, we argue that push notifications should not be a one-size-fits-all solution because marketing research suggests that among app users, just 50% accept push notifications from their favorite app, and 30% disable all push notifications [35].

In our review, user engagement was also influenced by the health goals that participants selected. This finding also aligns with that of a previous systematic review that further suggested that users could disengage at any time and re-engage at a later stage when needed. Thus, this feature might be particularly useful for addiction research targeting relapse prevention strategies [36].

**Barriers to User Engagement, Onboarding, or Adherence to the App**

From the studies included in this review, reasons for nonadoption included problems in app installation [24,30]; the use of other health apps that better suited participants’ needs and preferences; and other concerns, such as prioritizing COVID-19 over the condition addressed by the app [28]. Indeed, marketing can be influenced by competing health information targeting the same users. When people were concerned about contracting COVID-19 and seeking a pandemic-related app, they were less motivated to use an app being promoted to tackle other health issues [17,24]. This could lead to competition among mHealth apps for potential users or health care providers.

Other potential barriers to adoption included a lack of early active engagement with participants to verify that they could install the app, no direct contact with health care professionals,
and not enough interactive features. Indeed, as reported in another systematic review, being less informative and less interactive can lead to a very low engagement mean score [34].

The study by Roberts et al [28] reported that technical issues and concerns about data security reduced engagement. Similar concerns about data security with mHealth apps had been previously pointed out in the literature [37,38].

**Return on Investment in Marketing**

Although dependent upon the budget available, a decision threshold for the cost of marketing campaigns for mHealth apps to reach a certain level of engagement would be beneficial. According to a study on industry-specific Google benchmarks, for the health care sector overall, these costs amount to a CTR of 3.27% and a CPC of US $2.62 for Google Search and a CTR of 0.59% and a CPC of US $0.63 for the Google Display Network [39]. A similar study on industry-specific Facebook benchmarks produced a CTR of 0.83% and a CPC of US $1.32 for health care [40]. However, it is important to note that comparing CTR and CPC with industry-specific benchmarks for the entire health and medical field should not be the sole method of evaluating effectiveness because these benchmarks may be too broad [17].

**People Involved in Marketing mHealth Apps**

The most productive marketing team is a multitude of satisfied users championing the app on social media, encouraging their friends and colleagues to download it, and giving it 5-star reviews. Creating a strong user support system and feedback loop, regularly updating the app based on user feedback, and doing whatever it takes to keep users happy are the most important marketing tactics that can be deployed [12].

We found that in studies that investigated age and sex differences in engagement with apps, the age of research participants did not predict app engagement [24,25]. However, there were statistically significant differences in sex and app use, with more male participants using the apps in question than female participants, but not in the duration of app use [24,30]. This result is contradictory with that of another study that found that female sex positively influenced adherence [3]. Therefore, we could not draw conclusions on the effect of sex and age on app adherence and thus leave it to further research.

Marketers of mHealth apps should always consider the motivations of the app audience; for example, the study by Roberts et al [28] stated that apps promoting walking can be appealing to survivors of cancer. Similar findings were highlighted elsewhere [3]. In addition, consideration must be given to the timing of interventions intended to maintain engagement. Data should be collected to predict the moments that users will be available and receptive to in-app notifications.

**Limitations**

We have noted confusion surrounding the term *user engagement*. Engagement with digital health interventions and engagement with mHealth apps are not clearly separated. This study focuses on the latter, defined as a set of actions by a user within an mHealth app [18]. This differs from user engagement with digital health interventions, which is conceptualized in terms of both experience and behavior [41]. This confusion may explain why many of the studies identified during the database searches had to be excluded: they addressed outcomes related to changes in health behavior; for example, most of the excluded papers reported levels of user engagement with a health intervention, such as increased physical activity, but not engagement with an mHealth app. However, we recognize that these definitions are interlinked: research has shown that app engagement can motivate behavior change [25]. We have noted the same confusion with the term *user embedment*. In the 10 reviewed studies, only 1 (10%) referred to embedment as integrating a functionality within the app.

Most of the studies we reviewed (9/10, 90%) were conducted in high-income countries, with half being carried out in the United States (5/10, 50%). This may limit generalizability in low- and middle-income countries. Moreover, we only searched for papers written in French and English. We also note that the filters we applied with our search terms combination to avoid noise could have excluded some potentially useful papers. This could explain why the reviewed studies mostly reported research conducted in the United States and Europe (8/10, 80%).

Our findings may also be subject to observer bias [42] because in every reviewed study the research team members were involved in the diffusion processes. In some of the studies (2/10, 20%), participants received in-person physical assistance with app installation. This would be impossible for users in many settings. The generalizability of these findings is also limited by the fact that none of the included studies covered the dissemination of mHealth apps among health care personnel.

**Implications and Future Research**

**Integration of mHealth Apps Into Routine Clinical Practice**

None of the studies we reviewed aimed to address the embedment of mHealth apps in routine practice as part of their marketing strategy. To tackle the issue of the embedment of mHealth apps, researchers have proposed a framework for prescribing apps and outlined the key issues that need to be addressed to enable app dissemination in clinical care. This includes education and awareness, the creation of digital formularies, workflow and electronic health record integration, payment models, and patient or provider support [43]. As suggested by this framework, a starting point for the integration of mHealth apps into routine clinical practice would be education and awareness, meaning the promotion of mHealth apps, the aim of which would be to create a base of users downloading the apps because the number of app downloads and interactions over time also provides an indication of sustained uptake over time [11].

At this critical point of creating a user base, inspired by the synthesis of our findings, we offer a set of recommended uses of different channels (Table 4). This would first be applicable during the launching phase of the app and to lead users through their app onboarding stage. The use of these channels could change depending upon the objective of the marketing strategy.
### Table 4. Recommended uses of channels to promote mobile health apps.

<table>
<thead>
<tr>
<th>Channel and recommended use</th>
<th>Targeted people</th>
<th>Supportive marketing concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Email</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inform about the existence of the app, link for download, functionalities, app release notes and app updates, and general information regarding the app (developer, brand owner, and sponsor)</td>
<td>High-level users (influencers and decision-makers) and target base users (health care providers and day-to-day users of the app)</td>
<td>A/B testing</td>
</tr>
<tr>
<td>Give instructions on how to download and install and use the app</td>
<td>High-level users (influencers and decision-makers) and target base users (health care providers and day-to-day users of the app)</td>
<td>A/B testing</td>
</tr>
<tr>
<td>Provide technical support and answers to users’ questions; share user guide and tips</td>
<td>Effective users</td>
<td>A/B testing</td>
</tr>
<tr>
<td><strong>Social media</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inform about the existence of the app, and share the link for download</td>
<td>Potential target base users (health care providers and day-to-day users of the app)</td>
<td>A/B testing, incentivization, loyalty marketing, thought-leadership marketing, and content marketing</td>
</tr>
<tr>
<td>Engage in direct interactions to provide technical support and answers to users’ questions</td>
<td>Potential target base users (health care providers and day-to-day users of the app)</td>
<td>A/B testing, incentivization, loyalty marketing, thought-leadership marketing, and content marketing</td>
</tr>
<tr>
<td><strong>Television</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short promotional video report on the app and its functionalities, as well as 1- to 3-minute video spots with speech by high-level users (influencers and decision makers) recommending the apps</td>
<td>Potential target base users (health care users and day-to-day users of the app)</td>
<td>Incentivization and thought-leadership marketing</td>
</tr>
<tr>
<td><strong>Posters or flyers</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infographics and key text message to inform about the existence of the app and its main value; include a QR code and text to indicate link to download</td>
<td>Potential target base users (health care providers and day-to-day users of the app)</td>
<td>Predictive marketing and incentivization</td>
</tr>
<tr>
<td><strong>Face-to-face interaction: in-person training or meeting</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inform about the existence of the app, link for download, functionalities, app release notes and app updates, and general information regarding the app (developer, brand owner, and sponsor)</td>
<td>High-level users (influencers and decision-makers) and potential target base users (health care providers and day-to-day users of the app)</td>
<td>A/B testing</td>
</tr>
<tr>
<td>Give instructions on how to download, install, or use the app; share and explain user guide and discuss tips</td>
<td>High-level users (influencers and decision-makers) and potential target base users (health care providers and day-to-day users of the app)</td>
<td>A/B testing</td>
</tr>
<tr>
<td><strong>Face-to-face interaction: in-person ad hoc (unplanned) encounter</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engage in direct interactions to provide technical support and answers to users’ questions</td>
<td>Health care providers and day-to-day users of the app</td>
<td>Remarketing or retargeting and personalization</td>
</tr>
<tr>
<td>Sell the app (highlight its main value) and manage to install it on users' devices; explain how to use it and discuss tips</td>
<td>Health care providers and day-to-day users of the app</td>
<td>Remarketing or retargeting and personalization</td>
</tr>
</tbody>
</table>

Social media could be a beneficial entry point for motivating people to download an app, and human interaction is key during the engagement phase. Therefore, mHealth app promoters should provide users with training and support to start and continue using the apps. This can be done by maintaining communication through social media, including app-dedicated pages. At this point, content marketing—producing content that potential customers find useful—is valuable. It has been demonstrated that content marketing is highly effective in building a loyal user base and converting leads into customers [12].

It is important to note that although social media marketing also tends to attract people who are not the intended audience, communication through email requires a list of targeted email addresses. This entails contacting people directly and requesting their addresses or interacting with someone who will reveal potential users’ email addresses or share an app link with potential users; for instance, a hospital director may share information about an app with hospital staff or share an attendance list containing the email addresses of hospital staff.

Future public health campaigns targeting the parents of young children should consider crafting marketing messages for social media campaigns with animated GIF images as well as images of young children.

As time-varying push notifications have been shown to contribute to mHealth app user engagement, developers should...
interact with health care providers to implement this strategy. One approach to this is to apply mobile app attribution: the process of recording and measuring the actions of app users, such as installs, level completions, and in-app purchases [12].

**Future Research**

Finally, future research could be dedicated to developing a framework on how to disseminate mHealth apps. Such a framework, in addition to various marketing concepts presented in this review, should take into account additional considerations that are specific to mHealth apps, such as data confidentiality and privacy, and segment users on the marketing funnel [12] based on the best available evidence on engaging users with mHealth apps. One issue impeding the dissemination of apps that emerged in our study was the existence of competing apps. Some researchers have suggested that digital formularies or app libraries could help to address this. Digital formularies provide a short list of available apps, and providers could search these formularies and know what is available for a specific diagnosis or purpose [43]. Further research could explore and expand on the effectiveness of digital formularies as a dissemination channel for mHealth apps and the enablers of embedment of mHealth apps into routine practice. Finally, further research could aim to address the gap in identifying specific marketing strategies that would effectively drive the embedment of mHealth apps into routine practice.

**Conclusions**

The dissemination of mHealth apps takes place via face-to-face interactions, email, and printed posters and social media channels with diverse results. The effects of these strategies are reflected in download numbers and user engagement with mHealth apps. The results of this study will serve to guide future research and guide the marketing of mHealth apps for their routine use within the health sector.

The development of a framework for health care designers to promote their apps within health systems would be immensely beneficial. Such a framework would help systematize the dissemination of mHealth apps and guide the impact assessment of the dissemination strategies.

**Acknowledgments**

The study received funding from the Universitat Oberta de Catalunya, Barcelona, Spain, for proofreading and article processing fees.

**Conflicts of Interest**

None declared.

**Multimedia Appendix 1**

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.

[DOC File , 74 KB - mhealth_v12i1e50293_app1.doc ]

**Multimedia Appendix 2**

Search strategy.

[DOC File , 57 KB - mhealth_v12i1e50293_app2.doc ]

**References**


2. Be he@lthy, be mobile: personas toolkit. World Health Organization. 2019. URL: https://apps.who.int/iris/handle/10665/329947 [accessed 2023-04-22]


27. Rajani NB, Mastellos N, Filipidis FT. Impact of gamification on the self-efficacy and motivation to quit of smokers: observational study of two gamified smoking cessation mobile apps. JMIR Serious Games 2021 Apr 27;9(2):e27290 [FREE Full text] [doi: 10.2196/27290] [Medline: 33904824]


37. Tobias G, Spanier AB. Developing a mobile app (iGAM) to promote gingival health by professional monitoring of dental selfies: user-centered design approach. JMIR Mhealth Uhealth 2020 Aug 14;8(8):e19433 [FREE Full text] [doi: 10.2196/19433] [Medline: 32795985]


Abbreviations

CPC: cost per click
CTR: clickthrough rate
GIF: graphic interchange format
mHealth: mobile health
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SIGN: Scottish Intercollegiate Guidelines Network
UAC: Universal App Campaigns

©Henri Claude Moungui, Hugues Clotaire Nana-Djeunga, Che Frankline Anyiang, Mireia Cano, Jose Antonio Ruiz Postigo, Carme Carrion. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 05.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR

Please cite as:
Moungui HC, Nana-Djeunga HC, Anyiang CF, Cano M, Ruiz Postigo JA, Carrion C. Dissemination Strategies for mHealth Apps: Systematic Review. JMIR Mhealth Uhealth 2024;12:e50293
URL: https://mhealth.jmir.org/2024/1/e50293
doi:10.2196/50293
PMID:38180796
mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Review

**Effects of mHealth-Based Lifestyle Interventions on Gestational Diabetes Mellitus in Pregnant Women With Overweight and Obesity: Systematic Review and Meta-Analysis**

Yirong He\(^{1,2}\)*, MSc; Chuanya Huang\(^{1,2}\)*, PhD; Qiuyang He\(^{2,3}\)*, MSc; Shujuan Liao\(^{1,2}\), PhD; Biru Luo\(^{1,2}\), PhD

\(^{1}\)Department of Nursing, West China Second University Hospital, Sichuan University, Chengdu, Sichuan, China
\(^{2}\)Key Laboratory of Birth Defects and Related Diseases of Women and Children, Sichuan University, Ministry of Education, Chengdu, Sichuan, China
\(^{3}\)Department of Obstetric Nursing, West China Second University Hospital, Sichuan University, Chengdu, Sichuan, China

* these authors contributed equally

**Corresponding Author:**
Biru Luo, PhD
Department of Nursing
West China Second University Hospital
Sichuan University
#No. 20, Section 3, People’s South Road
Chengdu, Sichuan, 610041
China
Phone: 86 88570307
Email: luomr@scu.edu.cn

**Abstract**

**Background:** The increasing incidence of gestational diabetes mellitus (GDM) is a global health problem that is more likely to occur in pregnant women with overweight or obesity. Adhering to a healthy lifestyle is associated with a reduced risk of GDM. With the development of IT, mobile health (mHealth) interventions have become widely available in health care. However, there are no definitive conclusions on the effectiveness of mHealth-based lifestyle interventions in preventing GDM.

**Objective:** This study aims to evaluate the impact of mHealth-based lifestyle interventions on GDM and other pregnancy outcomes in pregnant women with overweight or obesity.

**Methods:** A systematic literature search was conducted in 5 English databases (MEDLINE, Embase, Web of Science, CENTRAL, and CINAHL) and 4 Chinese databases (CBM, CNKI, Vip, and Wanfang) to identify randomized controlled trials (RCTs) on the effectiveness of mHealth-based interventions for GDM from inception to January 10, 2023. In total, 2 authors independently screened the studies and extracted the data. The quality of the included studies was examined using the Cochrane risk-of-bias tool. Data synthesis was conducted using Review Manager (version 5.4; The Cochrane Collaboration).

**Results:** A total of 16 RCTs with 7351 participants were included in this study. The included studies were published between 2014 and 2021 and were conducted in China, the United States, Australia, New Zealand, the United Kingdom, Ireland, and Norway. The sample sizes of the studies ranged from 75 to 2202, and the duration of the mHealth-based lifestyle interventions ranged from 4 to 28 weeks. Compared with usual care, mHealth-based lifestyle interventions significantly reduced the incidence of GDM (odds ratio [OR] 0.74, 95% CI 0.56-0.96; \(P=0.03\); \(I^2=65\%\)), preterm birth (OR 0.65, 95% CI 0.48-0.87; \(P=0.004\); \(I^2=25\%\)), macrosomia (OR 0.59, 95% CI 0.40-0.87; \(P=0.008\); \(I^2=59\%\)), and gestational weight gain (mean difference=\(-1.12\) kg, 95% CI \(-1.44\) to \(-0.80\); \(P<0.001\); \(I^2=43\%\)). The subgroup analysis showed that interventions delivered via apps (OR 0.55, 95% CI 0.37-0.83; \(P=0.004\); \(I^2=44\%\)), provided by obstetricians (OR 0.69, 95% CI 0.51-0.93; \(P=0.02\); \(I^2=60\%\)), and targeted at Asian populations (OR 0.44, 95% CI 0.34-0.58; \(P<0.001\); \(I^2=0\%\)) and that used the International Association of Diabetes and Pregnancy Study Groups diagnostic criteria (OR 0.58, 95% CI 0.39-0.86; \(P=0.007\); \(I^2=69\%\)) showed a statistically significant reduction in the risk of GDM.

**Conclusions:** mHealth-based lifestyle interventions had a favorable impact on the prevention of GDM in pregnant women with overweight and obesity. Future studies need to further explore the potential of mHealth-based interventions for GDM through better design and more rigorous large-scale RCTs.
Introduction

Background

Gestational diabetes mellitus (GDM) is defined as a carbohydrate intolerance of varying severity with onset or first recognition during pregnancy [1]. GDM is one of the most common obstetric complications, with the prevalence varying from 7.5% to 27% in different areas [2]. GDM is associated with substantial adverse pregnancy outcomes such as neonatal hypoglycemia and macrosomia as well as long-term metabolic risk in pregnant women and their offspring [3]. Risk factors for GDM include age, ethnicity, overweight or obesity, and family history of diabetes [4]. As obesity becomes a global epidemic, perinatal overweight and obesity are also a growing concern [5]. In recent decades, the rates of overweight and obesity among women of reproductive age have increased dramatically [6]. In the United States, 60% of women are overweight or obese during pregnancy compared with 30% in Europe and 10% in Asia [7]. In pregnant women with overweight and obesity, it is estimated that the risk of GDM is more than twice that of other pregnant women [8]. Unhealthy lifestyle behaviors are a critical factor that affects overweight and obesity during the perinatal period [9]. Numerous randomized controlled trials (RCTs) have attempted to reduce the incidence of GDM through diet [10], physical activity [11], or combination interventions [12]. However, when lifestyle interventions are provided in a personalized way, they are commonly expensive and may lack scalability from the perspective of public health [13]. Therefore, there is an urgent need for effective real-world solutions that address the demand of pregnant women seeking personalized support, information, and guidance to help reduce the risk of GDM.

As IT advances and more people use electronic devices, mobile health (mHealth) has developed rapidly [14]. mHealth is commonly defined as health care services provided by health care professionals using telecommunications technology [15]. mHealth has been applied in many areas of perinatal health care, with medical staff providing health care services to pregnant women through multimedia services such as mobile apps, software, SMS text messages, email, web-based diaries, and integrated systems combining various components of digital communications technologies [16]. Studies have shown that mHealth care can reduce gestational weight gain (GWG); improve pregnant women’s health behaviors; and reduce the number of medical visits, thereby decreasing financial burden [13]. However, the impact of mHealth interventions on pregnancy outcomes in women with overweight or obesity is uncertain. More and more systematic reviews have found an effect of lifestyle interventions based on mHealth technology on diabetes prevention among adults with overweight and obesity [17,18]; however, little is known about their effectiveness in the perinatal population. Several previous reviews have attempted to synthesize the results of mHealth-based lifestyle programs for pregnant women, but none have evaluated the quantitative effects of these programs [19,20]. There is still no consensus on the impact of mHealth lifestyle interventions on preventing GDM and other pregnancy outcomes in women with overweight or obesity.

Objectives

Therefore, we conducted a systematic review and meta-analysis of RCTs to summarize mHealth interventions delivered in different ways and assess the effectiveness of mHealth-based lifestyle interventions in reducing the risk of GDM.

Methods

This systematic review and meta-analysis follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [21] and is presented in Multimedia Appendix 1. It was registered in PROSPERO on November 18, 2021, with registration number CRD42021286995.

Search Strategy

A systematic literature search was conducted in 5 English databases (MEDLINE, Embase, Web of Science, CENTRAL, and CINAHL) and 4 Chinese databases (CBM, CNKI, Vip, and Wanfang) for studies published in English and Chinese. A systematic search was conducted combining Medical Subject Headings (MeSH) and free-text terms, including overweight (MeSH) OR obesity (MeSH); pregnant* OR prenatal OR antenatal OR maternal OR gestational; gestational diabetes mellitus OR gestational diabetes OR GDM OR T2DM OR impaired fasting glucose OR impaired glucose tolerance; telemedicine (MeSH) OR telerehabilitation (MeSH) OR telecommunications (MeSH) OR electronic health OR eHealth OR ehealth OR mHealth OR mobile health OR telecare OR eHealthcare OR mcare OR telemonitor* OR telerehab* OR telemangement OR mobile communication OR remote consult OR mobile technolog* OR mobile devic* OR mobile app* OR internet (MeSH) OR web* OR online OR smartphone (MeSH) OR telephone (MeSH) OR cell phone (MeSH) OR cellular phone (MeSH) OR mobile phone OR messag* OR SMS. The searches were unlimited by time up to January 10, 2023, and were limited to RCTs. The full details of the search strategy for each database are provided in Multimedia Appendix 2. We complemented this strategy by manually searching the reference lists of included studies and related reviews.

Inclusion and Exclusion Criteria

In total, 2 researchers (YH and CH) independently screened the titles and abstracts and selected the studies in accordance with...
the eligibility criteria. Any disagreement was resolved through discussion with a third researcher (QH). Studies were included if they met the following inclusion criteria: (1) pregnant women with overweight (BMI ≥ 25 kg/m²) or obesity (BMI ≥ 30 kg/m²); (2) mHealth interventions including pregnancy nutrition, physical activity, weight management, and health behavior education delivered via the internet, websites, telephone, app, SMS text message, email, or other types of information and communications technologies (intervention); (3) usual care, routine care, conventional care, or standard care without mHealth (comparison); (4) incidence of GDM, postpartum hemorrhage, preterm birth, cesarean delivery, pregnancy-induced hypertension, macrosomia, neonatal gestational age, and GWG (outcome); (5) RCTs (study design); and (6) English or Chinese (language). There were no restrictions regarding the year of publication. We excluded studies that (1) included women with either type 1 or type 2 diabetes mellitus before pregnancy or with existing GDM; (2) contained incomplete data; (3) lacked data related to GDM; or (4) were study protocols, comments, editorials, and conference abstracts.

**Study Selection and Data Extraction**

The reference management program EndNote X9 (Clarivate Analytics) was used for data management. The studies were imported into EndNote after an extensive database search. After removing duplicates, 2 authors independently reviewed the titles and abstracts according to the eligibility criteria. Disagreements were resolved through discussion or consultation with a third researcher. Data were extracted by an independent researcher (YH) using the predesigned data collection forms, and the extracted data were verified by a second researcher (CH). Disagreements were resolved through consensus. The extracted data included the authors, year, country, study design, sample size, participant characteristics, intervention, control, GDM criteria, and outcomes.

**Quality Assessment**

In total, 2 researchers (YH and CH) independently assessed the studies’ risk of bias in accordance with the *Cochrane Handbook for Systematic Reviews of Interventions*. This tool consists of 6 items: selection bias, performance bias, detection bias, attrition bias, reporting bias, and other biases. Each item was judged as “low risk,” “high risk,” or “unclear risk.” A third researcher (QH) was available if there was a difference in opinion in assessing the risk of bias.

**Data Synthesis and Analysis**

Statistical analyses were conducted using Review Manager (version 5.4; The Cochrane Collaboration). The overall effect difference was considered statistically significant if the 2-tailed P value was < 0.05. Continuous variables were presented using the mean difference, and dichotomous variables were described using the odds ratio (OR) with a 95% CI. Heterogeneity was assessed using the Cochran Q test and the I² statistic. We considered ≤ 25%, 25% to 50%, 50% to 75%, and > 75% as low, moderate, high, and severe heterogeneity between the studies, respectively. If I² ≤ 50% and the P value was > 0.10, a fixed-effects model was considered; otherwise, a random-effects model was used. The sources of heterogeneity were explored using subgroup analysis. A funnel plot was constructed to check for potential publication bias.

**Results**

**Study Selection**

Figure 1 shows a PRISMA flowchart of the study selection process. A total of 1725 records were retrieved from 9 electronic databases. After removing duplicates (393/1725, 22.78% of the studies), 1332 studies were included for screening. Of these 1332 studies, we then excluded 1234 (92.64%) based on the relevance of the abstract and title, and the remaining 98 (7.36%) studies were assessed for eligibility. After full-text review, 84% (82/98) of the studies were excluded for the reasons outlined in Figure 1. Finally, 16 studies were included in the review and meta-analysis.
Characteristics of the Included Studies

Study Characteristics

The characteristics of the included studies are shown in Table 1. The 16 studies included in this review and meta-analysis were RCTs, 4 (25%) of which were multicenter RCTs. The included studies were published between 2014 and 2021 and were conducted in China (8/16, 50%) [22-29], the United States (2/16, 12%) [30,31], Australia (2/16, 12%) [32,33], New Zealand (1/16, 6%) [34], the United Kingdom (1/16, 6%) [35], Ireland (1/16, 6%) [36], and Norway (1/16, 6%) [37]. A total of 7351 participants were included in the studies, and the sample sizes varied from 74 [34] to 2153 [32]. In total, 38% (6/16) of the studies included pregnant women with obesity [23,24,26,28,33,35], and 62% (10/16) included both pregnant women with overweight and pregnant women with obesity [22,25,27,29,30,31,32,34,36,37]. A total of 81% (13/16) of the articles reported diagnostic criteria for GDM. The incidence of GDM was determined by screening pregnant women using an oral glucose tolerance test. The diagnostic criteria for GDM were inconsistent among the included studies. In total, 56% (9/16) of the RCTs used the International Association of Diabetes and Pregnancy Study Groups (IADPSG) criteria (2010) [22,23,25,27,29,34-36], 12% (2/16) used the World Health Organization 2013 criteria [33,37], 6% (1/16) used the Australasian Diabetes in Pregnancy Society criteria [32], and 6% (1/16) of the RCTs used the Carpenter-Coustan criteria [30]. All the RCTs (16/16, 100%) reported more than 1 outcome.
<table>
<thead>
<tr>
<th>Author, year, and country</th>
<th>Study design</th>
<th>Sample size</th>
<th>Participant characteristics</th>
<th>Intervention</th>
<th>Control</th>
<th>GDM criteria</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dodd et al [32], 2014, Australia</td>
<td>Multicenter RCT</td>
<td>2153; IG: 1080; CG: 1073 (1:1)</td>
<td>Women with a singleton pregnancy between 10 and 20 wk of gestation and a BMI of ≥25 kg/m²</td>
<td>Phone, in person</td>
<td>Standard care</td>
<td>ADPSC®</td>
<td>LGA¹, macrosomia, PIH², pre-eclampsia, GDM, PTB³, CS¹, and PPH¹</td>
</tr>
<tr>
<td>Sagedal et al [37], 2015, Norway</td>
<td>RCT</td>
<td>591; IG: 296; CG: 295 (1:1)</td>
<td>Women with a singleton pregnancy at ≤20 wk of gestation who had a prepregnancy BMI of ≥25 kg/m²</td>
<td>Phone, website, in person</td>
<td>Standard care</td>
<td>WHOk</td>
<td>GWG¹, GDM, LGA, SGA⁵, pre-eclampsia, PTB, PPH, and NW⁶</td>
</tr>
<tr>
<td>Seneviratne et al [34], 2015, New Zealand</td>
<td>RCT</td>
<td>74; IG: 37; CG: 37 (1:1)</td>
<td>Women aged 18-40 y with a BMI of ≥25 kg/m² and a singleton pregnancy of &lt;20 wk of gestation</td>
<td>Software, device, home-based, in person</td>
<td>Usual care</td>
<td>IADPSG⁰</td>
<td>NW, hypoglycemia, GWG, GDM, PIH, PTB, CS, PPH, MVPA⁷, and pre-eclampsia</td>
</tr>
<tr>
<td>Poston et al [35], 2015, United Kingdom</td>
<td>Multicenter RCT</td>
<td>1280; IG: 629; CG: 651 (1:1)</td>
<td>Women aged ≥16 y with a BMI of ≥30 kg/m² and a singleton pregnancy between 15 and 18 wk plus 6 d of gestation</td>
<td>Phone, email, DVD, pedometer, logbook, in person</td>
<td>Standard care</td>
<td>IADPSG</td>
<td>GDM, FBG⁵, pre-eclampsia, CS, PPH, GWG, HOMA-IR⁸, GL⁹, GI⁸, fat, MVPA, LGA, and NW</td>
</tr>
<tr>
<td>Simmons et al [33], 2017, Australia</td>
<td>Multicenter RCT</td>
<td>192; IG: 92; CG: 100 (1:1)</td>
<td>Women with a BMI of &gt;29 kg/m², ≤19 (~6 to ~6) d of gestation, and a singleton pregnancy</td>
<td>Email, phone, pedometer, device, in person</td>
<td>Usual care</td>
<td>WHO</td>
<td>GDM, FBG, HOMA-IR, insulin, NW, LGA, SGA, MVPa, and sugar intake</td>
</tr>
<tr>
<td>Chen [22], 2017, China</td>
<td>RCT</td>
<td>160; IG: 80; CG: 80 (1:1)</td>
<td>Women with a prepregnancy BMI of ≥24 kg/m² and a singleton pregnancy between 8 and 12 wk of gestation</td>
<td>Mobile apps, SMS text messages, in person</td>
<td>Usual care</td>
<td>IADPSG</td>
<td>BMI, GDM, CS, and NW</td>
</tr>
<tr>
<td>Kennelly et al [36], 2018, Ireland</td>
<td>RCT</td>
<td>498; IG: 241; CG: 257 (1:1)</td>
<td>Singleton pregnant women between 10 and 15 wk of gestation with a BMI between 25.0 and 39.9 kg/m²</td>
<td>Mobile app, email, in person</td>
<td>Usual care</td>
<td>IADPSG</td>
<td>GDM, GWG, GL, GL PIH, CS, HOMA-IR, NW, LGA, and SGA</td>
</tr>
<tr>
<td>Li [23], 2018, China</td>
<td>RCT</td>
<td>1000; IG: 500; CG: 500 (1:1)</td>
<td>Women with a BMI of ≥28 kg/m², aged ≥18 y, and with a singleton pregnancy between 8 and 12 wk of gestation</td>
<td>Phone, in person</td>
<td>Usual care</td>
<td>IADPSG</td>
<td>GDM, GWG, NW, and macrosomia</td>
</tr>
<tr>
<td>Tang et al [24], 2019, China</td>
<td>RCT</td>
<td>136; IG: 68; CG: 68 (1:1)</td>
<td>Pregnant women with a BMI of ≥28 kg/m², aged ≥18 y</td>
<td>Mobile app, software</td>
<td>Standard care</td>
<td>N/A⁸</td>
<td>Macrosomia, CS, hypoglycemia, and GDM</td>
</tr>
<tr>
<td>Ferrara et al [30], 2020, United States</td>
<td>Multicenter RCT</td>
<td>389; IG: 195; CG: 194 (1:1)</td>
<td>Women at 8-15 wk of gestation with singletons, with a prepregnancy BMI of 25-40 kg/m², and aged ≥18 y</td>
<td>Phone, device, in person</td>
<td>Usual care</td>
<td>Carpenter and Cousins criteria</td>
<td>GWG, caloric intake, MVPA, FBG, HOMA-IR, GDM, PIH, pre-eclampsia, CS, NW, and macrosomia</td>
</tr>
<tr>
<td>Author, year, and country</td>
<td>Study design</td>
<td>Sample size</td>
<td>Participant characteristics</td>
<td>Intervention</td>
<td>Control</td>
<td>GDM criteria</td>
<td>Outcomes</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------</td>
<td>-------------</td>
<td>-----------------------------</td>
<td>--------------</td>
<td>---------</td>
<td>--------------</td>
<td>----------</td>
</tr>
<tr>
<td>Cao [25], 2020, China</td>
<td>RCT</td>
<td>96; IG: 48; CG: 48 (1:1)</td>
<td>Women aged 22-38 y with a BMI of ≥24 kg/m² and a singleton pregnancy between 8 and 12 wk of gestation</td>
<td>Mobile apps, SMS text messages, in person</td>
<td>Usual care</td>
<td>IADPSG</td>
<td>FBG, GDM, BMI, CS, NW, and LGA</td>
</tr>
<tr>
<td>Wu and Guang [26], 2020, China</td>
<td>RCT</td>
<td>140; IG: 70; CG: 70 (1:1)</td>
<td>Pregnant women with a BMI of ≥28 kg/m² aged &gt;18 y</td>
<td>Mobile app, in person</td>
<td>Usual care</td>
<td>N/A</td>
<td>BMI, NW, GDM, PIH, CS, PTB, macrosomia, and PPH</td>
</tr>
<tr>
<td>Liu et al [31], 2021, United States</td>
<td>RCT</td>
<td>217; IG: 112; CG: 105 (1:1)</td>
<td>Pregnant women aged 18-44 y with a gestational age of ≤16 wk and a prepregnancy BMI of ≥25 kg/m²</td>
<td>Phone, software, SMS text messages, in person</td>
<td>Standard care</td>
<td>N/A</td>
<td>GWG, PTB, LBW, macrosomia, SGA, GDM, PIH, CS, and NW</td>
</tr>
<tr>
<td>Zhou et al [27], 2021, China</td>
<td>RCT</td>
<td>104; IG: 52; CG: 52 (1:1)</td>
<td>Singleton pregnant women with a BMI of ≥24 kg/m² and aged &gt;18 y</td>
<td>Mobile app, in person</td>
<td>Usual care</td>
<td>IADPSG</td>
<td>FBG, 2-hour BG, GDM, CS, PTB, and macrosomia</td>
</tr>
<tr>
<td>Kang and Sung [28], 2021, China</td>
<td>RCT</td>
<td>106; IG: 53; CG: 53 (1:1)</td>
<td>Women with a BMI of ≥28 kg/m², aged &gt;18 y, and with a singleton pregnancy between 12 and 20 wk of gestation</td>
<td>Mobile app, in person</td>
<td>Usual care</td>
<td>IADPSG</td>
<td>GDM, FBG, 2-hour BG, HbA₁c, CS, PTB, and macrosomia</td>
</tr>
<tr>
<td>Ding et al [29], 2021, China</td>
<td>RCT</td>
<td>215; IG: 104; CG: 111 (1:1)</td>
<td>Pregnant women with a BMI of ≥24 kg/m² at the onset of pregnancy, aged &lt;35 y, and at &lt;12 wk of gestation</td>
<td>Mobile app, in person</td>
<td>Usual care</td>
<td>IADPSG</td>
<td>Energy intake, GDM, FBG, GWG, CS, PTB, PIH, pre-eclampsia, PPH, NW, and macrosomia</td>
</tr>
</tbody>
</table>
**Characteristics of the mHealth Interventions**

The details of the mHealth interventions are presented in Multimedia Appendix 3 [22-37]. The interventions were divided into 3 groups: exercise only, diet only, and mixed interventions. Regarding the delivery mode, 12% (2/16) of the studies [23,32] provided mHealth interventions through phone counseling, 25% (4/16) of the studies [26-29] only provided the interventions through mobile apps, and 62% (10/16) of the studies [22,24,25,30,31,33-37] adopted a combination of methods. The duration of the mHealth interventions ranged from 4 to 28 weeks. The mHealth interventions were delivered by various types of personnel, including a single provider in 12% (2/16) of the studies [29,32] and multidisciplinary prenatal care providers in 88% (14/16) of the studies [22-28,30,31,33-37].

**Characteristics of the Comparators**

Most studies (14/16, 88%) briefly described the control group, which generally included regular maternity visits and education on diet and exercise during pregnancy. In the control groups, all studies provided standard or usual care for participants, which was based on the different countries’ perinatal practices and local hospital guidelines.

**Risk of Bias**

The Cochrane risk-of-bias tool was used to assess the risk of bias in each study, and the results are shown in Figure 2 [22-37]. A total of 75% (12/16) of the studies described the details of the randomization scheme, and the risk of random sequence generation was low. Only 31% (5/16) of the studies reported allocation concealment, with a low risk of selection bias. Blinding of participants or staff was difficult because of the nature of the mHealth interventions; thus, 25% (4/16) of the studies, which did not blind the participants or staff, were rated as having a high risk of bias, and 62% (10/16) of the studies were rated as unclear regarding the risk of bias. A total of 12% (2/16) of the studies did not report the outcome assessment and were rated as having a high risk of detection bias. In total, 81% (13/16) of the included studies described details of participant dropout and had a low risk of bias. A total of 56% (9/16) of the studies followed preregistered protocols for analysis and outcome reporting, and the remaining 44% (7/16) did not provide information on published protocols or registrations and had an unclear risk of reporting bias. The funnel plot showed no publication bias in any of the included studies (Multimedia Appendix 4).

---

**Figure 2.** Results of the risk-of-bias assessment of the included studies [22-37].

<table>
<thead>
<tr>
<th>Study Reference</th>
<th>Random sequence generation</th>
<th>Allocation concealment</th>
<th>Blinding of participants and personnel</th>
<th>Performance bias</th>
<th>Detection bias</th>
<th>Selective outcome reporting</th>
<th>Publication bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>He et al, 2020</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Chen, 2017</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Ding et al, 2021</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Dodd et al, 2014</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Ferrara et al, 2026</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Kang and Sang, 2021</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Kennedy et al, 2018</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Li et al, 2015</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Liu et al, 2021</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Poon et al, 2015</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Dzepedal et al, 2017</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Seneviratne et al, 2015</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Simmons et al, 2017</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Tang et al, 2016</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Wu and Ouang, 2020</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Zhou et al, 2021</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
Meta-Analysis Results

Primary Outcome: the Incidence of GDM

The pooled analysis of the 16 RCTs with 7351 participants showed that the mHealth interventions significantly decreased the incidence of GDM in women with overweight and obesity (OR 0.74, 95% CI 0.56-0.96; P=.03; $I^2=65$%; Figure 3 [22-37]).

Table 2 provides the results of the subgroup analyses, and the forest plots are presented in Multimedia Appendix 5 [22-37]. The subgroup analysis conducted based on the different interventions showed that diet (OR 0.61, 95% CI 0.29-1.28; $P=.19$; $I^2=61$%), exercise (OR 1.11, 95% CI 0.57-2.17; $P=.76$; $I^2=0$%), or a combination of the 2 (OR 0.71, 95% CI 0.51-0.99; $P=.05$; $I^2=74$%) were not statistically significantly associated with reduced risk of GDM. When compared with the control groups, app-based mHealth interventions were significantly effective in reducing GDM (OR 0.55, 95% CI 0.37-0.83; $P=.004$; $I^2=44$%). Another subgroup analysis conducted based on different providers indicated that interventions provided by obstetricians and nurses were effective in reducing GDM (OR 0.69, 95% CI 0.51-0.93; $P=.02$; $I^2=60$%). Interventions targeted at Asian populations (OR 0.44, 95% CI 0.34-0.58; $P<.001$; $I^2=0$%) and using the IADPSG as a diagnostic criterion for GDM showed a reduction in GDM compared with the control groups (OR 0.58, 95% CI 0.39-0.86; $P=.007$; $I^2=69$%).

Figure 3. Forest plot of the effect of mobile health interventions on gestational diabetes mellitus [22-37]. IV: inverse variance.
## Table 2. Subgroup analyses of the included studies (n=16).

<table>
<thead>
<tr>
<th>Subgroup analysis</th>
<th>References</th>
<th>Studies, n (%)</th>
<th>Sample size</th>
<th>Effect estimates (95% CI)</th>
<th>P value</th>
<th>I² (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intervention</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Diet              | • Tang et al [24]  
• Zhou et al [27]  
• Kang and Sung [28]  
• Kennelly et al [36] | 4 (25)  
844 | 0.61 (0.29-1.28) | .19 | 61 |
| Exercise          | • Ferrara et al [30]  
• Seneviratne et al [34] | 2 (12)  
463 | 1.11 (0.57-2.17) | .76 | 0 |
| Both              | • Chen [22]  
• Li [23]  
• Cao [25]  
• Wu and Guang [26]  
• Ding et al [29]  
• Liu et al [31]  
• Dodd et al [32]  
• Simmons et al [33]  
• Poston et al [35]  
• Sagedal et al [37] | 10 (62)  
6044 | 0.71 (0.51-0.99) | .05 | 74 |
| **mHealth® technology** |            |                |             |                           |         |        |
| Phone             | • Ferrara et al [30]  
• Simmons et al [33]  
• Poston et al [35]  
• Sagedal et al [37]  
• Liu et al [31]  
• Dodd et al [32] | 6 (38)  
4822 | 1.08 (0.92-1.26) | .37 | 1 |
| App               | • Chen [22]  
• Tang et al [24]  
• Cao [25]  
• Wu and Guang [26]  
• Zhou et al [27]  
• Kang and Sung [28]  
• Ding et al [29]  
• Kennelly et al [36] | 8 (50)  
1455 | 0.55 (0.37-0.83) | .004 | 44 |
| Computer          | • Li [23]  
• Seneviratne et al [34] | 2 (12)  
1074 | 0.72 (0.14-3.74) | .69 | 71 |
| **Provider**      |            |                |             |                           |         |        |
| Dietitian         | • Ding et al [29]  
• Dodd et al [32] | 2 (12)  
2368 | 0.84 (0.36-2.00) | .70 | 86 |
| Exercise physiologists | • Seneviratne et al [34] | 1 (6)  
74 | 2.12 (0.36-12.36) | .40 | N/A |
| Obstetricians     | • Chen [22]  
• Li [23]  
• Tang et al [24]  
• Cao [25]  
• Wu and Guang [26]  
• Zhou et al [27]  
• Kang and Sung [28]  
• Ferrara et al [30]  
• Liu et al [31]  
• Simmons et al [33]  
• Poston et al [35]  
• Kennelly et al [36]  
• Sagedal et al [37] | 13 (81)  
4909 | 0.69 (0.51-0.93) | .02 | 60 |

### Duration

---

*Note: P values are not provided for mHealth® technology and Provider subgroups due to insufficient data.*

---

https://mhealth.jmir.org/2024/1/49373  
JMIR Mhealth Uhealth 2024 | vol. 12 | e49373 | p.126  
(page number not for citation purposes)
<table>
<thead>
<tr>
<th>Subgroup analysis</th>
<th>References</th>
<th>Studies, n (%)</th>
<th>Sample size</th>
<th>Effect estimates (95% CI)</th>
<th>P value</th>
<th>$I^2$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short term</td>
<td>• Zhou et al [27]</td>
<td>3 (19)</td>
<td>1601</td>
<td>0.60 (0.26-1.39)</td>
<td>.24</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>• Liu et al [31]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Poston et al [35]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium to long term</td>
<td>• Chen [22]</td>
<td>11 (69)</td>
<td>5518</td>
<td>0.77 (0.53-1.10)</td>
<td>.15</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>• Li [23]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Wu and Guang [26]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Kang and Sung [28]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Ding et al [29]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Ferrara et al [30]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Dodd et al [32]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Simmons et al [33]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Seneviratne et al [34]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Kennelly et al [36]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Sagedal et al [37]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long term</td>
<td>• Tang et al [24]</td>
<td>2 (12)</td>
<td>232</td>
<td>0.62 (0.35-1.12)</td>
<td>.11</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>• Cao [25]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Ethnicity**

<table>
<thead>
<tr>
<th></th>
<th>References</th>
<th>Studies, n (%)</th>
<th>Sample size</th>
<th>Effect estimates (95% CI)</th>
<th>P value</th>
<th>$I^2$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian</td>
<td>• Chen [22]</td>
<td>8 (50)</td>
<td>1957</td>
<td>0.44 (0.34-0.58)</td>
<td>&lt;.001</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>• Li [23]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Tang et al [24]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Cao [25]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Wu and Guang [26]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Zhou et al [27]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Kang and Sung [28]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Ding et al [29]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>• Ferrara et al [30]</td>
<td>8 (50)</td>
<td>5394</td>
<td>1.09 (0.93-1.26)</td>
<td>.29</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>• Liu et al [31]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Dodd et al [32]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Simmons et al [33]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Seneviratne et al [34]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Poston et al [35]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Kennelly et al [36]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Sagedal et al [37]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**GDM<sup>c</sup> diagnostic criteria**
Subgroup analysis | References | Studies, n (%) | Sample size | Effect estimates (95% CI) | P value | $I^2$ (%) |
--- | --- | --- | --- | --- | --- | --- |
IADPSG<sup>d</sup> | • Chen [22]  
• Li [23]  
• Cao [25]  
• Zhou et al [27]  
• Kang and Sung [28]  
• Ding et al [29]  
• Seneviratne et al [34]  
• Poston et al [35]  
• Kennelly et al [36] | 9 (56) | 3533 | 0.58 (0.39-0.86) | .007 | 69 |
WHO<sup>e</sup> | • Simmons et al [33]  
• Sagedal et al [37] | 2 (12) | 783 | 1.20 (0.78-1.86) | .41 | 0 |
ADPSC<sup>f</sup> | • Dodd et al [32]  
• Ferrara et al [30] | 1 (6) | 2153 | 1.26 (0.98-1.63) | .08 | N/A |
Carpenter-Coustan criteria | | | | | | |

<sup>a</sup>mHealth: mobile health.  
<sup>b</sup>N/A: not applicable.  
<sup>c</sup>GDM: gestational diabetes mellitus.  
<sup>d</sup>IADPSG: International Association of Diabetes and Pregnancy Study Groups.  
<sup>e</sup>WHO: World Health Organization.  
<sup>f</sup>ADPSC: Australasian Diabetes in Pregnancy Society criteria.

**Secondary Outcomes: Maternal and Neonatal Outcomes**

**Effect on GWG**

A total of 50% (8/16) of the studies examined the effects of mHealth-based interventions on GWG [23,29-31,34-37]. There were statistically significant differences in decreases in GWG between the mHealth intervention groups and the control groups (mean difference=−1.12 kg, 95% CI −1.44 to −0.80; P<.001; $I^2$=43%; Figure 4 [23, 29-31,34-37]). The effects of the mHealth-based lifestyle interventions on maternal and neonatal outcomes are shown in Table 3.

**Figure 4.** Forest plot of the effect of mobile health interventions on gestational weight gain [23, 29-31,34-37]. IV: inverse variance.
### Table 3. Effectiveness of mobile health–based lifestyle interventions on maternal and neonatal outcomes.

<table>
<thead>
<tr>
<th>Maternal and neonatal outcomes</th>
<th>References</th>
<th>Studies, n (%)</th>
<th>Sample size</th>
<th>Statistical method</th>
<th>Effect estimates (95% CI)</th>
<th>P value</th>
</tr>
</thead>
</table>
| **Postpartum hemorrhage (%)** | • Wu and Guang [26]  
• Ding et al [29]  
• Dodd et al [32]  
• Seneviratne et al [34]  
• Porton et al [35]  
• Sagedal et al [37] | 6 (38) | 4664 | Odds ratio (IV<sup>a</sup>; fixed) | 1.06 (0.90 to 1.24) | .49 |
| **Cesarean delivery (%)**     | • Chen [22]  
• Tang et al [24]  
• Cao [25]  
• Wu and Guang [26]  
• Ferrara et al [30]  
• Dodd et al [32]  
• Seneviratne et al [34]  
• Porton et al [35]  
• Kennelly et al [36] | 9 (56) | 5224 | Odds ratio (IV; fixed) | 0.89 (0.79 to 1.00) | .05 |
| **Preeclampsia or PIH<sup>b</sup> (%)** | • Wu and Guang [26]  
• Ding et al [29]  
• Ferrara et al [30]  
• Liu et al [31]  
• Dodd et al [32]  
• Seneviratne et al [34]  
• Porton et al [35]  
• Kennelly et al [36]  
• Sagedal et al [37] | 9 (56) | 5829 | Odds ratio (IV; fixed) | 0.96 (0.80 to 1.15) | .63 |
| **GWG<sup>c</sup> (kg)**      | • Li [23]  
• Ding et al [29]  
• Ferrara et al [30]  
• Liu et al [31]  
• Seneviratne et al [34]  
• Porton et al [35]  
• Kennelly et al [36]  
• Sagedal et al [37] | 8 (50) | 4133 | Mean difference (IV; fixed) | −1.12 (−1.44 to −0.80) | <.001 |
| **Neonatal outcomes**         | • Wu and Guang [26]  
• Zhou et al [27]  
• Kang and Sung [28]  
• Ding et al [29]  
• Liu et al [31]  
• Dodd et al [32]  
• Seneviratne et al [34] |
| **Preterm birth (%)**         | • Li [23]  
• Tang et al [24]  
• Wu and Guang [26]  
• Zhou et al [27]  
• Kang and Sung [28]  
• Ding et al [29]  
• Liu et al [31]  
• Dodd et al [32]  
• Seneviratne et al [34] | 7 (44) | 2998 | Odds ratio (IV; fixed) | 0.65 (0.48 to 0.87) | .004 |
| **Macrosomia (%)**            | • Li [23]  
• Tang et al [24]  
• Wu and Guang [26]  
• Zhou et al [27]  
• Kang and Sung [28]  
• Ding et al [29]  
• Ferrara et al [30]  
• Liu et al [31]  
• Dodd et al [32]  
• Seneviratne et al [34] | 9 (56) | 4449 | Odds ratio (IV; random) | 0.59 (0.40 to 0.87) | .008 |
| **LGA<sup>d</sup> >90th percentile (%)** | • Dodd et al [32]  
• Simmons et al [33]  
• Porton et al [35]  
• Kennelly et al [36]  
• Sagedal et al [37] | 5 (31) | 4966 | Odds ratio (IV; random) | 0.80 (0.60 to 1.06) | .12 |
Effect on Cesarean Delivery
A total of 56% (9/16) of the studies explored the effects of mHealth-based interventions on cesarean delivery [22,24-26,30,32-34,36]. The pooled results showed no significant difference between the mHealth intervention and control groups (OR 0.89, 95% CI 0.79-1.00; \( P = .05; I^2 = 39\% \)). Forest plots of maternal and neonatal outcomes are shown in Multimedia Appendix 6 [22-37].

Effect on Pregnancy-Induced Hypertension
A total of 56% (9/16) of the studies assessed the effects of mHealth-based interventions on the risk of pregnancy-induced hypertension [26,29-32,34-37]. Compared with the control groups, the groups with mHealth-based interventions did not show statistically significant decreases in pregnancy-induced hypertension (OR 0.96, 95% CI 0.80-1.15; \( P = .63; I^2 = 31\% \)).

Effect on Postpartum Hemorrhage
A total of 38% (6/16) of the studies assessed the effects of mHealth-based lifestyle interventions on the risk of postpartum hemorrhage [26,29,31,32,34-37]. There were no statistically significant decreases in postpartum hemorrhage between the groups with mHealth-based interventions and the control groups (OR 1.06, 95% CI 0.90-1.24; \( P = .49; I^2 = 17\% \)).

Effect on Preterm Birth
A total of 44% (7/16) of the studies explored the effects of mHealth-based interventions on the risk of preterm birth [26-29,31,32,34]. Statistically significant differences in decreases in preterm birth were found between the mHealth intervention groups and the control groups (OR 0.65, 95% CI 0.48-0.87; \( P = .004; I^2 = 25\% \); Figure 5 [26-29,31,32,34]).

Effect on Macrosomia
A total of 56% (9/16) of the studies evaluated the effects of mHealth-based interventions on the risk of macrosomia [23,24,26-32]. Statistically significant differences in decreases in macrosomia were found between the mHealth intervention groups and the control groups (OR 0.59, 95% CI 0.40-0.87; \( P = .008; I^2 = 59\% \); Figure 6 [23,24,26-32]).
Effect on Large-for-Gestational-Age Infants

A total of 31% (5/16) of the studies explored the effects of mHealth-based lifestyle interventions on large-for-gestational-age infants [32,33,35-37]. Compared with the control groups, the mHealth intervention groups did not show statistically significant decreases in large-for-gestational-age infants (OR 0.80, 95% CI 0.60-1.06; P=.12; $I^2=51\%$).

Effect on Small-for-Gestational-Age Infants

In total, 25% (4/16) of the studies assessed the effects of mHealth-based lifestyle interventions on small-for-gestational-age infants [31,33,36,37]. Compared with the control groups, the mHealth intervention groups did not show statistically significant decreases in small-for-gestational-age infants (OR 1.10, 95% CI 0.78-1.55; P=.60; $I^2=0\%$).

Discussion

Methodological Quality of the Included Studies

The methodological quality of the included studies was assessed using the Cochrane risk-of-bias tool. A total of 25% (4/16) of the studies did not report details of randomization methods, and 75% (12/16) of the studies did not report allocation concealment and were at risk of selection bias. In total, 25% (4/16) of the studies did not blind participants and researchers. Therefore, there was a high risk of performance bias in these studies. A total of 12% (2/16) of the studies did not blind the outcome assessors and were rated as having a high risk of detection bias. The included studies reported complete data with low attrition bias. Finally, the funnel plot showed no substantial publication bias among the included studies. Therefore, the overall methodological quality was moderate, and larger samples and well-designed randomized trials are needed in the future.

Summary of Principal Findings

Effectiveness of mHealth Interventions

A total of 16 RCTs were included in this systematic review and meta-analysis. All of these studies offered mHealth lifestyle interventions for perinatal women with overweight or obesity. Pooled results showed that the mHealth interventions reduced the incidence of GDM, which was consistent with the results of a previous study [38]. Various strategies have been proposed to prevent GDM effectively, and the primary intervention strategy is to change the lifestyle. Lifestyle interventions, including dietary guidance, physical exercise, weight management, health education, and blood glucose self-monitoring, are first-line strategies for GDM prevention [39]. The benefits of lifestyle interventions are mediated by mechanisms that improve glycemic variables and outcomes in type 4 and type 2 diabetes by increasing insulin sensitivity and reducing oxidative stress, which has been demonstrated in studies in other populations [40]. Compared with conventional lifestyle interventions, mHealth-based lifestyle interventions can make health education more attractive by providing more intuitive and vivid education and consultation with the help of electronic devices. In addition, pregnant women need to maintain close and continuous contact with the medical team during pregnancy and communicate effectively in the event of pregnancy complications and conscious fetal abnormalities. The advantage of mHealth technology is its ability to provide time-sensitive connectivity and high-quality health care for pregnant women in all regions [41]. mHealth lifestyle interventions enable pregnant women to maintain a healthy diet and engage in appropriate physical activity on a daily basis to encourage a healthier lifestyle during pregnancy, thus reducing the incidence of GDM.

Subgroup Analysis

In this review, we classified mHealth lifestyle interventions into 3 groups based on delivery approach: app-based interventions, phone-based interventions, and mHealth interventions delivered via a computer. Subgroup analysis demonstrated that app-delivered mHealth interventions were highly effective in reducing the risk of GDM in women with overweight or obesity. mHealth telephone-based interventions and interventions delivered via a computer had no significant effect on the prevention of GDM. Perinatal women usually have different problems and specific needs during pregnancy. Personalized applications can focus on individual characteristics and tailor goals and actions to diverse populations [42]. In this way, medical staff can leverage app-based mHealth interventions to provide health care tailored to the specific needs of pregnant women with overweight or obesity. In contrast, the applications help medical staff remotely monitor real-time parameters related
to the health of pregnant women who are overweight or obese during pregnancy, observe whether they are adhering to a healthy and appropriate lifestyle, facilitate communication with them, and help them control their blood sugar and reduce the incidence of GDM [43].

Regarding the different providers of the interventions, the results of the subgroup analysis reported that mHealth interventions delivered by medical staff, including obstetricians and nurses, were effective in preventing GDM. It has been suggested that different models of care provided by various intervention providers may influence outcomes. mHealth interventions delivered by dietitians may focus on the food needs of women during pregnancy [44], whereas exercise physiologists are more concerned with physical activity during pregnancy [45]. Medical staff usually focus on the overall health of pregnant women, providing nutrition, exercise, weight management, and other comprehensive knowledge of pregnancy health care [46].

Subgroup analysis showed that mHealth interventions combining diet and exercise were effective in preventing gestational diabetes in women with overweight or obesity. Combined interventions with diet and exercise appeared to have a greater impact on GDM than interventions with diet or exercise alone, and the effects of diet and exercise on GDM were indistinguishable [47]. Type 2 diabetes has been proven to be preventable through combined diet, exercise, and weight loss interventions [48]. Another subgroup analysis found that mHealth-based lifestyle interventions were effective in preventing GDM in Asian populations. However, the pooled effects of interventions on the risk of GDM in predominantly White populations did not reach statistical significance. The disease burden of GDM varies by race because of multiple factors, including socioeconomic status, lifestyle, and culture, and the prevalence of GDM is significantly higher among Asian and Hispanic populations than among White populations [49]. The results of the subgroup analysis showed that, in studies using the IADPSG diagnostic criteria for GDM, mHealth-based lifestyle interventions had a significantly greater effect on GDM. The prevalence of GDM varies widely worldwide, at least in part because of a lack of consistency in screening and diagnostic criteria. Using lower glucose level thresholds as recommended by the IADPSG, significantly higher numbers of women with GDM were identified compared with using other diagnostic criteria [50].

Maternal and Neonatal Outcomes

Our meta-analyses revealed that the combined effect of mHealth interventions reduced the incidence of preterm birth, macrosomia, and excessive GWG. A review summarized the evidence regarding the influence of maternal diet before and during pregnancy on preterm birth. The results indicated that better maternal diet quality during pregnancy, characterized by a high intake of vegetables, fruits, whole grains, dairy products, and protein, played a significant role in reducing the risk of premature birth [51]. In addition, this study was in line with the findings of Fair and Soltani [52], who conducted a systematic review of the effectiveness of lifestyle interventions on weight gain in women with overweight or obesity during pregnancy and found that lifestyle interventions slightly reduced weight gain during pregnancy by 0.3 to 2.4 kg compared with standard care. There is no robust evidence that mHealth-based interventions are associated with a lower prevalence of postpartum hemorrhage, pregnancy-induced hypertension, cesarean sections, or any alteration in gestational age, consistent with a previous systematic review [53]. One reason for this nonsignificant effect may be the insufficient power of the combined sample size. Another possibility is the short duration of the interventions (median 18 wk), which might have been inadequate to affect some obstetric complications and neonatal outcomes. Further studies are needed to actively explore the associations between mHealth interventions and maternal and neonatal outcomes.

Strengths and Limitations

This systematic review and meta-analysis has several strengths. We combined MeSH terms and keywords covering pregnancy and mHealth to conduct a comprehensive search in 5 primary English electronic databases and 4 main Chinese electronic databases to minimize the possibility of publication bias. We used a robust 3-step search strategy to include databases containing published and unpublished RCTs. To minimize bias, the review methods were preregistered in accordance with the PRISMA statement. In addition, some of the studies included in this meta-analysis were conducted in Western, high-income countries, and some were conducted in Eastern, lower-income countries. Our meta-analyses provided an excellent synthesis of the responses of participants from different cultural backgrounds to mHealth interventions.

However, there were also some limitations to this review. First, we only retrieved studies published in English or Chinese owing to language limitations. Studies published in other languages were not included in the review, which may have led to some publication bias. Second, the methodological quality of the included studies was not optimal, and some studies had a risk of performance bias and detection bias. Third, the included studies varied in sample size, participant characteristics, components of the mHealth interventions, and intervention implementation methods, which may have led to high heterogeneity. Moreover, key variables such as the exact start of the intervention and intensity of physical activity were missing and incomplete in some studies and may have biased the pooled effects. Finally, not all the included studies reported the safety and cost-effectiveness of the mHealth interventions.

Conclusions

This systematic review and meta-analysis demonstrated that mHealth-based lifestyle interventions have a favorable impact on the prevention of GDM in pregnant women with overweight and obesity. mHealth interventions are a convenient and effective way to support pregnant women with overweight and obesity in out-of-hospital self-management in the context of rapid advances in IT and faster transmission speed. However, the potential of mHealth-based interventions for GDM needs to be further explored with better design and more rigorous large-scale RCTs.
Authors' Contributions
All authors made substantial contributions to the study. YH initiated the study. CH and QH conducted the data extraction and analyses. YH wrote the first draft of the manuscript. BL critically reviewed and revised the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest
None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File, 28 KB - mhealth_v12i1e49373_app1.docx ]

Multimedia Appendix 2
Search strategies.
[DOCX File, 22 KB - mhealth_v12i1e49373_app2.docx ]

Multimedia Appendix 3
The details of the mobile health interventions.
[DOCX File, 23 KB - mhealth_v12i1e49373_app3.docx ]

Multimedia Appendix 4
Funnel plot.
[DOCX File, 29 KB - mhealth_v12i1e49373_app4.docx ]

Multimedia Appendix 5
Forest plots of subgroup analysis.
[DOCX File, 172 KB - mhealth_v12i1e49373_app5.docx ]

Multimedia Appendix 6
Forest plots of secondary outcomes.
[DOCX File, 99 KB - mhealth_v12i1e49373_app6.docx ]

References

https://mhealth.jmir.org/2024/1/e49373


32. Dodd JM, Turnbull D, McPhee AJ, Deussen AR, Griell RM, Yelland LN. LIMIT Randomised Trial Group. Antenatal lifestyle advice for women who are overweight or obese: LIMIT randomised trial. BMJ 2014 Feb 10;348:g1285 [FREE Full text] [doi: 10.1136/bmj.g1285] [Medline: 24513442]


Abbreviations
- GDM: gestational diabetes mellitus
- GWG: gestational weight gain
- IADPSG: International Association of Diabetes and Pregnancy Study Groups
- MeSH: Medical Subject Headings
- mHealth: mobile health
- OR: odds ratio
- PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
- RCT: randomized controlled trial

Edited by L Buis; submitted 26.05.23; peer-reviewed by M Jiang, J Birks; comments to author 29.09.23; revised version received 22.11.23; accepted 28.11.23; published 17.01.24.

Please cite as:
He Y, Huang C, He Q, Liao S, Luo B
Effects of mHealth-Based Lifestyle Interventions on Gestational Diabetes Mellitus in Pregnant Women With Overweight and Obesity: Systematic Review and Meta-Analysis
JMIR Mhealth Uhealth 2024;12:e49373
URL: https://mhealth.jmir.org/2024/1/e49373
doi:10.2196/49373
PMID:38231555

©Yirong He, Chuanya Huang, Qiuyang He, Shujuan Liao, Biru Luo. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 17.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
**Review**

**Effectiveness of Telecare Interventions on Depression Symptoms Among Older Adults: Systematic Review and Meta-Analysis**

Man Wu¹, BSN; Chaoyang Li¹, BSN; Ting Hu¹, BSN; Xueyang Zhao¹, BSN; Guiyuan Qiao¹, BSN; Xiaolian Gao¹, BSN; Xinhong Zhu¹, PhD; Fen Yang¹,², PhD

¹School of Nursing, Hubei University of Chinese Medicine, Wuhan, China
²Hubei Shizhen Laboratory, Wuhan, China

**Corresponding Author:**
Fen Yang, PhD
School of Nursing
Hubei University of Chinese Medicine
North District, No 16, Huangjiahu West Road
Hongshan District
Wuhan, 430065
China
Phone: 86 18062003261
Email: fenyang@hbtcm.edu.cn

**Abstract**

**Background:** Depression is the most common psychiatric disorder among older adults. Despite the effectiveness of pharmacological and psychological therapies, many patients with late-life depression (LLD) are unable to access timely treatment. Telecare has been shown to be effective in addressing patients' psychosocial issues, while its effectiveness in serving patients with LLD remains unclear.

**Objective:** This study aimed to evaluate the effectiveness of telecare in reducing depression and anxiety symptoms and improving quality of life (QoL) in patients with LLD.

**Methods:** Databases including the Cochrane Library, Web of Science, PubMed, Embase, and EBSCO were searched for randomized controlled trials (RCTs) evaluating the effectiveness of telecare for LLD from database establishment to December 28, 2022.

**Results:** A total of 12 RCTs involving 1663 participants were identified in this study. The meta-analysis showed that (1) telecare significantly reduced depressive symptoms in patients with LLD compared to those in usual care (UC; standardized mean difference [SMD]=–0.46, 95% CI –0.53 to –0.38; \(P<.001\)), with the best improvement observed within 3 months of intervention (SMD=–0.72, 95% CI –1.16 to –0.28; \(P<.001\)); (2) other scales appeared more effective than the Patient Health Questionnaire-9 for LLD in telecare interventions (SMD=–0.65, 95% CI –0.96 to –0.35; \(P<.001\)); (3) telecare was more effective than telephone-based interventions for remote monitoring of LLD (SMD=–1.13, 95% CI –1.51 to –0.76; \(P<.001\)); (4) the reduction of depressive symptoms was more pronounced in patients with LLD with chronic conditions (SMD=–0.67, 95% CI –0.89 to –0.44; \(P<.001\)); (5) telecare was more effective for LLD in Europe and the Americas than in other regions (SMD=–0.73, 95% CI –0.99 to –0.47; \(P<.001\)); (6) telecare significantly reduced anxiety symptoms in patients with LLD (SMD=–0.53, 95% CI –0.73 to –0.33; \(P=.02\)); and (7) there was no significant improvement in the psychological components of QoL in patients with LLD compared to those receiving UC (SMD=0.30, 95% CI 0.18-0.43; \(P=.80\)).

**Conclusions:** Telecare is a promising modality of care for treatment, which can alleviate depression and anxiety symptoms in patients with LLD. Continued in-depth research into the effectiveness of telecare in treating depression could better identify where older patients would benefit from this intervention.

*(JMIR Mhealth Uhealth 2024;12:e50787) doi:10.2196/50787*

**KEYWORDS**
telecare; depression; anxiety; quality of life; older adults; meta-analysis
Introduction

Statistics show that the world’s population older than 60 years will double between 2015 and 2050, increasing from 12.0% to 22.0% [1]. With the rapid growth of the older population, late-life depression (LLD) has gradually emerged as a hot topic in the field of geriatric medical research. LLD refers to depressive disorders occurring in adults older than 60 years [2,3]. Research findings indicate a global prevalence of LLD of 28.4% [4], which could potentially be higher among individuals with concurrent physical ailments. As a geriatric syndrome with multifactorial etiology, LLD is highly associated with medical problems that pervade later life, including diabetes, hypertension, and dementia [2,5]. LLD is often chronic or recurrent and is associated with functional impairment, diminished health-related quality of life (QoL), and impaired social-psychological functioning [3,6]. A study confirmed that health care costs for patients with LLD were 43.0% to 52.0% higher for outpatient services and 47.0% to 51.0% higher when outpatient and inpatient services were combined, compared to those for individuals without LLD [7].

Despite its high prevalence and severe adverse outcomes, LLD is often overlooked and inadequately treated due to other complications resulting from aging-related issues. Psychopharmacotherapy and psychotherapy have been demonstrated to be effective for people with depression [6]; however, these treatments still have limitations, such as medical side effects and poor treatment adherence [8,9]. Due to mobility issues, geographic isolation, stigma associated with mental illness, and negative beliefs about treatment, older adults have limited access to health care or may be unwilling to seek help from health care institutions [10-12]. Additionally, underuse of professional mental health services, including low detection rates by health care providers and the lack of awareness among older patients regarding the severity of their condition [13,14], is also one of the factors that impede the treatment of LLD. Limited by these factors, only a minority of older adults receive appropriate treatment for depression. Therefore, there is an urgent need to study the clinical effectiveness of alternative therapies for depression, which are more socially acceptable and easily available.

In recent years, there has been increasing attention toward using telecare to support the management and well-being of mental health [15]. Telecare refers to the delivery of health care directly to users, typically in their own homes, supported by information and communication technologies such as telephone, videoconferencing, and applications [16,17]. Health care professionals can remotely provide consultation, assessment, and intervention services to patients [18]. These services include, but are not limited to, lifestyle monitoring, remote monitoring of vital signs for diagnosis, as well as long-distance assessment and education. The benefits of telecare are evident. Evidence suggests that as a promising strategy, telecare services can serve as a medium to overcome certain barriers, thereby enhancing mental health care and increasing opportunities to access evidence-based care under different conditions [19]. Particularly, telecare benefits older adults who are socially isolated or physically frail due to illness, disability, or other familial roles [17,20]. Currently, telecare has been widely used in the management of various chronic conditions among older adults, such as diabetes, hypertension, Parkinson disease, etc, yielding positive outcomes [21-23]. Depression is a commonly observed chronic condition among older adults, closely associated with an approximate 50% increase in chronic disease-related health care costs [24]. Given the significant impact of LLD on patients’ QoL and its potential consequences on decreased productivity or suicide, ensuring continuity of care is imperative. Telecare has been proposed as an effective alternative to help bridge this treatment problem. Considering the complexity and severity of LLD, it is necessary to further explore whether telecare is effective in improving health outcomes for patients with LLD.

Previous reviews have assessed the evidence related to the use of telecare for managing mental health issues [11,25]. In the field of psychiatry, telecare has been found to significantly impact mental health outcomes in older adults, including reducing emergency visits and hospitalizations, as well as improving cognitive function [11]. However, the efficacy of telecare for depression is inconsistent. Some studies suggest the effectiveness of telecare in reducing symptoms of depression [11,26], while others indicate that the impact of telecare on improving depressive symptoms is limited, even yielding contradictory results [27,28]. Previous meta-analyses examining the effectiveness of telecare on depression have mostly focused on adult populations [25-27]. However, compared to other age groups, LLD is considered to be different [14]. Differences in study design, intervention methods, and treatment intensity may contribute to varying clinical outcomes in telecare treatments for LLD. Despite recent meta-analyses demonstrating significant efficacy of telemedicine in alleviating depressive symptoms among older adults, the evaluation of its evidence remains limited [29]. Due to inherent heterogeneity in inclusion criteria, interpretation of these results should be approached cautiously.

The severe clinical outcomes and interfering factors often pose significant challenges in the treatment of LLD. Determining whether telecare management is effective for LLD is critical. It is unclear how effective telecare is in improving depression, anxiety symptoms, and QoL in patients with LLD. Therefore, this systematic review and meta-analysis explored the efficacy of telecare for LLD.

Methods

This systematic review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Multimedia Appendix 1) [30].

Search Strategy

We conducted searches in Cochrane Library, Web of Science, PubMed, Embase, and EBSCO for randomized controlled trials (RCTs) published from the inception of the databases up to December 28, 2022, without any language restriction. MeSH (Medical Subject Headings) and free search terms were both used in the literature search. The search terms included “cell Phones,” “telemedicine,” “smartphone,” “mobile applications,” “mobile phone*,” “telephone*,” “telehealth,” “tele-healthcare,” “electronic health*,” “application*,” “m-health,” “messaging,” “depression,” “depressive disorder,” “depress*,” and “LLD.”
“Major depression,” “sadness,” “late-life depression,” “LLD,” “aged,” “elder*,” “geriatric,” “senior people,” “RCTs,” etc. All titles, keywords, and abstracts have been reviewed in accordance with our search criteria. In this study, these research articles were exclusively published in English, focusing on telecare interventions for LLD. The specific search strategy is shown in Multimedia Appendix 2.

Study Selection and Data Exclusion
The inclusion criteria were the following: (1) studies were RCTs reported in full text with their title and abstract; (2) the average age of the study population was at least 60 years; (3) participants were diagnosed with depression in accordance with any established diagnostic criteria or with a score above a cutoff of any established depression rating scale at baseline; (4) the studies compare telecare (mobile phone, telephone, app, video, etc) participants with the control group receiving usual care (UC; routine, offline, or standard care); and (5) any health care professional providing care (ie, psychiatrists, family physicians, nurses, psychologists, etc).

Exclusion criteria were the following: (1) patients with manic or psychotic episodes or symptoms; (2) studies not related to the objective of this review and insufficient data, such as failure to report depression scale scores; and (3) books and studies without full text and studies in the format of abstracts of conference papers.

Data Extraction
Two authors independently reviewed all the databases, with specific search strategies for the relevant articles (MW and CYL). The software EndNote X9 (Clarivate) was used to import all the references and remove duplicates. After removing duplicates, the relevance of the title and abstract of the articles was evaluated. Any disagreements were discussed until a consensus was reached. After screening the title and abstract, the articles were selected for the next step of a full-text review. The 2 authors screened the full-text articles independently (MW and CYL). Finally, eligible articles included in the study were processed based on inclusion and exclusion criteria. Any discrepancies that arose during the assessment were resolved by a third reviewer (FY). Two authors independently extracted data from the included studies and entered them into a predesigned data extraction form. Data extracted for this study included the following: first author, year of publication, country, sample size, mean age, intervention approach, duration, presence or absence of comorbid chronic conditions, depression degree, and outcome measurement tools (Table 1).
Table 1. Basic characteristics of the included studies (N=12; all are randomized controlled trials).

<table>
<thead>
<tr>
<th>First author (year); country</th>
<th>Sample size, N (TC(^a)/UC(^b))</th>
<th>Age (years), mean (SD)</th>
<th>Duration</th>
<th>Comorbid chronic diseases</th>
<th>Depression degree</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rollman (2009); United States</td>
<td>302 (150/152)</td>
<td>TC: 64 (10.8); UC: 64 (11.2)</td>
<td>Baseline, 8 months</td>
<td>Yes</td>
<td>Moderate</td>
<td>HAM-D(^c) and SF-36(^d)</td>
</tr>
<tr>
<td>Aburizik (2013); United States</td>
<td>52 (29/23)</td>
<td>TC: 66.4 (7.9); UC: 64.1 (10.5)</td>
<td>Baseline, 10 weeks</td>
<td>Yes</td>
<td>Mild</td>
<td>PHQ-9(^e) and BDI(^f)</td>
</tr>
<tr>
<td>Lee (2014); Korea</td>
<td>25 (12/13)</td>
<td>TC: 66.7 (7.9); UC: 65.4 (8.6)</td>
<td>Baseline, 6 months</td>
<td>Yes</td>
<td>Mild, moderate</td>
<td>CES-D(^g)</td>
</tr>
<tr>
<td>Villani (2014); Italy</td>
<td>80 (40/40)</td>
<td>TC: 71 (4); UC: 73 (5)</td>
<td>Baseline, 12 months</td>
<td>Yes</td>
<td>Moderate, severe</td>
<td>PHQ-9 and STAI-6(^h)</td>
</tr>
<tr>
<td>Pickett (2014); United States</td>
<td>124 (60/64)</td>
<td>TC: 69.1 (10.9); UC: 68.6 (10.7)</td>
<td>Baseline, 12 weeks</td>
<td>No</td>
<td>Mild</td>
<td>PHQ-9</td>
</tr>
<tr>
<td>O'Neil (2014); Australia</td>
<td>121 (61/60)</td>
<td>TC: 61.0 (10.2); UC: 58.9 (10.7)</td>
<td>Baseline, 6 months</td>
<td>Yes</td>
<td>Mild, moderate</td>
<td>PHQ-9, CDS(^i), and SF-12(^j)</td>
</tr>
<tr>
<td>Gellis (2014); United States</td>
<td>94 (46/48)</td>
<td>TC: 80.1 (7.8); UC: 78.3 (6.9)</td>
<td>Baseline, 3 months, and 6 months</td>
<td>Yes</td>
<td>Mild, moderate</td>
<td>PHQ-9, HAM-D, and SF-12</td>
</tr>
<tr>
<td>Yang (2019); China</td>
<td>212 (107/105)</td>
<td>TC: 61.25 (8.6); UC: 60.85 (10.8)</td>
<td>Baseline, 12 months</td>
<td>Yes</td>
<td>Mild, moderate</td>
<td>HADS-D(^k) and SDS(^l)</td>
</tr>
<tr>
<td>Naik (2019); United States</td>
<td>225 (136/89)</td>
<td>61.9 (8.3)</td>
<td>Baseline, 6 months, and 12 months</td>
<td>Yes</td>
<td>Moderate</td>
<td>PHQ-9</td>
</tr>
<tr>
<td>Dobkin (2020); United States</td>
<td>72 (37/35)</td>
<td>TC: 65.62 (9.76); UC: 64.80 (9.62)</td>
<td>Baseline, 3 months, and 6 months</td>
<td>Yes</td>
<td>Moderate</td>
<td>HAM-D, BDI, HAM-A(^m), and SF-36</td>
</tr>
<tr>
<td>Almeida (2021); Australia</td>
<td>200 (79/121)</td>
<td>≥ 65</td>
<td>Baseline, and 52 weeks</td>
<td>No</td>
<td>Mild, moderate</td>
<td>PHQ-9, GAD-7(^n), and SF-12</td>
</tr>
<tr>
<td>Koehler (2021); Germany</td>
<td>156 (79/77)</td>
<td>TC: 68.30 (9.13); UC: 64.34 (11.35)</td>
<td>Baseline, 12 months</td>
<td>Yes</td>
<td>Moderate</td>
<td>PHQ-9 and SF-36</td>
</tr>
</tbody>
</table>

\(^a\)TC: telecare.
\(^b\)UC: usual care.
\(^c\)HAM-D: Hamilton Depression Rating Scale.
\(^d\)SF-36: 36-Item Short Form Survey.
\(^e\)PHQ-9: Patient Health Questionnaire-9.
\(^f\)BDI: Beck Depression Inventory.
\(^g\)CES-D: Center for Epidemiological Survey, Depression Scale.
\(^h\)STAI-6: Spielberger’s State Trait Anxiety Inventory.
\(^i\)CDS: Cardiac Depression Scale.
\(^j\)SF-12: 12-Item Short Form Survey.
\(^k\)HADS-D: Hospital Anxiety and Depression Scale.
\(^l\)SDS: Zung Self-Rating Depression Scale.
\(^m\)HAM-A: Hamilton Anxiety Rating Scale.
\(^n\)GAD-7: 7-item Generalized Anxiety Disorder Scale.

**Quality Assessment**

Two authors (MW and CYL) independently assessed the quality of the studies using the Cochrane Risk of Bias tool [40]. The assessment tool included 7 items (random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective reporting, and other bias), and authors judged each item individually as “low risk,” “high risk,” and “unclear risk.” The study was considered to be of high quality with a low risk score for at least 4 domains, of which 3 key areas had to be included (random sequence generation, allocation concealment, and incomplete outcome data). Consensus was reached by 2 authors (MW and CYL) through discussion with a third evaluator (FY).

**Statistical Analysis**

Data were analyzed using Stata (version 16.0; StataCorp) and Review Manager (version 5.3; The Cochrane Collaboration). Intervention effects were estimated by calculating Cohen \(d\) standardized mean differences (SMDs) and 95% CIs [41]. All studies reported outcomes as continuous data. The Cochran Q
statistic and \( I^2 \) statistic were used to assess the statistical heterogeneity between selected studies. Random-effects models were used when study heterogeneity was high (\( P<.10; I^2>50\% \)); otherwise, a fixed-effects model would be used. When heterogeneity identified across studies was high, we further performed subgroup analyses to explore possible explanations for heterogeneity. Publication bias was measured using a funnel plot and Egger linear regression analysis, and \( P<.05 \) on the Egger test indicated statistically significant publication bias [42].

**Results**

**Literature Search**

The database search yielded 15,265 articles, of which 14,249 publications were excluded. A total of 1016 full-text articles were assessed for eligibility. Finally, only 12 studies were eligible for inclusion in this meta-analysis [21-23,31-39], all of which were RCTs published between 2009 and 2021. The PRISMA flow diagram is shown in Figure 1.

**Figure 1.** PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. RCT: randomized controlled trial.

**Risks of Bias and Quality Assessment**

Overall, the quality of the included studies was moderate, of which 5 (41.7%) were of high quality. These studies show that the main bias in the blinding of participants and personnel may be caused by the nature of the intervention measures. All 12 articles reported adequate random sequence generation and, therefore, had a low risk of bias in this regard. In addition, 5 studies reported allocation concealment, which is a low risk of bias. As for detection bias, the assessors were blinded in 7 studies, the presence of blinding was unclear in 3 studies, and 2 studies were not blinded. The risks of study attrition bias and reporting bias were both low. Other risks of bias were also low but were unclear in 1 study. The specific risk of bias and quality assessment results are shown in Figures 2 and 3 [21-23,31-39].
Study and Patient Characteristics

The characteristics of the studies included are summarized in Table 1. A total of 1663 patients with LLD were involved, with an average age of over 60 years in each group. The sample size ranged from 25 [23] to 302 [31] participants. Studies were carried out across 6 countries, including the United States (n=6) [21,22,31,32,34,36], Korea (n=1) [23], Italy (n=1) [33], Australia (n=2) [35,38], China (n=1) [37], and Germany (n=1) [39]. Nine of these used telephone-based interventions, while the remaining studies used remote monitoring systems. Durations ranged from 10 weeks to 52 weeks. Depression, anxiety symptoms, and QoL were substantial influencing factors of treatment for older adults. Therefore, our primary outcome of interest was depression, and secondary outcomes were anxiety symptoms and QoL. Depression was evaluated using the Hamilton Depression Rating Scale, Patient Health Questionnaire-9 (PHQ-9), Beck Depression Inventory, Center for Epidemiological Survey, Depression Scale, Cardiac Depression Scale, Hospital Anxiety and Depression Scale, and Zung Self-Rating Depression Scale. Anxiety symptoms were assessed using Spielberger's State Trait Anxiety Inventory, Hamilton Anxiety Rating Scale, and the 7-item Generalized Anxiety Disorder Scale. QoL was assessed using the 12-Item Short Form Survey and the 36-Item Short Form Survey. A higher score on the scales indicated better QoL and...
greater severity of depression and anxiety symptoms. The specific interventions are available in Multimedia Appendix 3.

**Depression Symptoms**

A total of 12 RCTs involving 1663 participants were identified in this meta-analysis to calculate the effectiveness of telecare on depression, anxiety symptoms, and QoL in patients with LLD.

To compare the effects of telecare and UC in improving LLD, we included data from 12 of these studies. Our results show that telecare significantly reduced depressive symptoms in patients with LLD compared to those in UC (SMD=–0.46, 95% CI –0.53 to –0.38; P<.001). Fixed-effects model analysis revealed significant heterogeneity among the 12 included studies (I²=83.16%; P<.001; Figure 4) [21-23,31-39].

**Figure 4.** Forest plot for primary outcomes: depression. a: Hamilton Depression Rating Scale; b: Patient Health Questionnaire-9; c: Beck Depression Inventory; d: Center for Epidemiological Survey, Depression Scale; e: Cardiac Depression Scale; f: Hospital Anxiety and Depression Scale; g: Zung Self-Rating Depression Scale.

To address high heterogeneity, we performed subgroup analyses grouped by the type of scale (PHQ-9 or others), duration time (≤3 months or >3 months), device type (telephone-based or remote monitoring system), comorbid chronic diseases (presence or absence), and region (Europe and the Americas or others). Random-effects models indicated that telecare significantly reduced depressive symptoms in patients with LLD compared to the UC participants (SMD=–0.59, 95% CI –0.80 to –0.38; P<.001). Results of subgroup analysis by duration showed that short-term (≤3 months) interventions (SMD=–0.72, 95% CI –1.16 to –0.28; P<.001) were more effective than long-term (>3 months) interventions (SMD=–0.52, 95% CI –0.75 to –0.29; P<.001); other scales (SMD=–0.65, 95% CI –0.96 to –0.35; P<.001) were more effective than the PHQ-9 (SMD=–0.53, 95% CI –0.83 to –0.22; P<.001); the remote monitoring system (SMD=–1.13, 95% CI –1.51 to –0.76; P<.001) was more effective than telephone-based interventions (SMD=–0.38, 95% CI –0.56 to –0.20; P<.001); the effect on patients with LLD with chronic diseases (SMD=–0.67, 95% CI –0.89 to –0.44; P<.001) was better than that on patients with LLD without comorbid chronic diseases (SMD=–0.10, 95% CI –0.41 to 0.20; P=.07); and telecare was more effective in Europe and the Americas (SMD=–0.73, 95% CI –0.99 to –0.47; P<.001) than in other regions (SMD=–0.22, 95% CI –0.35 to –0.09; P=.42; Table 2).
Table 2. Subgroup meta-analysis for patients with late-life depression.

<table>
<thead>
<tr>
<th>Subgroups</th>
<th>Cohen d SMD(^a)</th>
<th>95% CI</th>
<th>P value</th>
<th>Heterogeneity (I(^2); %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>-0.59</td>
<td>-0.80 to -0.38</td>
<td>&lt;.001</td>
<td>86.42</td>
</tr>
<tr>
<td><strong>Duration</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤3 months</td>
<td>-0.72</td>
<td>-1.16 to -0.28</td>
<td>&lt;.001</td>
<td>86.91</td>
</tr>
<tr>
<td>&gt;3 months</td>
<td>-0.52</td>
<td>-0.75 to -0.29</td>
<td>&lt;.001</td>
<td>84.72</td>
</tr>
<tr>
<td><strong>Type of scale</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHQ-9(^b)</td>
<td>-0.53</td>
<td>-0.83 to -0.22</td>
<td>&lt;.001</td>
<td>87.13</td>
</tr>
<tr>
<td>Others</td>
<td>-0.65</td>
<td>-0.96 to -0.35</td>
<td>&lt;.001</td>
<td>86.21</td>
</tr>
<tr>
<td><strong>Device type</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Telephone-based</td>
<td>-0.38</td>
<td>-0.56 to -0.20</td>
<td>&lt;.001</td>
<td>75.30</td>
</tr>
<tr>
<td>Remote monitoring system</td>
<td>-1.13</td>
<td>-1.51 to -0.76</td>
<td>&lt;.001</td>
<td>78.32</td>
</tr>
<tr>
<td><strong>Comorbid chronic diseases</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presence</td>
<td>-0.67</td>
<td>-0.89 to -0.44</td>
<td>&lt;.001</td>
<td>85.31</td>
</tr>
<tr>
<td>Absence</td>
<td>-0.10</td>
<td>-0.41 to 0.20</td>
<td>.07</td>
<td>61.45</td>
</tr>
<tr>
<td><strong>Region</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Europe and the Americas</td>
<td>-0.73</td>
<td>-0.99 to -0.47</td>
<td>&lt;.001</td>
<td>86.38</td>
</tr>
<tr>
<td>Others</td>
<td>-0.22</td>
<td>-0.35 to -0.09</td>
<td>.42</td>
<td>0.00</td>
</tr>
</tbody>
</table>

\(^a\)SMD: standardized mean difference.
\(^b\)PHQ-9: Patient Health Questionnaire-9.

Meta-regression analysis showed that heterogeneity may not be related to the year of publication (P=.42), total sample size (P=.21), study area (P=.35), comorbid chronic disease (P=.47), duration (P=.75), and outcome measurement tools (P=.29). However, only the intervention device (P=.004) may have contributed to the heterogeneity.

**Sensitivity Analysis and Publication Bias**

The stability and reliability of the results of this meta-analysis and potential factors contributing to heterogeneity were explored by sensitivity analysis to assess the effect of the data of each study on the combined effect value (ie, SMD). The results of the sensitivity analysis showed that excluding each study individually had no significant effect on the combined effect value, and the study results were stable and reliable (Figure 5) [21-23,31-39]. Publication bias was assessed using funnel plots and Egger test indicators. The funnel plot was symmetrically distributed on both sides (Figure 6), and the Egger test showed no significant publication bias (P=.05).
Anxiety Symptoms
To examine the efficacy of telecare in reducing anxiety compared with that of UC, we included 3 articles on patients with LLD. The results showed that telecare significantly reduced anxiety symptoms in patients with LLD (SMD=−0.53, 95% CI −0.73 to −0.33; \( P = .02 \); Figure 7) [22,33,38].
Six studies assessed the mental components of QoL by using the Medical Outcomes Study Short Form survey. Our meta-analysis shows that the QoL of patients with LLD improved, but, overall, it was not significant (SMD = 0.30, 95% CI 0.18-0.43; \( P = .80 \); Figure 8) [22,31,35,36,38,39].

**QoL**

**Discussion**

**Principal Findings**

This meta-analysis shows that compared with UC, telecare significantly reduces symptoms of depression and anxiety but has no significant effect on improving QoL in patients with LLD.

**Primary Outcome Measures**

The pooled results show that telecare has a significant effect on reducing depressive symptoms in patients with LLD, which is consistent with the findings of previous studies [25,26,29]. Apart from dealing with depression itself, the increased severity of LLD is also related to factors such as aging, chronic disease, and socioeconomic stress [5]. Telecare offers unique and innovative opportunities for treating depression symptoms in older adults. Patients with LLD can leverage the advantages of telecare to connect with health care professionals, overcoming geographical distance and physical limitations, thereby reducing the psychological burden of coping with the disease [43]. Furthermore, professional psychological support is crucial for patients with LLD, and it can encourage patients to express their feelings and release stress [44]. However, it is worth noting that despite telecare offering more possibilities for treating LLD, the complexity of the medical population makes it challenging. Telecare can provide greater coverage for health care, yet considerations such as individual needs of older patients or environmental backgrounds need to be factored in [45]. Currently, offering targeted telecare services to a large population of older adults in rural, remote, or underserved areas remains a challenge [10]. In particular, older adults face significant barriers in using telephone and internet connections during the COVID-19 pandemic [46]. As a result, telecare management may not be as effective for this population as for others. The size of the research effect will depend on the nature of the intervention and the quality of the study [47]. High-quality telemedicine will help older adults benefit both physically and mentally. Further investigation and more research are necessary. Subgroup analysis indicates that the effectiveness of telecare in treating LLD can be influenced by measurement tools,
durations, intervention devices, comorbid chronic conditions, and regions involved. In terms of depression measurement tools, other scales appear to be more effective than PHQ-9 (0.65% vs 0.53%), which may be related to measurement errors caused by differences in specific items and the generalizability of different measurement tools [48]. Results from durations of ≤3 and >3 months showed a reduction in depressive symptoms in patients with LLD, with short-term interventions proving to be more effective (0.72% vs 0.52%). Short-term interventions focus more on addressing specific issues or symptoms, producing immediate effects. For older adults, short-term interventions might be more readily accepted as long-term treatments could induce fatigue or a lack of patience. Our findings differ slightly from those of another study [49], which implemented more targeted interventions based on different treatment responses, confirming the more significant effectiveness of long-term interventions. Therefore, there is insufficient evidence to conclusively establish that telecare is necessarily superior in short-term intervention efficacy for LLD compared to long-term interventions. In fact, for depression management, a combination of short-term and long-term interventions is often required to deliver comprehensive and enduring support and management [50].

Subgroup analysis also found that remote monitoring systems appear to be more effective than telephone-based management (1.13% vs 0.38%). The remote monitoring system ensures timely and accurate transmission of patients' symptom information and data to health care professionals, enabling patients to receive effective treatments [51]. Telecare was more effective in patients with LLD with comorbid chronic conditions compared to those without such comorbidities (0.67% vs 0.10%). Co-occurrence of chronic medical conditions and depression is common. Evidence suggests that older adults with chronic illness are more likely to be affected by depressive symptoms than those without chronic illness [2,5,7]. Older adults with chronic conditions are more likely to seek medical care and adhere to treatment [52]. Therefore, while actively treating chronic conditions, there might be a degree of alleviation in depressive symptoms among older adults. Telecare was more effective in Europe and the Americas in improving depressive symptoms in patients with LLD compared than in other regions (0.73% vs 0.22%). The health care systems in Europe and the Americas are generally more developed, which may lead to more comprehensive support for telecare [53]. In low- and middle-income countries, the resources available for geriatric mental health care are considered severely inadequate [54]. Nevertheless, telecare is beginning to have an important impact on many aspects of health care in transitional countries [55].

Secondary Outcome Measures

Telecare has a positive effect on improving anxiety symptoms of patients with LLD. This result is consistent with findings from other studies [56]. Telecare offers a more convenient access method, allowing patients to receive treatment at home, thereby circumventing the inconvenience and anxiety associated with hospital visits [16,17]. Health care professionals can engage with patients more frequently through telecare, gaining insights into their symptoms and emotional fluctuations. This allows for adjustments in the treatment plan to effectively meet the unique needs of this population [18,57]. Additionally, the symptoms of anxiety and depression are often co-occurring [58], particularly among older adults. Due to the similarity between depression and anxiety symptoms, many treatment approaches are shared between the two. A recent meta-analysis suggests that psychotherapy delivered remotely is as effective as face-to-face therapy for anxiety disorder [59]. This evidence is based on outcomes obtained from different age groups. It may be more challenging to create a trusting relationship remotely than in person [60]. Older adults have negative views about health IT performing accurately and dependably, which will have a significant impact on the acceptance of telecare [61]. In brief, when using telecare for addressing emotional disorders in older adults, closer supervision and guidance might be necessary. Health care professionals need to distinguish the appropriateness of using telecare for communication and, in turn, individually tailor patient care.

We found that the mental component of QoL in patients with LLD improved after using telecare; however, this difference was not significant compared to that with the use of UC. This finding aligns with results from other studies [62,63]. Improving QoL is a comprehensive issue that includes not only improvements in health care but also social, psychological, and emotional factors [64]. Influenced by these factors, it is difficult to compare the results of QoL considering different contexts. Several results from RCTs with older adults using telemonitoring systems showed an improvement in the participants' QoL [65,66]; other telemonitoring RCTs could not achieve congruent results [67]. Improvements in QoL often require deeper interactions and personalized care. In particular, participants with mental disorders may benefit from individual and tailored solutions provided by general practitioners [68]. When using telecare, it is crucial to acknowledge that each subpopulation of marginalized older adults has differing strengths and needs. The studies we included focused more on managing the disease itself, which may weaken overall effectiveness. It is not easy to present telemedicine to the older population. The limitations inherent in older adults may lead to difficulties in receiving telecare, including the lack of technical literacy, equipment access barriers, cognitive function, etc [11]. These reasons could explain why telecare is not significant in improving the QoL of patients with LLD. The potential value of telecare in maintaining the QoL for individuals with LLD warrants further exploration. While this study did not reveal a positive impact of telecare on the QoL for patients with LLD, it has been established that telecare can assist patients with LLD in gaining more information about health services.

Limitations

This study still had some limitations. First, most of the studies included in the review lacked sufficient measure detail, leading to irreversible bias. Our study mainly included 2 interventions based on telephone and remote monitoring to reduce this bias. Second, the measurement tools used in this study lacked standardization and heavily relied on self-reports from participants, introducing a degree of subjectivity and concealment that is not as rigorous as structured interviews. However, we attempted to validate the effectiveness of the results by using authoritative scales. Third, differences in the

https://mhealth.jmir.org/2024/1/e50787

JMR Mhealth UHealth 2024 | vol. 12 | e50787 | p.147

(page number not for citation purposes)
specific intervention methods, frequency, and content among the included studies may lead to clinical heterogeneity across different studies.

Conclusions
Our meta-analysis shows that telecare has a positive impact on depressive and anxiety symptoms, despite high heterogeneity in depression symptoms. Therefore, studies with larger sample sizes and homogeneity were required to determine the effects of telecare in patients with LLD. Future research can continue to refine telecare systems and assess the specific needs of older vulnerable populations for more accurate evidence.

Acknowledgments
This project was jointly supported by the Hubei Provincial Natural Science Foundation and the Innovative Development of Chinese Medicine of China (2023AFD160).

Conflicts of Interest
None declared.

Multimedia Appendix 1
PRISMA checklist.
[DOCX File, 28 KB - mhealth_v12i1e50787_app1.docx]

Multimedia Appendix 2
Search strategy.
[DOC File, 17 KB - mhealth_v12i1e50787_app2.doc]

Multimedia Appendix 3
The intervention content of the intervention group.
[DOC File, 26 KB - mhealth_v12i1e50787_app3.doc]

References


Abbreviations

LLD: late-life depression
MeSH: Medical Subject Headings
PHQ-9: Patient Health Questionnaire-9
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QoL: quality of life
RCT: randomized controlled trial
SMD: standardized mean difference
UC: usual care

Edited by L Buis; submitted 12.07.23; peer-reviewed by W He, X Tan, S Peng; comments to author 15.11.23; revised version received 01.12.23; accepted 22.12.23; published 17.01.24.

Please cite as:
Effectiveness of Telecare Interventions on Depression Symptoms Among Older Adults: Systematic Review and Meta-Analysis
JMIR Mhealth Uhealth 2024;12:e50787
URL: https://mhealth.jmir.org/2024/1/e50787
doi:10.2196/50787
PMID:38231546

©Man Wu, Chaoyang Li, Ting Hu, Xueyang Zhao, Guiyuan Qiao, Xiaolian Gao, Xinhong Zhu, Fen Yang. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 17.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use,
distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Review

Effects of Digital Physical Health Exercises on Musculoskeletal Diseases: Systematic Review With Best-Evidence Synthesis

Johanna Nagel¹, MSc; Florian Wegener¹, MSc; Casper Grim², PD, Dr med; Matthias Wilhelm Hoppe¹, Prof Dr

¹Movement and Training Science, Leipzig University, Leipzig, Germany
²Center for Musculoskeletal Surgery Osnabrück, Klinikum Osnabrück, Osnabrück, Germany

Corresponding Author:
Johanna Nagel, MSc
Movement and Training Science
Jahnallee 59
Leipzig, 04109
Germany
Phone: 49 3419731822
Email: jojo_nagel@yahoo.de

Abstract

Background: Musculoskeletal diseases affect 1.71 billion people worldwide, impose a high biopsychosocial burden on patients, and are associated with high economic costs. The use of digital health interventions is a promising cost-saving approach for the treatment of musculoskeletal diseases. As physical exercise is the best clinical practice in the treatment of musculoskeletal diseases, digital health interventions that provide physical exercises could have a highly positive impact on musculoskeletal diseases, but evidence is lacking.

Objective: This systematic review aims to evaluate the impact of digital physical health exercises on patients with musculoskeletal diseases concerning the localization of the musculoskeletal disease, patient-reported outcomes, and medical treatment types.

Methods: We performed systematic literature research using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The search was conducted using the PubMed, BISp, Cochrane Library, and Web of Science databases. The Scottish Intercollegiate Guidelines Network checklist was used to assess the quality of the included original studies. To determine the evidence and direction of the impact of digital physical health exercises, a best-evidence synthesis was conducted, whereby only studies with at least acceptable methodological quality were included for validity purposes.

Results: A total of 8988 studies were screened, of which 30 (0.33%) randomized controlled trials met the inclusion criteria. Of these, 16 studies (53%) were of acceptable or high quality; they included 1840 patients (1008/1643, 61.35% female; 3 studies including 197 patients did not report gender distribution) with various musculoskeletal diseases. A total of 3 different intervention types (app-based interventions, internet-based exercises, and telerehabilitation) were used to deliver digital physical health exercises. Strong evidence was found for the positive impact of digital physical health exercises on musculoskeletal diseases located in the back. Moderate evidence was found for diseases located in the shoulder and hip, whereas evidence for the entire body was limited. Conflicting evidence was found for diseases located in the knee and hand. For patient-reported outcomes, strong evidence was found for impairment and quality of life. Conflicting evidence was found for pain and function. Regarding the medical treatment type, conflicting evidence was found for operative and conservative therapies.

Conclusions: Strong to moderate evidence was found for a positive impact on musculoskeletal diseases located in the back, shoulder, and hip and on the patient-reported outcomes of impairment and quality of life. Thus, digital physical health exercises could have a positive effect on a variety of symptoms of musculoskeletal diseases.

(JMIR Mhealth Uhealth 2024;12:e50616) doi:10.2196/50616

KEYWORDS
mobile health; mHealth; electronic health; eHealth; digital health applications; DiGA; musculoskeletal; MSK; home-based; PROM; disorder; mobile phone
**Introduction**

**Background**

A total of 1.71 billion people are affected by musculoskeletal diseases worldwide [1]. They are characterized by chronic pain, functional disability, impairment, and reduced quality of life [1,2]. The most commonly affected body regions are the lower back and neck, with a period prevalence over the last 12 months of up to 61.3% and 45.7% [3], respectively, and a common disease is osteoarthritis, with a prevalence of up to 17.9% [4]. In addition to the high biopsychosocial burden [5], the evident increase in the incidence of musculoskeletal diseases over the last decades [6] results in high economic costs because of lost workdays and conservative or operative medical treatments [5]. To overcome such undesirable consequences, evidence-based, effective, and cost-saving health interventions are required. Therefore, the use of digital health interventions is a promising approach.

Digital health interventions aim to manage a wide range of diseases and health issues using digital devices such as smartphones, tablets, computers, or wearables, including mobile apps, telediagnosis and web-based physician visits, web-based interactive programs, or tracking tools [7]. The use of mobile apps is increasing, with common intervention types categorized as physical exercise and fitness, lifestyle and stress, diet and nutrition, or medication reminders and educational materials [7]. In some countries, such as Germany, so-called digital health applications are also supported by health insurers after being evaluated as medical devices [8]. However, owing to their cost-saving potential and the increasing number of commercially available digital health interventions [7], further research is needed to evaluate the impact of different types of digital health interventions on specific diseases.

Previous systematic reviews have extensively evaluated the impact of digital health interventions on internal diseases. Positive effects have been demonstrated in treating chronic obstructive pulmonary disease [9], cardiovascular disease [10], and diabetes [11]. These effects encompass improvements in clinically relevant outcomes such as quality of life, health-related impairments, amelioration of risk factors and their consequences, as well as the control and management of HbA1c levels. For musculoskeletal diseases, only 2 previous systematic reviews have evaluated the impact of digital health interventions as a primary outcome. One review [12] showed that there are substantial clinical benefits in the management of musculoskeletal diseases for the patient-reported outcomes of pain (9 out of 19 studies) and functional disability (10 out of 16 studies). The results show that digital health interventions as adjuncts and as stand-alone treatments are not inferior but partly superior compared with interventions based on standard therapy, nondigital self-management, noninteractive digital measures, or no intervention. However, in this previous review, no evidence synthesis was performed. In addition, a further review [13] conducted a meta-analysis and showed moderate-quality evidence that digital health interventions are effective in reducing pain and improving function and self-management in patients with musculoskeletal disease. The included studies considered digital health interventions as interventions that are to be used only at home and as adjuncts to standard clinical care, compared with standard care, noninteractive digital interventions, or no intervention. Taken together, the use of digital health interventions as an adjunct to regular therapy could have positive health-related effects for both internal and musculoskeletal diseases, although less evidence is available for the latter.

However, little is known about the relationship between clinically relevant factors, such as the localization of the musculoskeletal diseases, patient-reported outcomes, or the type of applied conservative or operative medical treatments, and the effects of different types of digital health interventions in the treatment of musculoskeletal diseases. In terms of evidence-based medicine, such relationships must first be clarified when using digital health interventions as a regular treatment option for specific musculoskeletal diseases. Because of the increasing number of original studies, more systematic research is needed to review and assess the existing evidence. Previous systematic reviews [12,13] have included all types of digital health interventions, providing a comprehensive overall result across all biopsychosocial domains. As physical exercise is the best clinical practice for the treatment of musculoskeletal diseases [14], digital physical health exercises could have a highly positive impact on musculoskeletal diseases. However, little is known about how the effects of digital physical health exercises are related to the aforementioned clinically relevant factors.

**Objective**

Therefore, this systematic review aimed to evaluate the impact of digital physical health exercises on patients with musculoskeletal diseases concerning the localization of the musculoskeletal disease, patient-reported outcomes, and medical treatment types. In addition, a best-evidence synthesis was conducted to estimate the direction and strength of the existing evidence.

**Methods**

**Research Design and Eligibility Criteria**

The systematic review was conducted according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [15]. Eligibility criteria according to the population, intervention, comparison, outcome, study design (PICOS) scheme [16] were applied. Table 1 presents the inclusion criteria according to the PICOS scheme. Textbox 1 presents the search line.

The corresponding keywords are also presented. Studies were not reviewed if they did not report on a specific musculoskeletal disease, if the digital health intervention included no physical exercises, if no control group was considered, or if none of the included patient-reported outcomes were assessed as a primary outcome. All methodological steps were performed by 1 author and validated by a second author. Uncertainties were discussed until consensus was reached. Because of the literary nature of this study, ethics approval was not required.
Table 1. PICOS\(^a\) scheme for the definition of the inclusion criteria and the presentation of the corresponding keywords.

<table>
<thead>
<tr>
<th>Inclusion criteria(^b)</th>
<th>Population</th>
<th>Intervention</th>
<th>Comparison</th>
<th>Outcome</th>
<th>Study design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients with any musculoskeletal disease according to the definition of the WHO(^c)</td>
<td>Any digital health intervention using home-based physical exercises</td>
<td>Any conventional or no therapy</td>
<td>Patient-reported outcomes pain, function, disability, and quality of life assessed by established and validated clinical questionnaires or scales</td>
<td>Randomized controlled trials</td>
<td></td>
</tr>
</tbody>
</table>

**Keywords**

- “Musculoskeletal disease” OR “Musculoskeletal disorder” OR “Musculoskeletal pain” OR “Chronic pain” OR “Acute pain” OR “Overuse pain” OR “Chronic injury” OR “Chronic injuries” OR “Acute injury” OR “Acute injuries” OR “Overuse injury” OR “Overuse injuries” OR “Chronic disease” OR “Acute disease” OR “Overuse disease” OR “Osteoporosis” OR “Osteoarthritis” OR “Rheumatoid arthritis” OR “Tendinopathy” OR “Tendinopathies” OR “Rotator cuff” OR “Lower extremity” OR “Lower extremities” OR “Upper extremity” OR “Upper extremities” OR “Hip” OR “Knee” OR “Foot” OR “Hand” OR “Ankle” OR “Wrist” OR “Elbow” OR “Low back” OR “Neck” OR “Back” OR “Spine” OR “Shoulder” OR “Arm” OR “Leg” OR “Muscle” OR “Tendon” OR “Ligament”

- “Digital movement therapy” OR “Digital movement therapies” OR “Mobile health” OR “eTherapy” OR “eTherapies” OR “Web-based intervention” OR “Digital intervention” OR “Computer-based intervention” OR “App-based intervention” OR “Digital health application” OR “Technology-assisted therapy” OR “Technology-assisted therapies” OR “Internet-based intervention” OR “Computer-assisted therapy” OR “Computer-assisted therapies” OR “health app” OR “mobile application” OR “Smartphone” OR “Mobile phone” OR “ehealth” OR “mhealth” OR “telerhabilitation” OR “Telemedicine” OR “online intervention” OR “internet-delivered intervention”

- “Osteopathy” OR “movement therapy” OR “movement therapies” OR “physical therapy” OR “physical therapies” OR “therapeutic exercise” OR “medical gymnastic” OR “traditional therapy” OR “traditional therapies” OR “manual therapy” OR “manual therapies” OR “physiotherapy” OR “No therapy” OR “No therapies” OR “conventional therapy” OR “conventional therapies” OR “no treatment” OR “no intervention” OR “watch-and-wait” OR “wait-and-see” OR “wait and see”

---

\(^a\)PICOS: Population, intervention, comparison, outcome, and study design.

\(^b\)Others: Studies in English or German language with free full access were included.

\(^c\)WHO: World Health Organization.

\(^d\)N/A: Not applicable.

**Textbox 1. Search line.**

```
("Musculoskeletal disease OR musculoskeletal disorder OR musculoskeletal pain OR chronic pain OR acute pain OR overuse pain OR chronic injury OR chronic injuries OR acute injury OR acute injuries OR overuse injury OR overuse injuries OR chronic disease OR acute disease OR overuse disease OR osteoporosis OR osteoarthritis OR rheumatoid arthritis OR tendinopathy OR tendinopathies OR rotator cuff OR lower extremity OR lower extremities OR upper extremity OR upper extremities OR hip OR knee OR foot OR hand OR ankle OR wrist OR elbow OR low back OR neck OR back OR spine OR shoulder OR arm OR leg OR muscle OR tendon OR ligament") AND (digital movement therapy OR digital movement therapies OR mobile health OR eTherapy OR eTherapies OR Web-based intervention OR Digital intervention OR Computer-based intervention OR App-based intervention OR Digital health application OR Technology-assisted therapy OR Technology-assisted therapies OR Internet-based intervention OR Computer-assisted therapy OR Computer-assisted therapies OR health app OR mobile application OR Smartphone OR Mobile phone OR ehealth OR mhealth OR telerhabilitation OR Telemedicine OR online intervention OR internet-delivered intervention) AND (osteopathy OR movement therapy OR movement therapies OR physical therapy OR physical therapies OR therapeutic exercise OR medical gymnastic OR traditional therapy OR traditional therapies OR manual therapy OR manual therapies OR physiotherapy OR no therapy OR no therapies OR conventional therapy OR conventional therapies OR no treatment OR no intervention OR watch-and-wait OR wait-and-see OR watch and wait OR wait and see) AND (randomized controlled trials)
```
Literature Search, Study Selection, and Risk of Bias

The literature search was performed on July 21, 2022, using the PubMed (MEDLINE), BISp (Federal Institute of Sport Science), Cochrane Library, and Web of Science databases. The search line included terms presented in Table 1. The “outcomes” category was not included in the search strategy but was considered in the subsequent study inclusion and selection process. No filters or other restrictions were used. The retrieved records were exported to a reference manager (EndNote 20, Clarivate). All duplicates were identified using the software and were removed after a manual review. On the basis of the defined eligibility criteria, studies were included or excluded by reviewing the titles, abstracts, and full texts. Full texts were accessed via public or open access and university accounts. If the full texts were not accessible, the authors were contacted. The study quality and the associated risk of bias were assessed using the Scottish Intercollegiate Guidelines Network checklist for randomized controlled trials [17]. The checklist consisted of 10 items related to the internal validity and 2 items related to the overall assessment of the studies. For each included study, all items were answered with “yes,” “no,” “can’t say,” or “not applicable.” The study quality was then finally rated throughout the “Scottish Intercollegiate Guidelines Network checklist for randomized controlled trials: Notes for completion of checklist” as “not acceptable,” “borderline,” “acceptable,” and “high,” as previously done [18]. The definitions of these quality classifications are presented in Textbox 2.

Textbox 2. Definitions for ratings of the overall methodological study quality.

<table>
<thead>
<tr>
<th>High quality</th>
<th>Most criteria met. Little or no risk of bias. Results unlikely to be changed by further research.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceptable quality</td>
<td>Most criteria met. Some flaws in the study with an associated risk of bias. Conclusions may change in the light of further studies.</td>
</tr>
<tr>
<td>Borderline quality</td>
<td>Crude effect estimates have been presented or have been calculated (thus no confounders have been considered), but the study is otherwise acceptably sound with respect to other possible biases.</td>
</tr>
<tr>
<td>Not acceptable quality</td>
<td>Either most criteria not met, or significant flaws relating to key aspects of study design. Conclusions likely to change in the light of further studies.</td>
</tr>
</tbody>
</table>

Note: Definitions according to Asker et al [18].

Data Extraction and Synthesis of Results

Data extraction was performed according to the PICOS scheme. A best-evidence synthesis was conducted to clarify the evidence for digital physical health exercises on clinically relevant factors clustered as (1) localization of the musculoskeletal diseases, (2) patient-reported outcomes (according to the eligibility criteria), and (3) medical treatment types (conservative vs operative). Within these clusters, the study results were individually classified as positive, negative, or equal for each clinically relevant factor. In accordance with a previous study [14], the study results were classified as positive or negative if the intervention or control group showed statistically better significant study results than the other group for >50% of the outcome parameters that were used to examine the respective clinically relevant factors. If no statistically significant differences were reported between the intervention and control groups, the study results were classified as equal. With regard to the best-evidence synthesis, the established criteria [18] are summarized in Table 2, and to increase the validity, only studies with at least acceptable study quality were included [19].

Table 2. Criteria of best-evidence synthesis according to Asker et al [18].

<table>
<thead>
<tr>
<th>Rating</th>
<th>Study quality</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong evidence</td>
<td>≥2 high-quality studies</td>
<td>≥75% consistent findings in these studies</td>
</tr>
<tr>
<td>Moderate evidence</td>
<td>1 high-quality study and/or ≥2 acceptable-quality studies</td>
<td>≥75% consistent findings in these studies</td>
</tr>
<tr>
<td>Limited evidence</td>
<td>1 acceptable-quality study and/or ≥1 borderline-quality study</td>
<td>N/A</td>
</tr>
<tr>
<td>Conflicting evidence</td>
<td>Conflicting results in several studies of any quality</td>
<td>&lt;75% of studies reported concordant results</td>
</tr>
<tr>
<td>No evidence</td>
<td>No admissible studies were found</td>
<td>N/A</td>
</tr>
</tbody>
</table>

aN/A: Not applicable.
Results

Literature Search, Study Selection, and Risk of Bias

Figure 1 shows the flowchart of the literature search including the study selection process according to the PRISMA guidelines. On the basis of the eligibility criteria, of 10,441 records, 30 (0.29%) studies were finally included in the risk of bias assessment. Although 1453 studies were identified as duplicates, 8958 studies that did not meet the inclusion criteria addressed no musculoskeletal diseases, were not randomized controlled trials, or addressed other outcomes. Table 3 summarizes the results of the risk of bias assessment.

There were 10 studies with high quality [20-29], 6 studies with acceptable quality [30-35], 10 studies with borderline quality [36-45], and 4 studies with not acceptable quality [46-49]. Thus, because of the not acceptable and borderline qualities of 14 studies, 16 studies were further analyzed and finally included in the best-evidence synthesis.

Figure 1. Flowchart of the literature search including the study selection process according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. RCT: randomized controlled trial.
<table>
<thead>
<tr>
<th>Study</th>
<th>1.1</th>
<th>1.2</th>
<th>1.3</th>
<th>1.4</th>
<th>1.5</th>
<th>1.6</th>
<th>1.7</th>
<th>1.8 (%)</th>
<th>1.9</th>
<th>2.1</th>
<th>2.2</th>
<th>2.3</th>
<th>Total</th>
<th>Study quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abadiyan et al [20]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>3</td>
<td>CS</td>
<td>N/A</td>
<td>++</td>
<td>Yes</td>
<td>Yes</td>
<td>8</td>
</tr>
<tr>
<td>Allen et al [21]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>13.1</td>
<td>Yes</td>
<td>N/A</td>
<td>++</td>
<td>Yes</td>
<td>Yes</td>
<td>9</td>
</tr>
<tr>
<td>Blanquero et al [22]</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
<td>Yes</td>
<td>N/A</td>
<td>++</td>
<td>Yes</td>
<td>Yes</td>
<td>9</td>
</tr>
<tr>
<td>Choi et al [23]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>CS</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
<td>Yes</td>
<td>N/A</td>
<td>++</td>
<td>Yes</td>
<td>Yes</td>
<td>9</td>
</tr>
<tr>
<td>Fatoye et al [24]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>16</td>
<td>CS</td>
<td>N/A</td>
<td>++</td>
<td>Yes</td>
<td>Yes</td>
<td>8</td>
</tr>
<tr>
<td>Fleischman et al [25]</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>16.6</td>
<td>Yes</td>
<td>N/A</td>
<td>++</td>
<td>Yes</td>
<td>Yes</td>
<td>9</td>
</tr>
<tr>
<td>Moffet et al [26]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>6.3</td>
<td>Yes</td>
<td>Yes</td>
<td>++</td>
<td>CS</td>
<td>Yes</td>
<td>9</td>
</tr>
<tr>
<td>Nelligan et al [27]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>12.6</td>
<td>Yes</td>
<td>Yes</td>
<td>++</td>
<td>Yes</td>
<td>Yes</td>
<td>11</td>
</tr>
<tr>
<td>Nelson et al [28]</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>CS</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>1</td>
<td>Yes</td>
<td>N/A</td>
<td>++</td>
<td>Yes</td>
<td>Yes</td>
<td>8</td>
</tr>
<tr>
<td>Özden et al [29]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>7</td>
<td>CS</td>
<td>N/A</td>
<td>++</td>
<td>Yes</td>
<td>Yes</td>
<td>9</td>
</tr>
<tr>
<td>Bennell et al [30]</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>10.1</td>
<td>Yes</td>
<td>N/A</td>
<td>_4 indifferent</td>
<td>Yes</td>
<td>Yes</td>
<td>9</td>
</tr>
<tr>
<td>Chhabra et al [31]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>0</td>
<td>Yes</td>
<td>N/A</td>
<td>+</td>
<td>Yes</td>
<td>Yes</td>
<td>8</td>
</tr>
<tr>
<td>Hardt et al [32]</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>CS</td>
<td>Yes</td>
<td>Yes</td>
<td>10</td>
<td>Yes</td>
<td>N/A</td>
<td>+</td>
<td>Yes</td>
<td>Yes</td>
<td>8</td>
</tr>
<tr>
<td>Hernando-Garijo et al [33]</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>18</td>
<td>Yes</td>
<td>N/A</td>
<td>+</td>
<td>Yes</td>
<td>No</td>
<td>8</td>
</tr>
<tr>
<td>Rodríguez Sánchez-Laulhé et al [34]</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>16</td>
<td>Yes</td>
<td>N/A</td>
<td>+</td>
<td>Yes</td>
<td>Yes</td>
<td>8</td>
</tr>
<tr>
<td>Tousignant et al [35]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>CS</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>15</td>
<td>CS</td>
<td>CS</td>
<td>+</td>
<td>Yes</td>
<td>Yes</td>
<td>8</td>
</tr>
<tr>
<td>Anan et al [36]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>CS</td>
<td>Yes</td>
<td>25.6</td>
<td>Yes</td>
<td>N/A</td>
<td>_c</td>
<td>CS</td>
<td>Yes</td>
<td>7</td>
</tr>
<tr>
<td>Bäcker et al [37]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>42</td>
<td>No</td>
<td>N/A</td>
<td>−</td>
<td>Yes</td>
<td>Yes</td>
<td>8</td>
</tr>
<tr>
<td>Bossen et al [38]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>24.6</td>
<td>Yes</td>
<td>N/A</td>
<td>−</td>
<td>Yes</td>
<td>Yes</td>
<td>9</td>
</tr>
<tr>
<td>Correia et al [39]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>36</td>
<td>No</td>
<td>N/A</td>
<td>−</td>
<td>Yes</td>
<td>Yes</td>
<td>8</td>
</tr>
<tr>
<td>del Pozo-Cruz et al [40]</td>
<td>Yes</td>
<td>CS</td>
<td>CS</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>10</td>
<td>Yes</td>
<td>Yes</td>
<td>−</td>
<td>Yes</td>
<td>CS</td>
<td>7</td>
</tr>
</tbody>
</table>
Study Characteristics

Table 4 presents the study characteristics of the 16 included studies according to the PICOS scheme.

The publication period ranged from 2011 [35] to 2022 [29,34], whereby 3 studies were published each in 2018 [21,31,32], 2019 [22,23,25], and 2021 [20,27,33]. The most common publication country was Australia, with 3 studies [27,28,30], followed by 2 studies each published by Spain [22,34], Canada [26,35], and the United States [21,25]. Across the 16 studies, 1840 patients were investigated, and the sample sizes ranged from 34 [33] to 350 patients [21]. The reported dropout rate was up to 18% (6/34) [33]. The average age of the patients varied from 38.5 [20] to 66 years [26,35], and the average female proportion across all studies reported was 61.35% (1008/1643) and varied from 51% (148/290) [25] to 100% (34/34) [33]. Regarding the localization of the musculoskeletal diseases, 7 studies were on knee-specific diseases such as total knee arthroplasty [25,26,32,35], knee osteoarthritis [21,27], and chronic knee pain [30]. This was followed by 4 studies on back-specific diseases such as low back pain [24,29,31] and chronic neck pain [20]. A total of 2 studies were on hand-specific diseases [22,34], whereas only 1 study was found for each full body [33], shoulder-specific diseases [23], and hip-specific diseases [28]. Regarding the patient-reported outcomes, 14, 12, 6, and 5 studies investigated pain [20-23,25-27,29-35], function [21,23,25-30,32-35], disability [20,22,24,29,31,34], and quality of life [20,27-30], respectively. In the 16 included studies, 26 different patient-reported outcomes were investigated. With regard to the digital health interventions, 7, 5, and 4 studies used app-based [20,22,23,28,31,32,34], web-based [21,25,27,29,30], and telerehabilitation-based physical exercises [24,26,33,35], respectively, whereby the duration of the digital health interventions ranged from 7 days [32] to 12 months [21]. As control groups, 9 studies used physiotherapy [20,21,23,25,26,28,31,32,35]; 4 studies used paper-based exercises [22,25,29,34]; 2 studies used internet-based information material [27,30]; and 1 study each used global postural re-education [20], waiting list [21], clinic-based McKenzie therapy [24], and no therapy [33].
<table>
<thead>
<tr>
<th>Study</th>
<th>Population and setting</th>
<th>Intervention and assessment</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abadiyan et al [20]</td>
<td>Sample size: n=60&lt;br&gt;Average age: 38.5 y&lt;br&gt;Female: 55%&lt;br&gt;Disease: chronic neck pain&lt;br&gt;Country: Iran</td>
<td>• I^4 (&quot;Seeb&quot; app+ GPR^5): n=20&lt;br&gt;C1^5 (GPR alone): n=20&lt;br&gt;C2^6 (conventional PT^7) n=20&lt;br&gt;Duration: 8 wk&lt;br&gt;Survey dates: baseline, 8 wk</td>
<td>• Drop out: 3%&lt;br&gt;I: 5%, C1: 5%, and C2: 0%&lt;br&gt;Pain: app+GPR&gt;PT&lt;br&gt;Neck disability index: app+GPR&gt;PT&lt;br&gt;Quality of life: GPR&gt;PT</td>
</tr>
<tr>
<td>Allen et al [21]</td>
<td>Sample size: n=350&lt;br&gt;Average age: 65.3 y&lt;br&gt;Female: 71.7%&lt;br&gt;Disease: knee osteoarthritis&lt;br&gt;Country: United States</td>
<td>• I (IBET^4): n=142&lt;br&gt;C1 (conventional PT): n=140&lt;br&gt;C2 (WL^5): n=68&lt;br&gt;Duration: 12 mo&lt;br&gt;Survey dates: baseline, 4, and 12 mo</td>
<td>• Drop out: 13.1%&lt;br&gt;I: 21.1%, C1: 7.9%, and C2: 7%&lt;br&gt;Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC^6) and other functional tests IBET=PT=WL</td>
</tr>
<tr>
<td>Bennell et al [30]</td>
<td>Sample size: n=148&lt;br&gt;Average age: 61.2 y&lt;br&gt;Female: 56.1%&lt;br&gt;Disease: chronic knee pain&lt;br&gt;Country: Australia</td>
<td>• I: internet-based education material supported by videoconferences with physiotherapist for home exercises (n=74)&lt;br&gt;Control group: internet-based education material only (n=74)&lt;br&gt;Duration: 9 mo&lt;br&gt;Survey dates: baseline, 3, and 9 mo</td>
<td>• Drop out: 10.1%&lt;br&gt;I: 11% and C: 10%&lt;br&gt;Pain and function: education+PT&gt;education&lt;br&gt;Quality of life: education+PT&gt;education</td>
</tr>
<tr>
<td>Blanquero et al [22]</td>
<td>Sample size: n=50&lt;br&gt;Average age: 50.0 y&lt;br&gt;Female: 82%&lt;br&gt;Disease: carpal tunnel release&lt;br&gt;Country: Spain</td>
<td>• I: ReHand app for physical home training (n=25)&lt;br&gt;Control group: paper and home-based physical exercise program (n=25)&lt;br&gt;Duration: 4 wk&lt;br&gt;Survey dates: baseline, 4 wk</td>
<td>• Drop out: 0%&lt;br&gt;Hand disability and pain: app based&gt;paper based</td>
</tr>
<tr>
<td>Chhabra et al [31]</td>
<td>Sample size: n=93&lt;br&gt;Average age: 41.2 y&lt;br&gt;Female: not reported&lt;br&gt;Disease: chronic low back pain&lt;br&gt;Country: India</td>
<td>• I: Snapcare app for physical home training (n=45)&lt;br&gt;Control group: conventional therapy (n=48)&lt;br&gt;Duration: 12 wk&lt;br&gt;Survey dates: baseline, 12 wk</td>
<td>• Drop out: 0%&lt;br&gt;Pain: app based&gt;conventional&lt;br&gt;Disability: app based&gt;conventional&lt;br&gt;Current Symptom Score: app based&gt;conventional</td>
</tr>
<tr>
<td>Choi et al [23]</td>
<td>Sample size: n=84&lt;br&gt;Average age: 54.5 y&lt;br&gt;Female: 68%&lt;br&gt;Disease: frozen shoulder&lt;br&gt;Country: Korea</td>
<td>• I: app (no name given) for physical home training (n=42)&lt;br&gt;Control group: conventional home-based self-exercises (n=42)&lt;br&gt;Duration: 3 mo&lt;br&gt;Survey dates: baseline, 4, 8, and 12 wk</td>
<td>• Drop out: 0%&lt;br&gt;Pain and range of motion: app based&gt;conventional</td>
</tr>
<tr>
<td>Fatoye et al [24]</td>
<td>Sample size: n=56&lt;br&gt;Average age: 48.7 y&lt;br&gt;Female: not reported&lt;br&gt;Disease: chronic low back pain&lt;br&gt;Country: Nigeria</td>
<td>• I: telerehabilitation home-based McKenzie therapy (TBMT^4; n=24)&lt;br&gt;Control group: clinic-based McKenzie therapy (CBMT^5; n=32)&lt;br&gt;Duration: 8 wk&lt;br&gt;Survey dates: baseline, 4, and 8 wk</td>
<td>• Drop out: 16%&lt;br&gt;I: 13% and C: 19%&lt;br&gt;Disability: TBMT=CBMT</td>
</tr>
<tr>
<td>Study</td>
<td>Population and setting</td>
<td>Intervention and assessment</td>
<td>Outcomes</td>
</tr>
<tr>
<td>-------</td>
<td>------------------------</td>
<td>-----------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>Fleischman et al [25]</td>
<td>Sample size: n=290&lt;br&gt;Average age: 65.0 y&lt;br&gt;Female: 51%&lt;br&gt;Disease: total knee arthroplasty&lt;br&gt;Country: United States</td>
<td>I: web-based PT at home (n=96)&lt;br&gt;C1: paper-based PT at home (n=97)&lt;br&gt;C2: formal outpatient PT (n=97)&lt;br&gt;Duration: 6 mo&lt;br&gt;Survey dates: baseline, 4-6 wk, 6 mo</td>
<td>Drop out: 15.9%&lt;br&gt;I: 17%, C1: 27%, and C2: 6%&lt;br&gt;Knee flexion and Knee Injury and Osteoarthritis Outcome Score (KOOS©): Web PT=paper PT=PT</td>
</tr>
<tr>
<td>Hardt et al [32]</td>
<td>Sample size: n=60&lt;br&gt;Average age: 65.9 y&lt;br&gt;Female: 57%&lt;br&gt;Disease: total knee arthroplasty&lt;br&gt;Country: Germany</td>
<td>I: PT+“GenuSport” app (PT+app; n=33)&lt;br&gt;Control group: PT (n=27)&lt;br&gt;Duration: 7 d&lt;br&gt;Survey dates: daily for 7 d</td>
<td>Drop out: 10%&lt;br&gt;I: 15% and C: 7%&lt;br&gt;Active range of motion, pain, function, KOOS, and Knee Society Score: PT+app&gt;PT</td>
</tr>
<tr>
<td>Hernando-Garjo et al [33]</td>
<td>Sample size: n=34&lt;br&gt;Average age: 53.4 y&lt;br&gt;Female: 100%&lt;br&gt;Disease: fibromyalgia&lt;br&gt;Country: Mexico</td>
<td>I: telerehabilitation with home-based aerobic exercises (n=17)&lt;br&gt;Control group: no additional intervention (n=17)&lt;br&gt;Duration: 15 wk&lt;br&gt;Survey dates: baseline, 15 wk</td>
<td>Drop out: 18%&lt;br&gt;I: 18% and C: 18%&lt;br&gt;Physical function: telerehabilitation=nothing</td>
</tr>
<tr>
<td>Moffet et al [26]</td>
<td>Sample size: n=205&lt;br&gt;Average age: 66.0 y&lt;br&gt;Female: 51.2%&lt;br&gt;Disease: total knee arthroplasty&lt;br&gt;Country: Canada</td>
<td>I: home-based telerehabilitation (n=104)&lt;br&gt;Control group: home-visiting PT (n=101)&lt;br&gt;Duration: 2 mo&lt;br&gt;Survey dates: baseline, 2, and 4 mo</td>
<td>Drop out: 6.3%&lt;br&gt;I: 9.6% and C: 2.9%&lt;br&gt;WOMAC, KOOS, function, and range of motion: telerehabilitation=PT</td>
</tr>
<tr>
<td>Nelligan et al [27]</td>
<td>Sample size: n=206&lt;br&gt;Average age: 60.0 y&lt;br&gt;Female: 61.2%&lt;br&gt;Disease: knee osteoarthritis&lt;br&gt;Country: Australia</td>
<td>I: website (information+active exercises) and text messages (n=103)&lt;br&gt;Control group: website with information only (n=103)&lt;br&gt;Duration: 24 wk&lt;br&gt;Survey dates: baseline, 24 wk</td>
<td>Drop out: 12.6%&lt;br&gt;I: 12.6% and C: 12.6%&lt;br&gt;Pain, WOMAC, KOOS, quality of life: website information+exercise&gt;website information only</td>
</tr>
<tr>
<td>Nelson et al [28]</td>
<td>Sample size: n=70&lt;br&gt;Average age: 64.5 y&lt;br&gt;Female: 63%&lt;br&gt;Disease: total hip replacement&lt;br&gt;Country: Australia</td>
<td>I: telerehabilitation and technology-based home exercise (n=35)&lt;br&gt;Control group: PT and paper-based home exercise (n=35)&lt;br&gt;Duration: 6 wk&lt;br&gt;Survey dates: baseline, 6 wk, 6 mo</td>
<td>Drop out: 1%&lt;br&gt;I: 3% and C: 0%&lt;br&gt;Quality of life and function: telerehabilitation+exercise=PT+exercise</td>
</tr>
<tr>
<td>Özden et al [29]</td>
<td>Sample size: n=50&lt;br&gt;Average age: 41.3 y&lt;br&gt;Female: 60%&lt;br&gt;Disease: low back pain&lt;br&gt;Country: turkey</td>
<td>I: telerehabilitation with Fizyoweb software (n=25)&lt;br&gt;Control group: same exercises with paper-based instructions (n=25)&lt;br&gt;Duration: 8 wk&lt;br&gt;Survey dates: baseline, 8 wk</td>
<td>Drop out: 7%&lt;br&gt;I: 7% and C: 7%&lt;br&gt;Pain, function, disability, and quality of life: telerehabilitation&gt;paper based</td>
</tr>
<tr>
<td>Rodríguez Sánchez-Laulhé et al [34]</td>
<td>Sample size: n=36&lt;br&gt;Average age: 59.8 y&lt;br&gt;Female: 61%&lt;br&gt;Disease: rheumatoid arthritis&lt;br&gt;Country: Spain</td>
<td>I: CareHand app for exercises and self-management and monitoring tools (n=14)&lt;br&gt;Control group: paper-based exercises (n=22)&lt;br&gt;Duration: 3 mo&lt;br&gt;Survey dates: baseline, 1, 3, and 6 mo</td>
<td>Drop out: 16%&lt;br&gt;I: 7% and C: 22%&lt;br&gt;Function: app based&gt;paper based&lt;br&gt;Pain and disability for upper extremity: app based&gt;paper based</td>
</tr>
<tr>
<td>Toussignant et al [35]</td>
<td>Sample size: n=48&lt;br&gt;Average age: 66.0 y&lt;br&gt;Female: not reported&lt;br&gt;Disease: total knee arthroplasty&lt;br&gt;Country: Canada</td>
<td>I: telerehabilitation by videoconference with a physiotherapist (n=24)&lt;br&gt;Control group: conventional PT (n=24)&lt;br&gt;Duration: 2 mo&lt;br&gt;Survey dates: baseline, 2, and 6 mo</td>
<td>Drop out: 15%&lt;br&gt;I: 12% and C: 17%&lt;br&gt;Disability: telerehabilitation=conventional PT&lt;br&gt;Function: telerehabilitation&gt;conventional PT&lt;br&gt;Functional activity, physical functioning, and physical pain: conventional PT&gt;telerehabilitation</td>
</tr>
</tbody>
</table>

**aI**: Intervention group.

**GPR**: Global postural re-education.
Synthesis of Results by Best Evidence Synthesis

Tables 5, 6, and 7 show the results of the best-evidence synthesis with regard to the cluster of the localization of the musculoskeletal diseases, patient-reported outcomes, and medical treatment types, respectively.

Regarding the localization of the musculoskeletal diseases, there was strong evidence that digital physical health exercises had a positive impact on the musculoskeletal diseases located in the back. Although moderate evidence was obtained for diseases located in the shoulder and hip, evidence for fibromyalgia (the entire body) is limited. Conflicting evidence was found for diseases located in the knee and hand. For the patient-reported outcomes, there was strong evidence that digital physical health exercises had a positive impact on disability and quality of life. Conflicting evidence was found for pain and function. Regarding the medical treatment types, operative and conservative therapies both achieved conflicting evidence. Figure 2 shows the evidence found across the 3 defined clusters for studies included in the best-evidence synthesis.

Table 5. Best-evidence synthesis for the localization of the musculoskeletal diseases.

<table>
<thead>
<tr>
<th>Localization</th>
<th>Study</th>
<th>Musculoskeletal disease</th>
<th>Results</th>
<th>Study quality</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Back</td>
<td>Abadiyan et al [20]</td>
<td>Chronic neck pain</td>
<td>+^a</td>
<td>High</td>
<td>Strong^b</td>
</tr>
<tr>
<td>Back</td>
<td>Chhabra et al [31]</td>
<td>Chronic low back pain</td>
<td>+</td>
<td>Acceptable</td>
<td>Strong^b</td>
</tr>
<tr>
<td>Back</td>
<td>Fatoye et al [24]</td>
<td>Chronic low back pain</td>
<td>^=c</td>
<td>High</td>
<td>Strong^b</td>
</tr>
<tr>
<td>Back</td>
<td>Ozden et al [29]</td>
<td>Chronic low back pain</td>
<td>+</td>
<td>High</td>
<td>Strong^b</td>
</tr>
<tr>
<td>Shoulder</td>
<td>Choi et al [23]</td>
<td>Frozen shoulder</td>
<td>=</td>
<td>High</td>
<td>Moderate</td>
</tr>
<tr>
<td>Hip</td>
<td>Nelson et al [28]</td>
<td>Total hip arthroplasty</td>
<td>=</td>
<td>High</td>
<td>Moderate</td>
</tr>
<tr>
<td>Full body</td>
<td>Hernando-Garijo et al [33]</td>
<td>Fibromyalgia</td>
<td>=</td>
<td>Acceptable</td>
<td>Limited</td>
</tr>
<tr>
<td>Knee</td>
<td>Allen et al [21]</td>
<td>Knee osteoarthritis</td>
<td>=</td>
<td>High</td>
<td>Conflicting^b</td>
</tr>
<tr>
<td>Knee</td>
<td>Bennell et al [30]</td>
<td>Chronic knee pain</td>
<td>+</td>
<td>Acceptable</td>
<td>Conflicting^b</td>
</tr>
<tr>
<td>Knee</td>
<td>Fleischman et al [25]</td>
<td>Total knee arthroplasty</td>
<td>=</td>
<td>High</td>
<td>Conflicting^b</td>
</tr>
<tr>
<td>Knee</td>
<td>Hardt et al [32]</td>
<td>Total knee arthroplasty</td>
<td>+</td>
<td>Acceptable</td>
<td>Conflicting^b</td>
</tr>
<tr>
<td>Knee</td>
<td>Moffet et al [26]</td>
<td>Total knee arthroplasty</td>
<td>=</td>
<td>High</td>
<td>Conflicting^b</td>
</tr>
<tr>
<td>Knee</td>
<td>Nelligan et al [27]</td>
<td>Knee osteoarthritis</td>
<td>+</td>
<td>High</td>
<td>Conflicting^b</td>
</tr>
<tr>
<td>Knee</td>
<td>Toussignant et al [35]</td>
<td>Total knee arthroplasty</td>
<td>+</td>
<td>Acceptable</td>
<td>Conflicting^b</td>
</tr>
<tr>
<td>Hand</td>
<td>Blanquero et al [22]</td>
<td>Carpal tunnel release</td>
<td>+</td>
<td>High</td>
<td>Conflicting^b</td>
</tr>
<tr>
<td>Hand</td>
<td>Rodríguez Sánchez–Laulhé et al [34]</td>
<td>Rheumatoid arthritis</td>
<td>=</td>
<td>Acceptable</td>
<td>Conflicting^b</td>
</tr>
</tbody>
</table>

^a>50% of the outcomes were significantly better in the intervention group than in the control group.

^bThe level of evidence was determined from all studies in the same localization.

^cNo statistically significant difference between the intervention and control groups.
Table 6. Best-evidence synthesis for the patient-reported outcomes of the musculoskeletal diseases.

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Study</th>
<th>Assessment tools</th>
<th>Results</th>
<th>Study quality</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disability</td>
<td>Abadiyan et al [20]</td>
<td>Neck Disability Index</td>
<td>+(^{a})</td>
<td>High</td>
<td>Strong(^{b})</td>
</tr>
<tr>
<td>Disability</td>
<td>Blanquero et al [22]</td>
<td>Disabilities of Arm, Shoulder and Hand Questionnaire</td>
<td>+</td>
<td>High</td>
<td>Strong(^{b})</td>
</tr>
<tr>
<td>Disability</td>
<td>Chhabra et al [31]</td>
<td>Modified Oswestry Disability Index</td>
<td>+</td>
<td>Acceptable</td>
<td>Strong(^{b})</td>
</tr>
<tr>
<td>Disability</td>
<td>Fatoye et al [24]</td>
<td>Oswestry Disability Index</td>
<td>=(^{c})</td>
<td>High</td>
<td>Strong(^{b})</td>
</tr>
<tr>
<td>Disability</td>
<td>Özden et al [29]</td>
<td>Oswestry Disability Index</td>
<td>+</td>
<td>High</td>
<td>Strong(^{b})</td>
</tr>
<tr>
<td>Disability</td>
<td>Rodríguez Sánchez-Laulhé et al [34]</td>
<td>Disabilities of Arm, Shoulder and Hand Questionnaire</td>
<td>=</td>
<td>Acceptable</td>
<td>Strong(^{b})</td>
</tr>
<tr>
<td>Quality of life</td>
<td>Abadiyan et al [20]</td>
<td>Short Form Health 36 Questionnaire</td>
<td>+</td>
<td>High</td>
<td>Strong(^{b})</td>
</tr>
<tr>
<td>Quality of life</td>
<td>Bennell et al [30]</td>
<td>Assessment of Quality of Life–2</td>
<td>+</td>
<td>Acceptable</td>
<td>Strong(^{b})</td>
</tr>
<tr>
<td>Quality of life</td>
<td>Nelligan et al [27]</td>
<td>Assessment of Quality of Life–6D</td>
<td>+</td>
<td>High</td>
<td>Strong(^{b})</td>
</tr>
<tr>
<td>Quality of life</td>
<td>Nelson et al [28]</td>
<td>Short Form Health 12 Questionnaire/European Quality of Life 5 Dimensions 5 Level Version</td>
<td>=</td>
<td>High</td>
<td>Strong(^{b})</td>
</tr>
<tr>
<td>Quality of life</td>
<td>Özden et al [29]</td>
<td>Short Form Health 36 Questionnaire</td>
<td>+</td>
<td>High</td>
<td>Strong(^{b})</td>
</tr>
<tr>
<td>Pain</td>
<td>Abadiyan et al [20]</td>
<td>Visual analog scale</td>
<td>+</td>
<td>High</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Pain</td>
<td>Allen et al [21]</td>
<td>WOMAC(^{d})</td>
<td>=</td>
<td>High</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Pain</td>
<td>Bennell et al [30]</td>
<td>Numeric rating scale</td>
<td>+</td>
<td>Acceptable</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Pain</td>
<td>Blanquero et al [22]</td>
<td>Visual analog scale</td>
<td>+</td>
<td>High</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Pain</td>
<td>Chhabra et al [31]</td>
<td>Numeric rating scale, Current Symptom Score</td>
<td>=</td>
<td>Acceptable</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Pain</td>
<td>Choi et al [23]</td>
<td>Visual analog scale</td>
<td>=</td>
<td>High</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Pain</td>
<td>Fleischman et al [25]</td>
<td>KOOS(^{e})</td>
<td>=</td>
<td>High</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Pain</td>
<td>Hardt et al [32]</td>
<td>Numeric rating scale</td>
<td>+</td>
<td>Acceptable</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Pain</td>
<td>Hernando-Garrio et al [33]</td>
<td>Visual analog scale</td>
<td>+</td>
<td>Acceptable</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Pain</td>
<td>Moffet et al [26]</td>
<td>WOMAC</td>
<td>=</td>
<td>High</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Pain</td>
<td>Nelligan et al [27]</td>
<td>Numeric rating scale</td>
<td>+</td>
<td>High</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Pain</td>
<td>Özden et al [29]</td>
<td>Visual analog scale</td>
<td>+</td>
<td>High</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Pain</td>
<td>Rodríguez Sánchez-Laulhé et al [34]</td>
<td>Visual analog scale</td>
<td>=</td>
<td>Acceptable</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Pain</td>
<td>Tousignant et al [35]</td>
<td>WOMAC</td>
<td>≠(^{f})</td>
<td>Acceptable</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Function</td>
<td>Allen et al [21]</td>
<td>WOMAC/30-s chair stand test/Timed up and go test/2-min step test, single-leg stand</td>
<td>=</td>
<td>High</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Function</td>
<td>Bennell et al [30]</td>
<td>WOMAC</td>
<td>+</td>
<td>Acceptable</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Function</td>
<td>Choi et al [23]</td>
<td>Range of motion</td>
<td>=</td>
<td>High</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Function</td>
<td>Fleischman et al [25]</td>
<td>KOOS</td>
<td>=</td>
<td>High</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Function</td>
<td>Hardt et al [32]</td>
<td>Range of motion/Timed up and go test/10-m walk test/30-s chair stand test/Knee Society Score</td>
<td>+</td>
<td>Acceptable</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Function</td>
<td>Hernando-Garrio et al [33]</td>
<td>Arm curl test, 6-min walk test</td>
<td>=</td>
<td>Acceptable</td>
<td>Conflicting(^{b})</td>
</tr>
<tr>
<td>Function</td>
<td>Moffet et al [26]</td>
<td>KOOS/Stair test/6-min walk test</td>
<td>=</td>
<td>High</td>
<td>Conflicting(^{b})</td>
</tr>
</tbody>
</table>
Table 7. Best-evidence synthesis for the medical treatment types.

<table>
<thead>
<tr>
<th>Therapy</th>
<th>Study</th>
<th>Musculoskeletal disease</th>
<th>Results</th>
<th>Study quality</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operative</td>
<td>Blanquero et al [22]</td>
<td>Carpal tunnel release</td>
<td>+</td>
<td>High</td>
<td>Conflicting</td>
</tr>
<tr>
<td>Operative</td>
<td>Fleischman et al [25]</td>
<td>Total knee arthroplasty</td>
<td>=</td>
<td>High</td>
<td>Conflicting</td>
</tr>
<tr>
<td>Operative</td>
<td>Hardt et al [32]</td>
<td>Total knee arthroplasty</td>
<td>+</td>
<td>Acceptable</td>
<td>Conflicting</td>
</tr>
<tr>
<td>Operative</td>
<td>Moffet et al [26]</td>
<td>Total knee arthroplasty</td>
<td>=</td>
<td>High</td>
<td>Conflicting</td>
</tr>
<tr>
<td>Operative</td>
<td>Nelson et al [28]</td>
<td>Total hip arthroplasty</td>
<td>=</td>
<td>High</td>
<td>Conflicting</td>
</tr>
<tr>
<td>Operative</td>
<td>Tousignant et al [35]</td>
<td>Total knee arthroplasty</td>
<td>+</td>
<td>Acceptable</td>
<td>Conflicting</td>
</tr>
<tr>
<td>Conservative</td>
<td>Abadiyan et al [20]</td>
<td>Chronic neck pain</td>
<td>+</td>
<td>High</td>
<td>Conflicting</td>
</tr>
<tr>
<td>Conservative</td>
<td>Allen et al [21]</td>
<td>Knee osteoarthritis</td>
<td>=</td>
<td>High</td>
<td>Conflicting</td>
</tr>
<tr>
<td>Conservative</td>
<td>Chhabra et al [31]</td>
<td>Chronic low back pain</td>
<td>+</td>
<td>Acceptable</td>
<td>Conflicting</td>
</tr>
<tr>
<td>Conservative</td>
<td>Choi et al [23]</td>
<td>Frozen shoulder</td>
<td>=</td>
<td>High</td>
<td>Conflicting</td>
</tr>
<tr>
<td>Conservative</td>
<td>Fatoye et al [24]</td>
<td>Chronic low back pain</td>
<td>=</td>
<td>High</td>
<td>Conflicting</td>
</tr>
<tr>
<td>Conservative</td>
<td>Hernando-Garijo et al [33]</td>
<td>Fibromyalgia</td>
<td>=</td>
<td>Acceptable</td>
<td>Conflicting</td>
</tr>
<tr>
<td>Conservative</td>
<td>Nelligan et al [27]</td>
<td>Knee osteoarthritis</td>
<td>+</td>
<td>High</td>
<td>Conflicting</td>
</tr>
<tr>
<td>Conservative</td>
<td>Özden et al [29]</td>
<td>Chronic low back pain</td>
<td>+</td>
<td>High</td>
<td>Conflicting</td>
</tr>
<tr>
<td>Conservative</td>
<td>Rodríguez Sánchez-Laulhé et al [34]</td>
<td>Rheumatoid Arthritis</td>
<td>=</td>
<td>Acceptable</td>
<td>Conflicting</td>
</tr>
</tbody>
</table>

*a* >50% of the outcomes were significantly better in the intervention group than in the control group.

*b* The level of evidence was determined from all studies in the same therapy.

*c* No statistically significant difference between the intervention and control groups.

dOMA: Western Ontario and McMaster Universities Osteoarthritis Index.

eKOOS: Knee Injury and Osteoarthritis Outcome Score.

"Conflicting" indicates a lack of consensus among the studies on the effectiveness of the treatment.
Discussion
Principal Findings
This systematic review aimed to evaluate the impact of digital physical health exercises on patients with musculoskeletal diseases concerning the localization of the musculoskeletal disease, patient-reported outcomes, and medical treatment types. In addition, a best-evidence synthesis was conducted to estimate the direction and strength of the existing evidence. The main findings were that (1) strong evidence was found for a positive impact on musculoskeletal diseases located in the back and on the patient-reported outcomes of disability and quality of life (Figure 2). Our findings are partly supported by a previous systematic review with a meta-analysis [13] showing moderate-quality evidence for the positive impact on the patient-reported outcome of disability. In contrast to the previous review [13] and to another systematic review [12], conflicting evidence for the patient-reported outcomes of pain and function was found. It should be noted that 1 study [12] found some clinical benefits for pain and function but did not conduct an evidence synthesis or a meta-analysis. In addition, the outcomes of pain and function represent health-related outcomes, and the outcomes of disability and quality of life are the resulting consequences. Therefore, pain acts as a protective mechanism and can lead to disability [50]. With appropriate exercises,

![Figure 2](image-url)

Figure 2. Visualization of the evidence found across the 3 defined clusters for studies included in the best-evidence synthesis. ↑↑↑: strong evidence, ↑↑: moderate evidence, ↑: limited evidence, and ↑↓: conflicting evidence.

<table>
<thead>
<tr>
<th>Study characteristics</th>
<th>Digital health interventions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General information</strong></td>
<td><strong>Duration:</strong> 7 days to 12 months</td>
</tr>
<tr>
<td>Included studies: n=16</td>
<td><strong>Type:</strong> App-based interventions (n=7)</td>
</tr>
<tr>
<td>Sample size: n=34 to n=350</td>
<td>Internet-based exercises (n=5)</td>
</tr>
<tr>
<td>Total patients: n=1840 (female 61.35%)</td>
<td>Telerehabilitation (n=4)</td>
</tr>
<tr>
<td>Patient age: 38.5 to 66.0 years</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evidences for intervention effects</th>
<th>Investigated outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathology localization</td>
<td>Disability ↑↑↑ (n=6)</td>
</tr>
<tr>
<td>Entire body ↑ (n=1)</td>
<td>Quality of life ↑↑↑ (n=5)</td>
</tr>
<tr>
<td>Shoulder ↑ (n=1)</td>
<td>Pain ↑ (n=14)</td>
</tr>
<tr>
<td>Hand ↑ (n=2)</td>
<td>Function ↑ (n=12)</td>
</tr>
<tr>
<td>Back ↑↑↑ (n=4)</td>
<td></td>
</tr>
<tr>
<td>Hip ↑ (n=1)</td>
<td></td>
</tr>
<tr>
<td>Knee ↑ (n=7)</td>
<td></td>
</tr>
</tbody>
</table>

| Medical treatment type | |
|-----------------------| |
| Operative ↑ (n=6)     | |
| Conservative ↑ (n=10) | |

The first main finding was that strong evidence was found for a positive impact on musculoskeletal diseases located in the back and on the patient-reported outcomes of disability and quality of life (Figure 2). Our findings are partly supported by a previous systematic review with a meta-analysis [13] showing moderate-quality evidence for the positive impact on the patient-reported outcome of disability. In contrast to the previous review [13] and to another systematic review [12], conflicting evidence for the patient-reported outcomes of pain and function was found. It should be noted that 1 study [12] found some clinical benefits for pain and function but did not conduct an evidence synthesis or a meta-analysis. In addition, the outcomes of pain and function represent health-related outcomes, and the outcomes of disability and quality of life are the resulting consequences. Therefore, pain acts as a protective mechanism and can lead to disability [50]. With appropriate exercises,
patients learn to compensate for their disabilities [20,30], whereas exercise alone can provoke pain [51]. As disabilities are part of the concept of health-related quality of life [52], these outcomes are mutually dependent, and identical strong evidence is plausible.

In addition, it should be mentioned that both previous systematic reviews included all types of digital health interventions, and we explicitly focused our systematic review on the impact of digital physical health exercises. Regarding this, our findings add that this type of intervention shows strong evidence to have an overall positive impact on the musculoskeletal diseases located in the back, independent of the investigated outcomes [20,24,29,31]. Back-related musculoskeletal diseases usually arise because of muscular causes and are often caused by a lack of physical activity [53]. Participants recruited in back pain–related studies are often middle-aged and have an office occupation [20]. The use of digital physical health exercises in such patients can be considered highly effective because of the increased physical activity targeting muscle strengthening and the teaching of exercise techniques [20,24,29,31]. Overall, the application of digital physical health exercise in patients with musculoskeletal diseases shows versatile positive effects, especially for musculoskeletal diseases located in the back and for the improvement of disabilities and quality of life. However, the type of digital health interventions seems to influence the effects on the specific patient-reported outcome, and more studies to investigate this relationship are needed.

The second main finding was that moderate evidence was found for a beneficial effect on musculoskeletal diseases of the shoulder and hip (Figure 2). As this systematic review is the first to evaluate the association between digital physical health exercises and different localizations of musculoskeletal diseases, no evidence levels from previous research is available for clarification. There is only 1 other systematic review on the effectiveness of digital health interventions for total hip arthroplasty [54]. The review found no significant improvements in the studied patient-reported outcomes. For the shoulder, another systematic review examined the effectiveness of telerehabilitation for musculoskeletal diseases compared with normal in-person physiotherapy [55] and found very low to low evidence. In this context, our findings suggest that digital physical health exercises may also be effective in treating musculoskeletal diseases of the shoulder and hip. However, it should be noted that only 1 study each was found for shoulder- and hip-specific musculoskeletal diseases, whereas several studies were found for back- or knee-specific musculoskeletal diseases (Table 5). Therefore, our results must be interpreted with caution, as a small number of high-quality studies may result in stronger evidence, according to the definitions of the best-evidence synthesis [18], than the presence of many lower-quality studies. Overall, the results demonstrated that digital physical health exercises could have a positive effect on a variety of health-related outcomes, regardless of the localization of the musculoskeletal diseases. However, the number of studies investigating the relationship between the effectiveness of digital health interventions and the localization of musculoskeletal diseases is small, and more studies are needed, especially for localizations other than the knee and back.

An additional interesting finding is the conflicting evidence in the medical treatment types concerning operative and conservative approaches (Figure 2). Although the underlying reasons remain unknown, it can be stated that the operative treatment (ie, carpal tunnel release and total knee arthroplasty) has no impact on the overall stimulus-response mechanism of the digital intervention, requiring further studies for clarification.

Limitations

Although this systematic review increases knowledge of the positive impacts of digital physical health exercises on musculoskeletal diseases, there are a few limitations. Because of the heterogeneity of the included studies (eg, different numbers of patients, interventions, body regions, and control groups), a meta-analysis could not be performed. Instead, and as an established alternative approach, a best-evidence synthesis [18] was used. A strength of this approach is that it is possible to estimate an evidence level for various categories despite the large study heterogeneity. However, a limitation is that no quantitative analysis (eg, in terms of statistical significance) can be conducted [56]. An additional limitation of our review is that we did not register the study plan in PROSPERO a priori. The reason is that according to the PRISMA guidelines, registration is currently recommended but not mandatory [15]. Furthermore, all studies that included digital health interventions beyond active exercises were not included. Therefore, some studies could be lost, but the aspect of physical exercise as an established clinical treatment for musculoskeletal diseases could be focused on for the first time.

Conclusions

There is strong to moderate evidence for the beneficial impact of digital physical health exercises for musculoskeletal diseases located in the back, shoulder, and hip. There is limited or conflicting evidence for other localizations. In addition, strong evidence was found for the patient-reported outcomes of disability and quality of life, whereas conflicting evidence exists for other commonly investigated patient-reported outcomes such as pain and function. Thus, digital physical health exercises could have a positive effect on a variety of health-related outcomes of musculoskeletal diseases. To implement digital physical health exercises in evidence-based medicine for musculoskeletal diseases, more high-quality randomized controlled trials are needed to clarify the relationship between the impact of digital physical health exercises and clinically relevant factors such as localization, patient-reported outcomes, and medical treatment types.
Acknowledgments

The study was funded by the Open Access Publishing Fund of Leipzig University supported by the German Research Foundation within the Open Access Publication Funding program. The authors acknowledge the support from Leipzig University for Open Access Publishing.

Authors' Contributions

JN, FW, and MWH conceived the study and were involved in the conceptual process. JN performed the literature search and data analysis, and MWH validated these steps. JN wrote a raw version of the manuscript, and FW and MWH revised the raw version and wrote the final manuscript together with JN. CG critically revised the final manuscript and made additional changes. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA 2020 Checklist.

References

5. Musculoskeletal health status in Europe - summery report. v5.0. eumusc.net. URL: http://www.eumusc.net/workpackages_wp4.cfm [accessed 2023-07-03]


with rheumatoid arthritis of the hands: randomized controlled trial. JMIR Mhealth Uhealth 2022 Apr 07;10(4):e35462 [FREE Full text] [doi: 10.2196/35462] [Medline: 35389367]


Abbreviations

- **PICOS**: Population Intervention Comparison Outcome Study design
- **PRISMA**: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

©Johanna Nagel, Florian Wegener, Casper Grim, Matthias Wilhelm Hoppe. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 23.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Review

Mobile and Computer-Based Applications for Rehabilitation Monitoring and Self-Management After Knee Arthroplasty: Scoping Review

Sabhya Pritwani¹, BDS, MPH; Purnima Shrivastava¹, MPT; Shruti Pandey¹, BPT, MPH; Ajit Kumar², MPT; Rajesh Malhotra³, MS; Ralph Maddison³, PhD; Niveditha Devasenapathy¹, MBBS, PhD

¹Department of Research & Development, The George Institute for Global Health India, Delhi, India
²Department of Orthopaedics, All India Institute of Medical Sciences, Delhi, India
³Department of School of Exercise & Nutrition, Institute for Physical Activity and Nutrition, Deakin University, Geelong, Australia

Corresponding Author:
Niveditha Devasenapathy, MBBS, PhD
Department of Research & Development
The George Institute for Global Health India
308, Third floor, Elegance Tower
Plot no 8, Jasola District Centre
Delhi, 110025
India
Phone: 91 11 4158 8091 93
Email: NDevasenapathy@georgeinstitute.org.in

Abstract

Background: Successful post-knee replacement rehabilitation requires adequate access to health information, social support, and periodic monitoring by a health professional. Mobile health (mHealth) and computer-based technologies are used for rehabilitation and remote monitoring. The extent of technology use and its function in post-knee replacement rehabilitation care in low and middle-income settings are unknown.

Objective: To inform future mHealth intervention development, we conducted a scoping review to map the features and functionality of existing technologies and determine users’ perspectives on telerehabilitation and technology for self-management.

Methods: We followed the Joanna Briggs Institute methodology for scoping reviews. We searched the Embase, Medline, PsycINFO via OVID, and Cochrane Central Register of Controlled Trials databases for manuscripts published from 2001 onward. We included original research articles reporting the use of mobile or computer-based technologies by patients, health care providers, researchers, or family members. Studies were divided into the following 3 categories based on the purpose: validation studies, clinical evaluation, and end user feedback. We extracted general information on study design, technology features, proposed function, and perspectives of health care providers and patients. The protocol for this review is accessible in the Open Science Framework.

Results: Of the 5960 articles, 158 that reported from high-income settings contributed to the qualitative summary (64 studies on mHealth or telerehabilitation programs, 28 validation studies, 38 studies describing users’ perceptions). The highest numbers of studies were from Europe or the United Kingdom and North America regarding the use of a mobile app with or without wearables and reported mainly in the last decade. No studies were from low and middle-income settings. The primary functions of technology for remote rehabilitation were education to aid recovery and enable regular, appropriate exercises; monitoring progress of pain (n=19), activity (n=20), and exercise adherence (n=30); 1 or 2-way communication with health care professionals to facilitate the continuum of care (n=51); and goal setting (n=23). Assessment of range of motion (n=16) and gait analysis (n=10) were the commonly validated technologies developed to incorporate into a future rehabilitation program. Few studies (n=14) reported end user involvement during the development stage. We summarized the reasons for satisfaction and dissatisfaction among users across various technologies.

Conclusions: Several existing mobile and computer-based technologies facilitate post-knee replacement rehabilitation care for patients and health care providers. However, they are limited to high-income settings and may not be extrapolated to low-income settings. A systematic needs assessment of patients undergoing knee replacement and health care providers involved in rehabilitation,
involving end users at all stages of development and evaluation, with clear reporting of the development and clinical evaluation can make post-knee replacement rehabilitation care in resource-poor settings accessible and cost-effective.

**(JMIR Mhealth Uhealth 2024;12:e47843) doi:10.2196/47843**

**KEYWORDS**

knee arthroplasty; telerehabilitation; mHealth; rehabilitation; monitoring; self-management; knee; arthroplasty; social support; mHealth intervention; development; scoping review; knee replacement

**Introduction**

Knee arthroplasty is the gold standard treatment for end-stage osteoarthritis when conservative treatments fail to relieve symptoms [1]. Wound care and postarthroplasty physiotherapy are essential components of this treatment. Poor adherence to physiotherapy could delay the recovery and lead to suboptimal functional outcomes [2]. Beyond in-hospital clinical care and initiation of physical therapy before discharge, continued and reliable access to information, support from health care providers, awareness of the recovery pathway, easy access to rehabilitation centers, and periodic monitoring are influential factors for optimal recovery [3-6]. In addition to an uneventful surgery, postarthroplasty outcomes are associated with several patient-related factors such as their preoperative physical and mental state, comorbidities, social support, and socioeconomic status, emphasizing the need for personalized approaches [7]. Hence, monitoring of the rehabilitation phase is essential, whether at clinics, in rehabilitation units, or at home [8-10].

Technology-assisted remote monitoring methods are increasingly being advocated in high-income countries. There is low to moderate-quality evidence on the superiority of telehealth compared with unsupported home-based rehabilitation and noninferiority compared with clinic-based monitoring with respect to range of motion (ROM), pain, function, quality of life, and cost-effectiveness at 3 months between clinic-based and home-based rehabilitation strategies using technology [11-17]. Hence, current evidence supports the adaptation of technology-based rehabilitation as feasible, as safe, and as good as clinic-based monitoring with an additional benefit of saving out-of-pocket expenditure. Technology-based approaches are diverse, varying from telehealth [17] to virtual reality techniques [13] aimed at improving adherence to physical therapy and facilitating remote monitoring [12] of patient progress during the post-acute rehabilitation phase [18].

Therefore, the aim of this scoping review was to summarize the extent, range, and nature of technology used for provision of rehabilitation or to monitor progress following knee arthroplasty. This scoping review aimed to address the following objectives:

1. To map the characteristic features and functionality of the technologies, guiding or theoretical framework for designing the technology, and evaluation methodologies of mobile technology–based apps for rehabilitation monitoring and self-management following knee arthroplasty
2. To understand the patient and physical therapist perspectives regarding the use of mobile technology–based apps for rehabilitation monitoring and self-management following knee arthroplasty

To our knowledge, there are no existing scoping reviews that address our aims [19]. The information from this review will help us and other researchers make an informed decision on future mobile health (mHealth) interventions for monitoring post-knee arthroplasty rehabilitation care by physiotherapists and orthopedic surgeons and to promote self-management by individuals. This review will also help highlight existing gaps in the context of low and middle-income countries (LMICs).

**Methods**

We conducted this scoping review following the Joanna Briggs Institute (JBI) methodology for JBI Scoping Reviews [20] and consulted the PRISMA-ScR (Preferred Reporting Items for Systematic Review and Meta-Analyses Extension for Scoping Reviews) checklist for reporting [21]. The protocol was registered at the Open Science Framework [22].

**Data Sources and Searches**

To identify relevant studies, an electronic database literature search was conducted in the Embase, Medline, PsycINFO via OVID, and Cochrane Central Register of Controlled Trials (CENTRAL) databases using the following key terms: "Knee arthroplasty OR Knee replacement," “mobile,” “web,” “remote sensor,” “computer,” “telerehabilitation,” and “m-health” (Tables S1 and S2 in Multimedia Appendix 1). The search was executed in October 2021 and updated in August 2023. The search was restricted to 2001 onward. There were no language restrictions during the search. We searched the reference list of included articles to identify potentially eligible studies.

**Study Selection**

Predefined inclusion criteria were articles reporting the use of mobile or computer apps or any other technologies such as sensor-based devices for delivering or monitoring rehabilitation either scheduled or following knee joint replacement. We also included proof- of-concept papers that described the development process of a mobile or technology-based solution for rehabilitation. The purpose of technology could be for a health care provider to monitor rehabilitation adherence, to aid patient-health care provider communication, to promote self-management, to act as reminders, or to act as a source of education or any other function that is aimed at rehabilitation care following knee replacement. The app or technology could be used by patients, health care providers, researchers, or a family member. Included studies could have been conducted in the community or home for any clinical setting in any geographic region. The studies were required to be original research articles, and we included experimental and observational studies using quantitative or qualitative research methods. Reviews (narrative or systematic reviews), non-English
articles, and articles without abstracts or full texts were excluded.

**Data Extraction**

Screening of manuscript titles and abstracts was conducted by 2 independent reviewers using the web app Rayyan [23]. Prior to screening, reviewers discussed inclusion and exclusion criteria to ensure consistency between individuals. Two reviewers assessed the eligibility of the full text, and disagreements were resolved by discussion. Systematic reviews were not included in the review but were used to obtain potentially relevant references. Multiple publications originating from a single technology were grouped and presented as 1 study.

For data charting purposes, the studies were divided into the following 3 categories: (1) studies that had no rehabilitation program but included an app or a technology to assess ROM or gait and were validation studies, (2) studies reporting the use of a mobile or computer app or a telehealth delivery platform for a rehabilitation program with or without sensor-based devices and wearable sensors, (3) studies that reported end users’ perceptions of the technology used for rehabilitation monitoring. Data on the general information for the studies, features of the technology, the proposed function, and perspectives of health care providers and patients were extracted and entered in Microsoft Excel. If only the protocol of a planned study was available, there was no information on clinical evaluation, or the study included <6 individuals, we did not extract data beyond the general information.

**Results**

**Search Results**

The database search, including the ad hoc search, yielded 5960 articles. Of these articles, 300 articles were considered potentially relevant. Of these, 158 articles were included for data extraction, 131 articles were excluded, and 11 articles were not available (Table S3 in Multimedia Appendix 1). Of the 158 articles, 91 articles (64 studies) reported the clinical evaluation of a technology-based rehabilitation program, 29 articles (28 studies) reported the validation or a proof of concept of technology intended to be used for rehabilitation, and 13 articles were protocols of evaluation studies. In addition, 25 articles reported end users’ perceptions on technology (Figure 1) as stand-alone articles or part of clinical evaluation studies (n=13), totaling 38 studies. The 13 studies that reported the perceptions of technology that were also included in rehabilitation program studies were removed from the final list of included full-text articles to avoid double counting.

**Figure 1.** Process of identifying and including studies according to PRISMA-ScR (Preferred Reporting Items for Systematic Review and Meta-Analyses Extension for Scoping Reviews).
Technology for Rehabilitation

Characteristics of the 105 Studies

Studies were reported from Europe and the United Kingdom (n=45) [11, 24-60, 62-66, 169, 170], North America (n=34) [6, 67-99], Australia and New Zealand (n=10) [100-109], and Asia (n=16) [110-125]. None of the studies were from LMICs.

Reports of mobile-based technologies represented the highest number (54/105, 51.4%) [6, 25-27, 31, 32, 36, 37, 41-50, 52, 55, 56, 58-60, 62, 64, 67-69, 74, 75, 78, 80, 82, 83, 90-92, 95, 101, 103, 106-108, 112, 113, 117, 121-126, 169], followed by computer applications (31/105, 29.5%) [24, 29, 30, 33-35, 39, 53, 54, 57, 65, 66, 70, 73, 76, 79, 84-86, 89, 93, 97, 98, 100, 102, 104, 111, 114, 116, 120, 127], and tele/video/web conferencing (20/105, 19%) for rehabilitation monitoring [28, 38, 51, 63, 71, 72, 77, 81, 87, 88, 94, 96, 99, 105, 109, 110, 115, 118, 119, 170]. The highest use of mobile apps associated with or without a wearable was in Europe and the United Kingdom, followed by North America. Tele/video/web conferencing was used across regions, with the highest number in North America (Figure 2).

Validation Studies

There were 28 validation studies. Studies that validated stand-alone technologies included those to assess ROM (n=16) [24, 29, 43, 45, 48, 49, 52, 68, 73, 84, 100, 101, 111, 112, 116, 125] or gait or posture (n=10) [29, 30, 33, 53, 57, 74, 89, 97, 102, 124], and 2 studies involved technologies to monitor exercises [98, 114]. The technologies involved were either wearables (n=20) [24, 29, 30, 33, 45, 48, 49, 57, 68, 73, 84, 89, 97, 98, 100, 111, 112, 114, 116, 125], sensor-based devices (nonwearables; n=4) [53, 66, 102, 124], or inbuilt sensors available within a smartphone (n=4) [43, 52, 74, 101] (Table S4 in Multimedia Appendix 1).

In terms of study design, 9 were cross-sectional studies [33, 48, 52, 57, 84, 89, 97, 101, 116], 7 were cohort or longitudinal studies [45, 53, 68, 74, 100, 111, 125], 5 were pre-post studies [29, 30, 43, 73, 102], 1 was an uncontrolled trial [112], 1 was a randomized controlled trial (RCT) [66], and 5 were articles that described the proof of concept or development plan for the technologies [24, 49, 98, 114, 124]. The participant sample size ranged from 1 to 60. Most studies reported reliability between a standard or universal goniometer and smartphone app goniometry and the clinical evaluation of sensors to measure gait parameters (Table S4 in Multimedia Appendix 1). In 7 studies, gait was measured using sensors provided by a health care provider in a hospital setting [29, 33, 57, 74, 89, 97, 102], and 3 studies did not describe the measurement setting [30, 53, 124].

Clinical Evaluation Studies

There were 64 clinical evaluation studies. The technology consisted of a mobile or computer app with a wearable device (n=18) [6, 26, 31, 32, 39, 44, 46, 50, 54, 64, 67, 69, 90, 92, 95, 106, 108, 169], a mobile or computer app with a sensor-based device (n=13) [25, 34, 35, 40, 42, 65, 70, 76, 79, 85, 86, 93, 120], only a mobile app (n=14) [36, 37, 55, 56, 62, 75, 78, 80, 83, 107, 113, 117, 123, 128], or only telephone or videoconferencing (n=19) for remote monitoring [28, 38, 51, 63, 71, 72, 77, 81, 88, 94, 96, 99, 105, 109, 110, 115, 118, 119, 170]. Of the studies that used a mobile app, 9 studies were developed only for iOS [55, 67, 69, 71, 77, 92, 106, 107, 109], 1 was an Android app [42], 7 were for both Android and iOS devices [28, 36, 56, 88, 108, 115, 117], and 21 studies did not specify the platform (Multimedia Appendix 2). A web-based clinician portal for synchronous or asynchronous remote monitoring of patients was reported by 36 studies (Table 1). The number of published studies and the intervention arm sample size (ranging from 7 to 2292), especially for those that included wearable sensors and mobile apps, steadily increased over the last 2 decades (Figure 3).
<table>
<thead>
<tr>
<th>First author, year</th>
<th>Web portal Monitoring</th>
<th>Devices Wearables</th>
<th>Sensor-based devices</th>
<th>Peer</th>
<th>App name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexander, 2023 [67]</td>
<td>✓</td>
<td>Apple Watch</td>
<td>—</td>
<td>—</td>
<td>mymobility</td>
</tr>
<tr>
<td>An, 2021 [110]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Argent, 2019 [169]</td>
<td>—</td>
<td>IMU</td>
<td>Avatar</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Bade, 2020 [166]</td>
<td>—</td>
<td>In-shoe sensors</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Bell, 2020 [90]</td>
<td>✓</td>
<td>InterACTION IMU</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Bini, 2017 [71]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Capture proof</td>
</tr>
<tr>
<td>Blasco, 2022 [28]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>WeChat app</td>
</tr>
<tr>
<td>Campbell, 2019 [72]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>StreaMD</td>
</tr>
<tr>
<td>Chughtai, 2018 [76]</td>
<td>✓</td>
<td>—</td>
<td>VERA</td>
<td>—</td>
<td>VERA</td>
</tr>
<tr>
<td>Chughtai, 2019 [75]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>PReHab</td>
</tr>
<tr>
<td>Colomina, 2021 [31]</td>
<td>✓</td>
<td>Fitbit Flex 2</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>De Berardinis, 2022 [26]</td>
<td>✓</td>
<td>Magnetic sensors with Velcro bands</td>
<td>—</td>
<td>—</td>
<td>kari</td>
</tr>
<tr>
<td>Doiron-Cadrin, 2020 [77]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Reacts Lite</td>
</tr>
<tr>
<td>Duong, 2023 [106]</td>
<td>✓</td>
<td>Fitbit, ActivPal, Go-niometer Pro</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Eichler, 2019 [34]</td>
<td>✓</td>
<td>Kinect sensor</td>
<td>—</td>
<td>—</td>
<td>MainReha app</td>
</tr>
<tr>
<td>Fung, 2012 [79]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Gohir, 2021 [36]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>i-Beat app</td>
</tr>
<tr>
<td>Gray, 2022 [37]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Digital Joint School using GoWell health program</td>
</tr>
<tr>
<td>Gunduz, 2021 [38]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Hadamus, 2022 [40]</td>
<td>—</td>
<td>—</td>
<td>Kinetic camera</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Hardwick-Morris, 2022 [107]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Physitrack</td>
</tr>
<tr>
<td>Hong, 2022 [80]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Digital Musculoskeletal Surgical Care Program app</td>
</tr>
<tr>
<td>Huang, 2017 [113]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Yishu</td>
</tr>
<tr>
<td>Janhunen, 2023 [42]</td>
<td>—</td>
<td>—</td>
<td>Kinect sensor with TV and tablet</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Juhl, 2016 [44]</td>
<td>✓</td>
<td>IMU</td>
<td>—</td>
<td>—</td>
<td>ICURA app</td>
</tr>
<tr>
<td>Klement, 2019 [81]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Knapp, 2021 [83]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Kramer, 2003 [99]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>First author, year</td>
<td>Web portal Monitoring</td>
<td>Devices Wearables</td>
<td>Sensor-based devices</td>
<td>Peer</td>
<td>App name</td>
</tr>
<tr>
<td>--------------------</td>
<td>-----------------------</td>
<td>-------------------</td>
<td>----------------------</td>
<td>------</td>
<td>----------</td>
</tr>
<tr>
<td>Kuether, 2019 [85]</td>
<td>✓</td>
<td>—</td>
<td>VERA</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Lam, 2016 [86]</td>
<td>✓</td>
<td>IMU</td>
<td>—</td>
<td>—</td>
<td>ReHab system</td>
</tr>
<tr>
<td>Lebleu, 2023 [46]</td>
<td>✓</td>
<td>Activity tracker</td>
<td>—</td>
<td>—</td>
<td>moveUP Therapy</td>
</tr>
<tr>
<td>LeBrun, 2022 [78]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>MyChart app</td>
</tr>
<tr>
<td>Li, 2023 [115]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
</tr>
<tr>
<td>Lu, 2021 [117]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
</tr>
<tr>
<td>McDonall, 2022 [147]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Milliren, 2022 [88]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Ubicare Smart X</td>
</tr>
<tr>
<td>Nuevo, 2023 [50]</td>
<td>✓</td>
<td>Accelerometer, gyroscope, magnetometer (DyCare)</td>
<td>—</td>
<td>—</td>
<td>ReHab</td>
</tr>
<tr>
<td>Osterloh, 2023 [51]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>YOLii</td>
</tr>
<tr>
<td>Park, 2017 [118]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Park, 2023 [119]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Piquerás, 2013 [54]</td>
<td>✓</td>
<td>(WAGYRO)</td>
<td>Avatar</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Pournajaf, 2022 [65]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Pronk, 2020 [55]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Pain coach app</td>
</tr>
<tr>
<td>Prvu Bettger, 2019 [70]</td>
<td>✓</td>
<td>—</td>
<td>VERA</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Ramkumar, 2019 [92]</td>
<td>✓</td>
<td>Motion sensors</td>
<td>—</td>
<td>—</td>
<td>Focus ventures RPM</td>
</tr>
<tr>
<td>Russell, 2011 [105]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Scheper, 2019 [56]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Woundcare app</td>
</tr>
<tr>
<td>Su, 2015 [120]</td>
<td>—</td>
<td>Kinect sensor</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Summers, 2023 [93]</td>
<td>✓</td>
<td>—</td>
<td>Electro-mechanical device</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Szöts, 2016 [170]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Timmers, 2019 [62]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>The Patient Journey app</td>
</tr>
<tr>
<td>Torpil, 2022 [63]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Toussignant, 2011 [94]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Tripuraneni, 2021 [95]</td>
<td>✓</td>
<td>Smart watch</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>van Dijk-Huisman, 2020 [64]</td>
<td>✓</td>
<td>MOX activity monitor</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Visperas, 2021 [96]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Wang, 2023 [121]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>WeChat app</td>
</tr>
<tr>
<td>Zhang, 2021 [123]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>WeChat app</td>
</tr>
</tbody>
</table>

*aNot applicable.

*bIMU: inertial motion unit.

*cVERA: Virtual Exercise Rehabilitation Assistant.
Although most studies described the features and functionality of the technology to deliver the intervention, they lacked details about the technological aspects that could benefit future researchers. For example, 2 studies [31, 86] explicitly reported information on the software, programming language and tools used, or calibration procedures either along with the main study or cited the article that described the development phase. Information on conceptualization of the technology-based intervention was described in only 3 studies [37, 38, 120]. End users’ involvement was typically late during the development phase (ie, prototype stage) and involved refining the functionalities and features of technology [32, 51, 64, 65, 75-77, 80, 86, 88, 115, 118, 169] prior to deployment. Patient feedback on their needs at an early development phase was reported only by Blasco et al [28].

Clinical effectiveness was tested using an RCT design in 57.8% (37/64) of the studies [6, 25, 28, 34-36, 39, 42, 44, 45, 51, 54, 55, 62, 63, 65, 67, 70-72, 77, 79, 90, 94-96, 99, 105, 106, 108-110, 117-119, 128, 170], and the rest of the studies were either retrospective comparative cohort studies (n=3) [26, 37, 78], uncontrolled cohort studies (n=9) [46, 56, 69, 76, 81, 83, 85, 92, 169], cross-sectional studies (n=1) [86], or non-RCTs (n=14) [31, 32, 38, 40, 64, 75, 80, 88, 93, 107, 113, 115, 120, 123].

We found 13 study protocols, of which 12 were RCTs published between 2013 and 2023 [11, 41, 47, 58-60, 82, 87, 91, 103, 104, 122], for which we could not find a published report and hence were not included in this summary. User experience was measured in trials using quantitative (n=9) [32, 34, 38, 50, 55, 94, 96, 105, 123], qualitative (n=2) [61, 128], and mixed methods (n=3) [39, 90, 109] approaches.

**Application Functionality for Rehabilitation Programs**

The key functionalities of the telerehabilitation technologies extracted from 64 studies are summarized under 4 themes, namely education and enablement, monitoring progress, communication, and goal setting (Table 2).
Table 2. Themes of the key functionalities of the telerehabilitation technologies.

<table>
<thead>
<tr>
<th>First author, year</th>
<th>Exercise</th>
<th>Monitoring progress</th>
<th>Functions</th>
<th>Communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repository</td>
<td>Diary</td>
<td>Tracker or reminder</td>
<td>Biofeedback</td>
<td>VR</td>
</tr>
<tr>
<td>Alexander, 2023 [67]</td>
<td>✓</td>
<td>✓</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>An, 2021 [110]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Argent, 2019 [169]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Bade, 2020 [166]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Bell, 2020 [90]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>—</td>
</tr>
<tr>
<td>Bini, 2017 [71]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Campbell, 2019 [72]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>—</td>
</tr>
<tr>
<td>Chughtai, 2018 [76]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Chughtai, 2019 [75]</td>
<td>✓</td>
<td>✓</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Doiron-Cadrin, 2019 [77]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>—</td>
</tr>
<tr>
<td>Duong, 2023 [106]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>—</td>
</tr>
</tbody>
</table>

https://mhealth.jmir.org/2024/1/e47843

JMIR Mhealth Uhealth 2024 | vol. 12 | e47843 | p.178
(page number not for citation purposes)
<table>
<thead>
<tr>
<th>First author, year</th>
<th>Exercise</th>
<th>Repository</th>
<th>Diary</th>
<th>Tracker or reminder</th>
<th>Biofeedback</th>
<th>VR</th>
<th>Feedback to patient</th>
<th>Pain</th>
<th>ROM</th>
<th>Knee function</th>
<th>Physical activity</th>
<th>Somatic time</th>
<th>Sleep</th>
<th>Triggers</th>
<th>Goal setting</th>
<th>Communication</th>
<th>Mode</th>
<th>Goal direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eichler, 2019 [34]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>SA, AP</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Exercise</td>
<td>1-way, 2-way Audio, video, text, F2F</td>
<td></td>
</tr>
<tr>
<td>Fung, 2012 [79]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>SA</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>2-way Lower extremity function</td>
<td>Text, F2F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gray, 2022 [37]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SP</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1-way, 2-way Text</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gunduz, 2021 [38]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Hadamus, 2020 [40]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>SA, SP</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Exercise</td>
<td>2-way F2F</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hardwick-Morris, 2022 [107]</td>
<td>✓</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>SP</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SC</td>
<td>2-way Video, text</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hong, 2022 [80]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SP</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Recovery goals</td>
<td>2-way Video</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Huang, 2017 [113]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Juhl, 2016 [44]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>SP</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SC</td>
<td>2-way Unclear</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Klement, 2019 [81]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1-way, 2-way Text, video, F2F</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knapp, 2021 [83]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>NU^k</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

https://mhealth.jmir.org/2024/12/e47843

JMIR Mhealth Uhealth 2024 | vol. 12 | e47843 | p.179

(page number not for citation purposes)
<table>
<thead>
<tr>
<th>First author, year</th>
<th>Exercise</th>
<th>Repository</th>
<th>Diary</th>
<th>Tracker or reminder</th>
<th>Biofeedback</th>
<th>VR</th>
<th>VRa</th>
<th>Feedback to patient</th>
<th>Pain</th>
<th>ROM</th>
<th>Knee function</th>
<th>Physical activity</th>
<th>Solitary time</th>
<th>Sleep</th>
<th>Triggers</th>
<th>Goal setting</th>
<th>Communication</th>
<th>Functions</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kuether, 2019 [85]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SA, SP</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>2-way</td>
<td>F2F, video</td>
<td></td>
</tr>
<tr>
<td>Lam, 2016 [86]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SA, SP</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>ROM, strength</td>
<td>—</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>LeBrun, 2022 [78]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SP</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SC</td>
<td>2-way</td>
<td>Audio, video</td>
<td></td>
</tr>
<tr>
<td>Li, 2023 [115]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SP</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>2-way</td>
<td>Video, text</td>
<td></td>
</tr>
<tr>
<td>Lu, 2021 [117]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SP</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SC</td>
<td>2-way</td>
<td>Video</td>
<td></td>
</tr>
<tr>
<td>McDonnell, 2022 [147]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Milliren, 2022 [88]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Discharge goal</td>
<td>1-way (automatic)</td>
<td>Text, F2F</td>
</tr>
<tr>
<td>Nuevo, 2023 [50]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SA</td>
<td>✓</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>DS, NU</td>
<td>Video</td>
<td>text</td>
<td></td>
</tr>
<tr>
<td>Osterloh, 2023 [51]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SP</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SC</td>
<td>2-way</td>
<td>Video</td>
<td></td>
</tr>
<tr>
<td>Park, 2017 [118]</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SC</td>
<td>1-way, 2-way</td>
<td>Text, audio (tele),</td>
<td></td>
</tr>
<tr>
<td>Park, 2023 [119]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SP</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SC</td>
<td>2-way</td>
<td>Audio calls</td>
<td></td>
</tr>
<tr>
<td>Piqueras, 2013 [54]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SA, AP</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>2-way</td>
<td>Audio (tele)</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>First author, year</td>
<td>Exercise</td>
<td>Repository</td>
<td>Diary</td>
<td>Tracker or reminder</td>
<td>Biofeedback</td>
<td>VR</td>
<td>Feedback to patient</td>
<td>Pain</td>
<td>ROM</td>
<td>Knee function</td>
<td>Physical activity</td>
<td>Solitary time</td>
<td>Sleep</td>
<td>Triggers</td>
<td>Goal setting</td>
<td>Communication Direction</td>
<td>Mode</td>
<td>Functions</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>----------</td>
<td>------------</td>
<td>-------</td>
<td>---------------------</td>
<td>-------------</td>
<td>----</td>
<td>---------------------</td>
<td>------</td>
<td>------</td>
<td>----------------</td>
<td>-------------------</td>
<td>---------------</td>
<td>-------</td>
<td>---------</td>
<td>--------------</td>
<td>--------------------</td>
<td>------</td>
<td>-----------</td>
<td></td>
</tr>
<tr>
<td>Pournajaf, 2022 [65]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
<td>SA</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>VR-based balance board</td>
<td>Exercise</td>
<td>✓</td>
<td>Exercise</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pronk, 2020 [55]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Un</td>
<td>clear</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Exercise</td>
<td>2-way Video</td>
<td>F2F</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prvu Bettger, 2019 [70]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
<td>SA, SP, AP</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>Exercise</td>
<td>1-way Text</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ramkumar, 2019 [92]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>SA</td>
<td>✓</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>DS</td>
<td>Exercise</td>
<td>2-way Video</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russell, 2011 [105]</td>
<td>✓</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>SP</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SC</td>
<td>Un</td>
<td>clear</td>
<td>2-way Text</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scheper, 2019 [56]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>DS</td>
<td>—</td>
<td>1-way Text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Su, 2015 [120]</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
<td>SA</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Exercise</td>
<td>2-way Video</td>
<td>Audio (tele)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summers, 2023 [93]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
<td>—</td>
<td>SA, SP</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>DS</td>
<td>—</td>
<td>2-way Audio (tele)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Szüts, 2016 [170]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Torpil, 2022 [63]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>SC</td>
<td>Occupation related</td>
<td>2-way Video</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tripuranen, 2021 [95]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>AA</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>1-way Task</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>van Dijk-Huisman, 2020 [64]</td>
<td>✓</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
<td>—</td>
<td>SA, AP</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>SC</td>
<td>2-way Video</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wang, 2023 [121]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>AP</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>Task</td>
<td>2-way Text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Education and Enablement

An exercise repository in the form of videos, text, or infographics was one of the main features in the studies (n=53), of which only 20 studies described the list of exercises (Table S5 in Multimedia Appendix 1). Education for patients was part of the rehabilitation program in 17 studies. Table S6 in Multimedia Appendix 1 lists the topic areas covered in the education materials. Regarding exercise, 6 studies reported using an e-diary for maintaining an exercise log, 11 studies reported using reminders to perform exercises, and 13 studies reported using a tracker for exercise adherence (Multimedia Appendix 2). Feedback on the appropriateness of exercise performance was synchronous (biofeedback or virtual reality) from the app (n=19), directly from the health care provider via a video call with the patients (patient performing exercise live, measurement of ROM during video call, transmission of virtual avatar data to health care provider; n=14), or provided via both (n=6; Table 2). Feedback to the patient, which was either in the form of push notifications or a progress summary, was asynchronous from the app using automated programs in 2 studies. Asynchronous feedback from a health care provider in the form of instructions, messages, or an exercise regimen was reported in 13 studies. Feedback via both the app and a health care provider was provided in 3 studies (Table 2). Only 7 studies [6, 51, 75, 115, 123, 128] had an option for peer support for patients.

Measuring Progress

Measurement of patient-reported outcomes such as pain (n=19) was an inbuilt feature in the app. Changes in knee function and activity were monitored directly via wearables or captured using patient-reported outcome measures. These included ROM in 15 studies, knee function in 8 studies, physical activity in 20 studies, sedentary behavior in 5 studies, and sleep in 4 studies. Automatic alerts were provided to the health care provider for any danger signs such as knee pain, wound health, opioid consumption, function, ROM, number of steps, exercise adherence, and any negative response to questions after entering the postoperative follow-up in 9 studies; for non-use of the technology by patients in 4 studies; and for scheduled consultations in 18 studies (Table 2).

Communication

Mobile app–enabled 1-way communication included push messages, notifications, reminders, patients’ replies to inbuilt questions in the app, information sent to the patient by the health care team, and an SMS text messaging bot (n=10). Two-way communication, either via an app or in face-to-face visits, was reported in 41 studies. In addition, 11 studies reported a combination of both 1 and 2-way communication, and 1 study did not provide sufficient information about communication. Electronic communication was delivered in the form of text, audio or video messages, and direct communication (Table 2).

Goal Setting

Goal setting for exercises, activity, pain management, knee function, ROM, muscle strength, rehabilitation, and discharge as part of the rehabilitation program was reported in 23 studies. The goals were set by either the health care provider or the patient (Table 2).

End Users’ Perceptions

Of the 38 studies that reported user perspectives, 2 focused on the perspectives of health care providers, 27 focused on the perspectives of patients and caregivers, and 9 focused on the perspectives of both groups (health care providers and patients and caregivers). The approach for data collection was quantitative (n=23), qualitative (n=9), or mixed methods (n=6). The sample size ranged from 2 to 200 health care providers and from 5 to 2292 patients (Tables S7 and S8 in Multimedia Appendix 1).
Commonly used quantitative questionnaires to assess satisfaction were the System Usability Scale [129] and the net promoter score [130]. To ratify the experience with telerehabilitation, the Telemedicine Perception Questionnaire was used [131]. Acceptability and usability were assessed using the acceptability of information technology questionnaire [132] and the Telemedicine Usability Questionnaire [133]. Some studies used bespoke questionnaires to report user experience and satisfaction [32, 39, 61, 90, 94, 105, 109, 134-146].

Overall, health care providers perceived telerehabilitation and the use of technology such as biosensors as a way of improving efficiency in providing care [146], patient adherence to exercises [39, 136, 146], patient-physician communication [136], and case management [137, 146]. The main factors associated with user satisfaction with e-consultations were reliable technology, good voice or image synchronization, the refresh rate of images, sound quality, and operability of the peripherals [94, 96, 138, 139]. The key factors they perceived would influence use and uptake of technology were decreased workload (rather than increased) [140], reliability of measurements aided by technology [146], ability to measure functional outcomes objectively [141], clearer criteria when choosing appropriate patients to be enrolled in the program [140], self-efficacy in the use of technology [94, 138, 146], and ease of reporting and tracking of patient data [90]. Patients and health care providers felt e-learning modules, push notifications, and appropriate feedback from sensors and virtual reality improved self-management [138, 142-144] (Figure 4).

Figure 4. Perceptions of patients and health care providers about the technology used. ADL: activities of daily living; FAQs: frequently asked questions; KR: knee replacement; mHealth: mobile health; VR: virtual reality.

Patient satisfaction levels were reported when teleconsultation was provided via a computer, smartphone, or tablet [34, 39, 55, 56, 80, 92, 105, 121, 123, 134, 135, 145, 147-149]; telephone [61], videoconferencing [38, 77, 94, 105, 139, 141, 150-152], a web-based system [32, 50, 90, 96, 140, 153], and an mHealth-enabled integrated care model [46, 88, 138]. Patients were satisfied with telemonitoring due to improved access to services, continued support after discharge from hospital, ability for self-management, reduced need for clinic visits, reduction in cost and travel time, ability of health care providers to provide personalized care [32, 61, 94, 121, 136, 138, 140, 141, 145, 153-155], ease of use [34, 50, 55, 56, 92, 105, 135, 138, 147, 148], motivation to perform exercises [134, 135], sense of security with remote monitoring [134, 155], and empathetic communication by a health care provider [121, 135, 136, 145, 152, 155]. The reasons for dissatisfaction were lack of an in-person examination, shorter appointment times, delay in receiving reports (eg, x-ray), and an inability to transfer pictures from one technology to another [140, 145, 149, 153]. Patients provided suggestions for the app functionalities to improve the ease of use such as minimal clicks, an instructional video for app navigation, and restriction of commercial advertisements [149]. Home modifications [149], emotional well-being, information related to activities of daily living in simple text,
dietary advice, frequently asked questions, and use of traditional medicine for postoperative pain management were a few of the suggestions for app content [121].

Patients were generally satisfied with the telerehabilitation program and were ready to recommend it to others [39, 80, 85, 96, 121, 135, 151]. The use of technology for rehabilitation was influenced by computer literacy [141, 150]. However, interruption of virtual physiotherapy sessions due to poor internet issues [139] was not commonly reported (Figure 4).

Discussion

Principal Findings

This scoping review summarized the extent, user perceptions, range, and nature of technologies used to support rehabilitation following knee arthroplasty. All studies reported in this review were from upper and middle-to-upper–income countries, with a steep increase in studies in the last decade. The technologies focused on enabling patients to remember prescribed exercises as well as be able to perform them appropriately by providing synchronous and asynchronous feedback via biosensors or virtual reality. Motivation and support during recovery via technology-enabled 1-way or 2-way communication gave patients access to health care providers. Self-management and monitoring of progress were dependent on active input using e-diaries by patients or passive input through wearables. In the context where these technologies were evaluated, end users were satisfied and found remote monitoring to be acceptable for routine use.

The last decade has seen an exponential increase in the number of arthroplasties worldwide [156]; however, a corresponding increase in technological solutions to facilitate remote monitoring is nonexistent in resource-limited settings such as LMICs where the need for monitoring and a continuum of care may be higher due to lower literacy levels and lack of access to rehabilitation clinics. Research on this topic that can inform clinical practice is nonexistent in the LMIC context. Despite a high penetration of the smartphone market [157] in LMICs, a higher initial investment to develop the technology, especially in the health care sector [158], or a lack of publication of such efforts could be reasons. In LMICs, there is an increasing trend of lower joint replacement procedures [156]. High out-of-pocket expenditures incurred due to home visits by physiotherapists or clinic visits by patients [159] dictate the need for a cost-effective and feasible technology-based strategy to fit the context while using lessons learned from available research.

There is unequivocal evidence that there is a need for physical and psychological support from professionals during the recovery period for pain management, adherence to exercises, and modifications to therapy planning based on one’s progress [3, 160, 161]. The apps were either focused on a single function (such as communication or knowledge transfer) or were multifunctional. They were generally received well by end users; however, the usability and acceptability of these applications or remote monitoring modalities cannot be extrapolated to low health literacy and tech literacy settings. The challenges we expect with using remote monitoring in the LMIC context could be inequitable smartphone access or tech literacy, internet speed, affordability of wearables, the burden to the health system if these needs are provided free of cost, and the need for educational content in multiple languages in countries with a non-native English-speaking, multilingual population such as in India [162].

Implications for Future Research

mHealth interventions have the potential to expand the reach and effectiveness of health support by facilitating behavior change. However, to ensure these “digital behavior health interventions” effectively engage users and are effective, both microengagement (the mHealth interface) and macroengagement (evidence-based behavior change techniques) are essential [163, 164]. However, we found only a handful of studies that reported user involvement during the development stage [28, 32, 51, 58, 64, 65, 75-77, 80, 86, 88, 115, 118, 169]. Studies rarely provided an adequate explanation of the theoretical behavioral framework behind the technology-based interventions [165].

Since the context and technologies are so varied, any new applications that are developed, especially in the LMIC context, should undertake formative research with end users to understand their needs, understand their preferences, and study the local digital regulatory requirements before investing time and effort. Feasibility and pilot testing by a multidisciplinary team should be crucial steps before a full-scale evaluation [69, 166], and embedding end users’ involvement and documenting their experiences at every stage are vital to refining future interventions [164]. Further, the rehabilitation protocols should map the application features with the desired function [167, 168], and this should be confirmed by means of a process evaluation embedded within the clinical evaluation to inform the mechanism of the impact in a real-life setting [147].

Limitations

This review needs to be interpreted in light of the following limitations. This scoping review focused only on technology interventions for post-knee replacement rehabilitation and hence cannot be extrapolated to other orthopedic procedures. We did not include articles for which the full text was not available. Further, incomplete reporting on the features and functions of the technology is possible and may have affected our qualitative summary and conclusion.

We did not perform a consultation phase as per the guidelines [20], and the research question was formulated upon discussion between the researchers of the scoping review team, physiotherapists, and clinicians. We limited our search from 2001 onward; however, since knee arthroplasty and mHealth came into practice in the last 2 decades, this restriction in the search may not have an implication for our review findings.

Conclusion

Several technologies have been identified to promote adherence, increase self-efficacy, enhance self-management, and support remote monitoring. However, all the available technologies have been developed and used in developed countries. The need for remote monitoring is compelling in resource-limited
countries where knee arthroplasty is on the rise. However, irrespective of the context, it is important to involve a multidisciplinary team and include users’ perspectives during the development stage.

**What Was Already Known About the Topic**

Computer and mobile technologies to support rehabilitation following knee arthroplasty are in wide use. Telerehabilitation and remote monitoring are as effective and safe as clinic-based rehabilitation programs. They reduce out-of-pocket expenditure or health cost expenditure by reducing the time to discharge following surgery and the number of clinic visits after discharge.

**What This Study Adds**

This study provides a map of the types of technology and the functionality of mobile and computer-based multifunction applications. We summarized end users’ perceptions and reasons for satisfaction or dissatisfaction with available technology. The findings reflect the lack of research and readily available technologies for LMICs.

**Acknowledgments**

The authors would like to thank distinguished Professor Gordon Guyatt, in the Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada, for providing expert comments. We also thank Dr. Shyamashree Biswas, Research Intern at the George institute, for assisting with the quality check of the extracted data.

This scoping review is a part of a fellowship funded by DBT/Wellcome Trust India Alliance (grant number: IA/CPHI/20/1/505224).

**Authors' Contributions**

ND conceptualized the protocol and conducted the search. ND, SP, PS, and ShP screened and extracted the data. The first draft was written by SP. RM and AK interpreted the study findings and gave significant feedback to the early drafts. RMad provided expert comments and suggestions and edited the later version of manuscript. ND takes responsibility for the data. All authors read and agreed to the final version of the manuscript.

**Conflicts of Interest**

None declared.

**Multimedia Appendix 1**

Supplementary information on search strategy, included and excluded studies.

[DOCX File, 546 KB - mhealth_v12i1e47843_app1.docx]

**Multimedia Appendix 2**

Raw data extraction file for rehabilitation program studies.

[XLSX File (Microsoft Excel File), 39 KB - mhealth_v12i1e47843_app2.xlsx]

**Multimedia Appendix 3**

PRISMA-ScR checklist.

[PDF File (Adobe PDF File), 105 KB - mhealth_v12i1e47843_app3.pdf]

**References**


23. Rayyan. URL: https://www.rayyan.al/ [accessed 2023-12-11]


https://mhealth.jmir.org/2024/1/e47843

JMR Mhealth Uhealth | vol. 12 | e47843 | p.190

(page number not for citation purposes)


112. Huang P, He J, Zhang YM. The mobile application of patient management in education and follow-up for patients following total knee replacement patients using the WeChat app: survey study. JMIR Mhealth Uhealth 2021 Mar 18;9(3):e18763 [FREE Full text] [Medline: 33734097]


130. Net Promoter System. URL: https://www.netpromotersystem.com/about/measuring-your-net-promoter-score/ [accessed 2023-12-12]


156. Hip and knee replacement. OECD iLibrary. URL: http://tinyurl.com/vx339863 [accessed 2023-12-12]


Abbreviations

- CENTRAL: Cochrane Central Register of Controlled Trials
- JBI: Joanna Briggs Institute
- LMIC: low and middle-income countries
- mHealth: mobile health
- PRISMA-ScR: Preferred Reporting Items for Systematic Review and Meta-Analyses Extension for Scoping Reviews
- RCT: randomized controlled trial
- ROM: range of motion

©Sabhya Pritwani, Purnima Shrivistava, Shruti Pandey, Ajit Kumar, Rajesh Malhotra, Ralph Maddison, Niveditha Devasenapathy. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 26.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Abstract

Background: Smart home technology (SHT) can be useful for aging in place or health-related purposes. However, surveillance studies have highlighted ethical issues with SHTs, including user privacy, security, and autonomy.

Objective: As digital technology is most often designed for younger adults, this review summarizes perceptions of SHTs among users aged 50 years and older to explore their understanding of privacy, the purpose of data collection, risks and benefits, and safety.

Methods: Through an integrative review, we explored community-dwelling adults’ (aged 50 years and older) perceptions of SHTs based on research questions under 4 nonmutually exclusive themes: privacy, the purpose of data collection, risk and benefits, and safety. We searched 1860 titles and abstracts from Ovid MEDLINE, Ovid Embase, Cochrane Database of Systematic Reviews, and Cochrane Central Register of Controlled Trials, Scopus, Web of Science Core Collection, and IEEE Xplore or IET Electronic Library, resulting in 15 included studies.

Results: The 15 studies explored user perception of smart speakers, motion sensors, or home monitoring systems. A total of 13 (87%) studies discussed user privacy concerns regarding data collection and access. A total of 4 (27%) studies explored user knowledge of data collection purposes, 7 (47%) studies featured risk-related concerns such as data breaches and third-party misuse alongside benefits such as convenience, and 9 (60%) studies reported user enthusiasm about the potential for home safety.

Conclusions: Due to the growing size of aging populations and advances in technological capabilities, regulators and designers should focus on user concerns by supporting higher levels of agency regarding data collection, use, and disclosure and by bolstering organizational accountability. This way, relevant privacy regulation and SHT design can better support user safety while diminishing potential risks to privacy, security, autonomy, or discriminatory outcomes.

(JMIR Mhealth Uhealth 2024;12:e48526) doi:10.2196/48526

KEYWORDS

smart homes; privacy; surveillance; ambient assisted living; smart speakers; Internet of Things; sensors; sensor; smart home; perception; perceptions; elderly; older adult; older adults; review methods; review methodology; home monitoring; security; safety; ageing; ageing-in-place; integrative review; integrative reviews
Introduction

Smart home technologies (SHTs) typically consist of one or more devices connected through the Internet of Things, which can transmit user data to various stakeholders [1]. Commonly used SHTs include Wi-Fi–enabled cameras, smart speakers with embedded voice assistants, or ambient assisted-living networks of sensors. SHTs are often controllable through smartphones, web platforms, or voice interaction [2]. These networked devices can be useful to the general population for a variety of reasons, but specifically for the aging population, they allow monitoring health status and enable information sharing with health care practitioners, family, or caregivers, potentially alleviating pressure on such networks [3-6]. Until recently, researchers have noted a gap in user perception studies focusing on older adults’ (aged 65 years and older) unique needs, preferences, and ethical factors in SHT adoption or decision-making [4]. Others have noted the need for further research that involves users from older age groups outside of the laboratory [2]. Overall, researchers have an active interest in better understanding user perceptions to remove the barriers to SHT adoption for aging populations.

Related studies have also focused on the pressing ethical implications of SHTs in terms of privacy, autonomy, and security [3,5,7]. Insights from surveillance studies and gerontology literature warn that such systems can limit user autonomy by flagging spontaneous behavior as “abnormal or deviant” in ways that could discourage users from deviating from daily routines where movements are continually monitored [3,5]. Others have noted the potential for exploiting vulnerable SHT users through surveillance capitalism, in which user behavioral data are commodified by commercial actors, resulting in exacerbated power imbalances [7,8]. For instance, some commercial-grade smart devices have the potential to make behavioral data available to advertisers, third parties, and insurance companies in ways that can disproportionately and negatively affect vulnerable individuals and groups [1,7,8]. Moreover, security issues with any internet-enabled technology can lead to unauthorized data access by malicious actors, exacerbating the potential for harm [9,10].

With these insights in mind, the aim of this paper is to explore the potential benefits and drawbacks of SHTs from the perspective of users aged 50 years and older. Despite the abovementioned privacy and security risks, it has been well established that SHT users are often limited in their knowledge of the purpose of SHT data collection [11,12]. On the other hand, SHTs are often seen as safety-enhancing [13]. Moreover, as mentioned in our related larger review paper on SHT users of all ages (Percy Campbell et al, unpublished data, January 2024), user perception studies frequently pertain to younger populations and such technology is more often designed for younger groups [14]. Because of the usual emphasis on younger age groups and technology, our goal is to incorporate the views of older demographics regarding the paradoxical benefits and drawbacks of SHTs. To do so, we collected user perception studies related to 4 nonmutually exclusive themes: privacy, the purpose of data collection, risk and benefits, and safety. To our knowledge, we are the first to compile research findings spanning these 4 categories, leading to unique insights that can inform private sector data protection regulation and SHT design, especially for older adults. We constructed four research questions prior to our literature search. (1) Privacy: What are SHT users’ privacy attitudes? (2) Purpose: What are SHT users’ understandings of the purpose of why and how their data are collected? (3) Risk or benefits: What do users think about the possible benefits and potential risks of harms of SHTs? (4) Safety: What are SHT users’ safety perceptions?

Methods

Overview

This section outlines our search strategy and the inclusion and exclusion criteria for paper selection. Research questions were crafted to examine the interdisciplinary literature on user perceptions of smart home surveillance. We used an integrative review framework to provide an established, rigorous, and comprehensive review method. An integrative approach is well suited to consolidating an expansive range of articles from varied theoretical backgrounds and empirical methods, allowing for a deeper understanding of a given phenomenon [15].

Search Strategy

The search for peer-reviewed English studies was conducted in October 2021. The research team selected relevant keywords based on 4 research questions listed in the previous section. A health information specialist helped to further identify and refine the search keywords (Multimedia Appendix 1) and selected the following databases to find relevant journal articles using the following databases: Ovid MEDLINE, Ovid Embase, Cochrane Database of Systematic Reviews (Ovid), Cochrane Central Register of Controlled Trials (Ovid), Scopus; Web of Science Core Collection, and IEEE Xplore or IET Electronic Library (IEL). No time frame for the publication date was specified. The results were imported into Covidence reference management software to manage the screening process. The duplicate studies were removed automatically by Covidence and manually by team members.

Following duplicate removal, 4 team members were involved in the review process, which included a title or abstract screening round and a full-text review screening round. Articles were eligible for full-text review if they initially appeared to meet inclusion criteria in the title and abstract phase. Next, in the full-text review phase, each article was read in full and subsequently accepted or rejected based on inclusion and exclusion criteria. To ensure reliability and to mitigate subjective biases, article selection in each research phase required acceptance from 2 team members working independently. The rare instance of disagreement between researchers over whether to accept or reject an article was resolved through the involvement of other team members in weekly team meetings. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram (Figure 1) outlines the screening processes.
Inclusion and Exclusion Criteria

Eligible studies were those focused predominantly on smart home users’ perceptions of privacy, data collection purposes, perceived risks or benefits, and physical safety impacts of SHTs. Selected studies included SHT users aged 50 years and older, aside from 1 study where the participant age range spanned from 48 to 84 years with a mean age of 67.8 years and the aging population was the key focus [16]. The studies were included if participants were community dwelling rather than residing in care home facilities (eg, nursing homes), assisted living facilities, or hospitals. These clinical settings produce additional complexities associated with SHTs (eg, legal requirements, staff, and patient or resident consent), which were beyond the scope of this review. Qualitative, quantitative, and mixed methods empirical peer-reviewed studies that measured user perception of smart home surveillance were included. Common methods from accepted studies included questionnaires, surveys, interviews, and focus groups.

The following enumerates our exclusion criteria: (1) studies focused exclusively on wearables or smartphones due to their use outside of the home and further surveillance potential such as location tracking; (2) smart energy grid studies focused solely on cost or energy sustainability; (3) technical articles (eg, algorithm or system development) or theoretical articles; (4) system feasibility studies that were unrelated to user perception; (5) other review papers; (6) usability or acceptability studies that were unrelated to privacy, data collection purpose, risk, or safety.

Thematic Analysis

Once the full-text screening phase had been completed, key details from each accepted study were entered into a shared Microsoft Excel sheet. The sheet was used to record the article title, author, publication year, country, method, demographic breakdown of participants (age or gender), and a short summary of key findings under the corresponding themes of privacy, the purpose of data collection, risk or benefits, and safety. The articles were classified under 1 or more themes when insights could be linked to our related research questions. These findings were then reported throughout the results section, which has been organized by theme. Summarizing articles by theme allows for patterns and contradictions to emerge from the data, ultimately facilitating analysis in the discussion section.
Results

Study Characteristics
Throughout the first phase, 2398 titles and abstracts were screened by our research team. The team selected 146 articles for full-text review, of which 78 were excluded based on the screening criteria mentioned in the previous section. The 68 remaining articles were selected for a larger user perception review paper on users of all age groups. Of those 68 studies, the 15 studies that focused primarily on adults aged 50 years and older were included here. Publication dates ranged from 2011 to October 2021. The results included 6 qualitative, 5 quantitative, and 4 mixed methods studies. Sample sizes ranged from 8 to 447 participants. Studies were conducted in the United States (n=4), the United Kingdom (n=3), Philippines (n=1), South Korea (n=1), Germany (n=1), Finland (n=1), the Netherlands (n=1), and 3 studies sampled participants from India, Thailand, Indonesia, and Malaysia. Participants’ mean ages ranged from 66.4 years to 86.67 years. Further demographic details are available in Table 1.
<table>
<thead>
<tr>
<th>Reference</th>
<th>Location</th>
<th>SHT type</th>
<th>Demographic of participants</th>
<th>Method</th>
<th>Key themes</th>
</tr>
</thead>
</table>
| Albina and Hernandez [17]  | Philippines      | Sensors, cameras                                                         | • N=118  
  • Age range: 60 to ≥81 years  
  • 34.3% female  
  • 65.7% male                                                  | Survey          | Enhanced safety and privacy concerns                           |
| Arthanat et al [18]        | United States    | Thermostats, voice assistants, home security systems, cameras, and remote controlled lights and appliances | • N=447  
  • Age range: 65-95 years  
  • 68.8% female  
  • 31.2% male                                                  | Survey          | Enhanced safety                                              |
| Choi et al [2]             | United States    | Smart speakers, cameras, door or window sensor, and multisensor          | • N=37  
  • Age range: N/A  
  • 75% female  
  • 22% male                                                  | Semi-structured interview | Privacy concerns                                             |
| Chung et al [19]           | United States    | Smart speakers (voice assistant)                                         | • N=47  
  • Age range: N/A  
  • 49% female  
  • 51% male                                                  | Survey          | Privacy concerns and risk or benefits (lack of affordability) |
| Ghorayeb et al [13]        | United Kingdom   | Nonmedical sensors: (SPHERE system: environmental and wearable and video sensors) | • N=13  
  • Group 1 (users): N=6  
  • Age range: 66-88 years  
  • 67% female  
  • 33% male  
  • Group 2 (nonuser): N=7  
  • Age range: 67-89 years  
  • 57% female  
  • 43% male                                                  | Focus groups     | Privacy concerns, safety concerns, and purpose of data collection (unaware or forgetful) |
| Jo et al [20]              | South Korea      | Environmental sensors, Bluetooth smart bands, and receivers              | • N=9  
  • Age range: 68-87 years  
  • 100% female                                                  | Focus groups     | Fewer privacy concerns, purpose of data collection (aware), and enhanced safety |
| Kirchbuchner et al [16]    | Germany          | Sensors                                                                  | • N=60  
  • Age range: 48-84 years  
  • 70% female  
  • 30% male                                                  | Survey          | Privacy concerns, risk or benefits, and enhanced safety       |
| Leikas and Kulju [21]      | Finland          | Sensors                                                                  | • N=8  
  • Age range: 70-85 years  
  • 88% female  
  • 12% male                                                  | Focus groups     | Enhanced safety, risk or benefits (improved independent living), privacy concerns, and purpose of data collection (unaware) |
| O’Brien et al [22]         | United States    | Smart speakers (voice assistant)                                         | • N=125  
  • Age range: N/A  
  • Sex: N/A                                                    | Thematic analysis of Amazon smart speaker reviews | Enhanced safety |
| Pal et al [4]              | India, Thailand, Indonesia, and Malaysia                               | Smart homes                                                             | • N=239  
  • Age range: 55 to 75+ years  
  • 34.3% female  
  • 65.7% male                                                  | Survey          | Privacy concerns and risk or benefits (security concerns)     |
Thematic Results

Overview

Table 1 shows the same papers classified under our research questions related to privacy, the purpose of data collection, risks and benefits, and safety. A total of 13 studies related to user privacy perception, 4 studies explored users’ understandings of the purpose of their data collection, 7 studies related to the benefits and risks of SHTs, while 9 studies pertained to user safety perception.

In summary, our results show that users display a diverse range of perspectives on privacy, the purpose of data collection, risks and benefits, and safety. Although privacy is often seen as an important value in need of protection, nuanced perspectives showed that users were more comfortable with certain types of devices over others, and more comfortable with certain groups of data recipients than others. Participants were sometimes unaware of SHT data collection purposes, although others expressed higher levels of understanding when their SHT functions were adequately explained to them. However, in some cases, the details of SHTs’ purpose and function were forgotten over time. Often, users lacked confidence in explaining what data their devices collected or why. Security risks, including hacking and data breaches, were often cited user concerns, while SHT convenience was frequently seen as a major benefit.

Overall, despite varying levels of concern in the aforementioned areas, users were generally enthused about safety-enhancing benefits of SHTs. These safety benefits may be especially important to older adults with health concerns in particular, as SHTs can act as emergency contact tools or direct lines of support with health care providers, caregivers, and family. In what follows, the key findings are explained in further detail and separated by theme.

Privacy: What Are SHT Users’ Privacy Attitudes?

A total of 13 studies discussed privacy perspectives in general, with some flagging privacy as an important consideration in SHT adoption [2,17]. This section explores a range of user privacy attitudes, where environmental or motion sensors were the most accepted type of SHT [20,26] compared to smart speakers or cameras which were considered invasive [2,16]. Participants indicated the need to control access to their SHT data [13,21]. Many participants were skeptical about the handling of their personal information by service providers and third parties [4,23,24]. However, over time, some users forgot about the presence of sensors that collected electricity, light, temperature, and movement data [26]. In another study, most participants forgot about the presence of nonmedical sensors that collected environmental and activity data in the home [13].

Certain SHTs were generally viewed as nonthreatening, such as door and window sensors, multisensors [2], fall detection and health monitoring sensors [20], or light, temperature, and movement sensors [26]. Some participants considered sensor systems to be preferable in comparison to alternative arrangements such as institutionalization, which was generally seen as undesirable due to a lack of privacy and restricted visitation rights [27]. In early smart home models, networks of motion sensors, fall detectors, emergency voice response, and fire detection sensors were seen preferably by users, except for 1 participant who removed the technology from her home due...
to privacy and autonomy concerns [27]. This participant enjoyed standing in the hallway which would set off alarms, and in turn, alerted staff. However, although most participants did not feel they were being “watched or monitored” the authors also noted that “some are even not fully aware of the presence of the UAS (Unattended Autonomous Surveillance) system at home” [25].

In terms of data sharing, participants expressed mixed attitudes. In a UK study, interviews (n=41) showed older adults were open to having behavioral data collected and shared with family members or health care providers. Here, researchers noted that older adults were more open to sharing data for health care purposes than younger people [25]. One study from South Korea reported that participants (n=9) used environmental sensors for energy management and health-related sensors for fall detection and activity monitoring and reported willingness to share their health-related data with friends, family, and health care practitioners [20]. By contrast, others expressed the need to limit their data sharing to select parties. Those equipped with environmental, wearable, and video sensors in the United Kingdom preferred to share data with health care practitioners instead of family or friends (n=7) [13]. In a survey of 118 older adults (aged 60 years and older) in the Philippines, participants were concerned about assistive technology data access and sharing from environmental sensors and cameras [17].

In Germany, older adults (n=60) perceived cameras to be privacy-invasive compared to other SHTs. Here, privacy was prioritized above other potentially relevant adoption considerations, such as ease of use. Again, study participants preferred to limit data sharing, rejecting commercial service providers as legitimate data recipients [16]. Additionally, in a US study (n=37), the IP web cameras were considered more invasive than other SHTs such as smart speakers, door and window sensors, or multisensors [2]. Some participants expressed concern over smart speaker developers listening in to private conversations, while others were unperturbed [2]. Users were comfortable using smart speakers for certain purposes such as alarms, reminders, and searching for information. However, many were hesitant to use medical SHTs that shared their health data; 1 user specifically noted their discomfort with the potential for pharmaceutical companies to profile them with targeted advertisements based on health-related data [2]. In another US survey with 47 Amazon smart speaker (Alexa) users, some participants expressed concern over their conversations being monitored, while others were indifferent [19]. The following section discusses the extent to which users understood the purpose of their SHT data collection.

**Purpose: What Are SHT Users’ Understandings of the Purpose of Why and How Their Data Are Collected?**

A total of 4 studies revealed insight into what participants understood about the purposes of their SHT data collection. In using SHTs for health care purposes, participants in Jo et al [20] were generally aware of the purposes of their sensor data collection. In this study, researchers had explained to participants what data were collected, how they were stored, and who had access to their data. However, study participants most often relied on support networks such as friends, family, or neighbors to help with their privacy decisions regarding SHTs. One issue arose, however, when participants in another study were taught about the functions of their SHTs. They eventually forgot the purposes of why those sensors were installed and, by extension, what information was being transmitted [13]. Focus group participants expressed a lack of confidence in their knowledge of whether sensors were measuring water consumption levels, humidity levels, body movement, the number of people in the room, and how alarms are triggered by artificial intelligence (AI) [13]. Similarly, focus group participants in Finland lacked confidence in their knowledge of who had access to their movement sensor data, for what purposes, and whether they had access to it themselves [21]. As noted elsewhere, SHT users often have a sense that privacy issues are present, but are unsure of “what data is collected, or how or why” [19]. Overall, apart from 1 study under this category [20], participants expressed limited understanding of the purposes of the collection of their SHT data [13,21,25]. As will be further discussed, a lack of privacy literacy around the types of surveillance SHTs contribute to is an issue with users of all age groups, younger demographics included. A lack of literacy in this area may result in difficulties in obtaining ongoing consent and informed decision-making regarding SHT use.

**Risk and Benefits: What Do Users Think About the Possible Benefits and Potential Risks of Harm of SHTs?**

A total of 7 studies explored user perceptions of SHT risks and benefits. Overall, participants identified data security threats as significant risks [16,19,21,23,24]. These perceived risks were mostly divided between disquiet over malicious data breaches, such as through hacking and misuse of personal data by smart home providers. Malicious data breaches were generally characterized by users as the unauthorized access of data by criminal parties; 1 survey (n=60) found that participants, the majority of whom had no prior experience with SHTs, were mainly concerned with criminal access to their data [16]. Similarly, the larger survey sample (n=237) in the study by Pal et al [23] showed that older adults did not trust smart home companies to securely handle their data or prevent data breaches. Specifically, the authors found that SHT users wanted their personal data to be anonymized and did not trust SHT providers to provide adequate or desired protections [23].

Alongside malicious data breaches, the misuse of personal data by SHT providers was consistently described as a risk by study participants. These concerns were mirrored in commercial contexts: focus group discussions (n=14) showed that SHT users were knowledgeable about the collection of their consumer data and were uncomfortable with their lack of agency in the use of the data [21]. Another study by Pal et al [24] (n=239) further affirmed that SHT users are uncomfortable with corporate access to their personal information. From specific medical and commercial contexts to overarching sentiments, users appear dissatisfied with the current levels of data protection offered by SHTs.

Users described additional risks beyond data collection. These included concerns over SHT dependence: semistructured interviews (n=12) demonstrated that power outages or system
failures were flagged as risks by SHT users [27]. Participants also expressed concern over steep learning curves with new SHTs, as well as a potential lack of agency in selecting their own devices and controlling use of the devices [24]. Additionally, affordability was consistently identified as a risk, with concerns that SHTs would not offer benefits worth their price [4]. This finding was reiterated by Chung et al [19] where 24 of 47 (51%) surveyed users reported that affordability surpassed other risks. Overall, malfunction, affordability, and user trust represent additional risks identified by older adults using SHT.

Alongside risks, participants aged 50 years and older noted distinct positive benefits conferred by SHTs. Users were commonly enthusiastic about assistive smart home devices, including mobility monitoring, voice response, fire detection, and wandering prevention technology. Participants believed these SHTs gave them greater independence and reduced the burden on supportive family members and caregivers [21,27]. Similarly, the survey responses (n=239) in the study by Pal et al [23] indicated that users enjoyed home automation, which increased their daily convenience, especially those users who experienced or expected to experience physical or cognitive ailments. Survey participants simultaneously identified the abovementioned risks while reportedly appreciating SHTs’ value, thus creating the need to trade their reservations for SHT convenience [24]. Finally, users gained self-confidence with digital technology by mastering newly installed SHTs; however, learning to use the devices was sometimes perceived as a barrier and a deterrent to use [19]. These varied findings comprise the social benefits identified by users; however, enhanced physical safety was among the most noted. The following section explores safety perceptions in greater detail.

Safety: What Are SHT Users’ Safety Perceptions?

A total of 9 studies discussed the role of SHTs in safety enhancement, where participants were generally enthused about their devices’ safety features and support for aging in place [16,17,20-22,26,27]. Safety has been viewed as an important component of smart home adoption, especially for older adults with health issues [18]. Sensor users have even expressed the need to trade their privacy for increased safety through SHTs, especially for older adults living alone who experience memory problems [23].

All types of SHTs were considered useful for safety purposes, especially in an accident or emergency. In one example, stroke survivors in the United Kingdom felt safer using motion sensors in the home, as these gave them the feeling that they were being looked after [26]. In South Korea, participants found environmental and wearable sensors to be beneficial for aging in place and reported enhanced feelings of safety [20]. The ability to share their behavioral data was seen as a form of safety assurance among users [20]. Participants using ambient intelligent systems also reported enhanced feelings of safety in the home in the Netherlands, especially in the event of a fall or when feeling unwell and unable to access the phone [27]. Fall detection and other health-related safety features and burglary detection were generally well accepted [16]. Likewise, in the Philippines, assistive technology users reported enthusiasm about increased feelings of safety in the home through emergency response features [17]. Sensor Platform for Healthcare in Residential Environment users were subject to environmental, wearable, and video sensors. They were mainly concerned about the limited ability of human operators to react quickly enough in the event of an emergency [13].

Smart speakers embedded with voice assistants were also perceived to improve safety [22]. In a study examining 125 Amazon smart speaker reviews, safety features were commonly mentioned by older adults and caregivers. For example, emergency contact features such as “Ask My Buddy” were popular among reviewers [22]. In the words of one reviewer [22],

If I call out “Alexa, tell My Buddy to alert contacts,” she sends an alert via cell phone voice and text to my contact list telling them to check on me. This is great in the event of anything from a slip in the shower to any medical or emergency issue or if I feel in danger.

Throughout our collected studies, this sentiment appears to reflect the preference for the safety-enhancing features that SHTs may provide among many adults aged 50 years and older.

Discussion

Paradoxical Nature of SHTs

Throughout the analysis of 15 studies, study participants identified numerous benefits and drawbacks of SHTs. Overall, our findings indicated that SHT users aged 50 years and older found value in SHTs for several reasons beyond the superficial purposes of convenience or entertainment. Perceived benefits included enhanced independence levels for older adults and increased confidence levels with technology [19,21,27]. They were also widely perceived to support health and well-being through fall prevention or emergency contact features and were seen to enhance physical safety levels at home [16,17,20-22,26,27].

Conversely, participants voiced several concerns pertaining to device affordability [4,19], device reliability, criminal data breaches [16], or a lack of trust in SHT companies in securing user data [23]. If user consent over SHT data collection is to be considered meaningful, it should be ongoing, which poses an issue in cases where users expressed limited understanding of data flows and access [21]. As mentioned by homecare field professionals and related employees, this challenge can be particularly difficult for those who develop memory issues in later years [21]. This is complicated by the fact that privacy concerns sometimes fade over time and participants sometimes forget about the existence of their SHTs altogether [13,26]. Low levels of understanding around SHT data collection purposes, use, and disclosure span all age groups [28] but may be especially detrimental to older people who have SHTs installed by others for health and safety purposes.

It follows that the need for higher levels of user autonomy regarding data access is a consistent finding that requires further attention [4,16,21,23-25]. This is especially important because SHTs are often marketed in ways that promote increased autonomy for older adults. However, if SHT settings are not
carefully configured and managed, they may increase independence in some ways while simultaneously diminishing it in others [28]. Data sharing and intrusive monitoring may create issues related to privacy, autonomy, or attempts at behavioral control [29]. Higher levels of user autonomy would require a strong understanding and access to controls over data monitoring and use. Otherwise, these issues can be mitigated by designers embedding tightly controlled “privacy by default” settings. The SHTs should occasionally prompt users to review and manage privacy settings and restrict data flows where unnecessary for device functionality. Finally, some older adult users may consider creating a set of guidelines and privacy preferences for caregivers to follow should memory or cognitive capacities diminish.

Despite a general unease with data sharing among third parties or service providers, one major limitation of our selected studies is the lack of detailed participant discussion on the potential for SHTs to influence insurance rates, targeted ads, or the increased difficulty in differentiating consumer data from health data. Recall that participants rejected commercial providers as data recipients [2,16]. Commercial SHTs such as smart speakers commodify user data [1,30,31], potentially inferring user health data in the process [32] and sharing such information with third parties with unknown goals or incentives. In cases where commercial-grade SHTs are used as care or safety devices for older people with health issues (eg, [19,22]), should such data still be commodified by commercial actors? SHTs can reveal mental and physical health status, mood, personality traits, and sensitive activity recognition, among other personal details [32-34]. In some cases, SHT or wearable data can also be used to influence personalized insurance rates in ways that may be disadvantage older adults with health issues [7].

Moreover, SHT developers in health care spaces have noted the difficulty in differentiating what is or should be considered medical and health data versus what is not [5]. If users are unaware of what types of data they are sharing (eg, [13,21]), to what extent is autonomous decision-making enabled or respected? When commercial-grade SHTs are used to infer health data, they may be treated as consumer data, facilitating access by public and private sector actors outside of user knowledge or meaningful consent. Many people would likely object to commercial actors gaining access to health care data from hospital settings for the purposes of third-party advertising, yet inferring user behavioral and health patterns through SHTs and wearables is possible. The discriminatory issues with targeted advertising, data brokers, and marketer classifications of different groups of people are well known [35]. Currently, studies linking the ways that SHT data contribute to targeted ads through behavioral patterns or biometric markers such as voice are in their infancy [30,33]. Additional research is needed on how SHT data from older adults are treated by SHT companies; what the subsequent targeted advertising or personalized insurance outcomes may be, either now or in the future; and whether such outcomes are discriminatory in nature.

For these reasons, the ethical implications of inferring health-related data from commercial SHT products should be considered alongside the abovementioned user privacy concerns. In short, the challenges in protecting SHT user privacy and autonomy are ongoing [3] and can be further complicated by the involvement of inferred or self-reported health-related data. As has been recommended elsewhere [2,6,14,29,36] SHT developers should prioritize design choices that better support members of all age groups through user-centric design, considering multiple stakeholders, such as older adults, nurses, and caregivers. Others have advocated for an ethical by-design (EbD) approach to implementing digital technology not only through co-design and product development but also through transdisciplinary research [29,37]. Alongside EbD choices, private sector privacy regulation could further protect users through a data justice approach that privileges human rights over commercial interests.

At a global level, the technological ability to collect and aggregate data for surveillance has outsized regulatory mechanisms [38]. Using a data justice framework is a logical path forward to the ethical use of technology in ways that benefit both individuals and groups without further disempowering them through surveillance imperatives that do not suit their needs. Taylor’s [38] data justice framework includes three pillars: (1) visibility, (2) engagement, and (3) antidiscrimination. The first pillar, visibility, refers to the understanding that representation in certain databases can be beneficial to individuals and groups, such as in health care or welfare services. However, it also recognizes the right to privacy and the need to opt out of databases, such as those aggregated by commercial bodies [38]. As shown throughout our findings, many users indicated preferences in sharing their data with health care providers instead of family or friends [13] and preferred not to share with manufacturers, marketers, or other third parties [16]. The second pillar, digital engagement and disengagement, supports individual autonomy by encouraging personalized decisions regarding a user’s preferred level of technological engagement and control over circumstances [38]. For our purposes, the right to digital disengagement would help support older adult SHT users in situations where personalized human care is their preferred option for certain purposes or where only select SHT functions were preferred. The third pillar, the right to challenge data-driven discrimination, allows for the ability to challenge bias in algorithmic decision-making and outcomes [38]. This last pillar may be particularly important as AI capabilities continue to develop alongside rising SHT popularity. Although issues with gender and racial bias with AI platforms are well documented within the literature, digital ageism is currently understudied and is thus in need of further critical analysis [28,36]. The ability to evaluate and challenge ageist bias is an important task as consumer-grade devices become more popular among aging populations. Taken together, regulatory frameworks following Taylor’s [38] 3 pillars of data justice can be used to construct meaningful guidelines around how SHT data should be managed by private sector actors. This way, those who choose to engage with such technology in their homes can enjoy the potential health and safety benefits of SHTs while preventing or mitigating challenges to privacy, autonomy, and discrimination that can be detrimental to older age groups.

Strengths and Limitations
To the best of our knowledge, we are the first to research SHT user perception under the 4 themes of privacy, the purpose of
data collection, risk and benefits, and safety. Previous reviews have largely focused on rehabilitation or health care settings exclusively, whereas we have also incorporated user perception of commercial SHT surveillance. Our review engaged with interdisciplinary fields across the social sciences, computer sciences, engineering, legal studies, and nursing. We have also applied insights from the surveillance studies literature to findings from gerontology research. In terms of limitations, as we excluded studies that focused solely on nonusers, we may have missed potential insight into why individuals do not adopt SHTs. We also excluded studies on smartphones or wearable devices, due to their ability to be used outside the home, which may have further limited our findings. We did not include other search methods such as hand searching for references and did not reconduct the search after October 2021 both of which may have resulted in additional relevant studies. We did not conduct a quality appraisal of our included studies, resulting in another potential limitation. As many of these studies were written about users in global North countries, the extent to which these findings are representative of other regions requires further inquiry. Finally, only English language studies were reviewed, so relevant non-English papers may have been omitted.

Conclusions
In conclusion, through our review of 15 studies, we have demonstrated a variety of perceived benefits and drawbacks from research participants over the age of 50 years. Although SHTs are seen as beneficial for safety enhancement such as emergency contact and convenience purposes, many users are also concerned about the privacy and or security risks, such as a lack of knowledge over where their data were going or a lack of control over who had access. These findings add to the growing body of literature highlighting the need for more age-inclusive technology design. This becomes especially important as commercial-grade SHTs are increasingly positioned to be used for care or health-related purposes for aging populations. In tandem with age-inclusive efforts such as EbD approaches [29], we further encourage the use and development of technology that enhances home safety while respecting the need for user privacy and autonomy. To do so, we have recommended data justice [38] as an equitable approach to these issues through regulatory guidelines.

Future directions for research in this area include studies on how privacy regulators can better support adults aged 50 years and older who use SHT or wearable devices for health or safety purposes. Further work is also needed on how privacy settings can be made more easily accessible and flexible to support everyday users in various contexts. As mentioned, robust analysis is needed where there is a current gap in the literature pertaining to the link between older adults, targeted advertisements or personalized insurance pricing, and SHTs or wearables [28], both in the practical application of such commercial relationships and through user perception studies. For further insight on this topic, subsequent user perception research on SHTs in general should actively include participants over the age of 50 years, especially in the oldest age categories, as opposed to targeting younger populations exclusively. Finally, beyond privacy and security, user perception studies on related ethical issues such as AI discrimination and the potential impacts on user autonomy should be further explored.

Acknowledgments
This work was supported by the Social Sciences and Humanities Research Council (435-2021-0879). The authors thank the research assistant Alicia Natividad for the assistance with article screening and selection. They would also like to extend their thanks to Ani Orchanian-Cheff for assistance with database selection and keyword searches.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File, 13 KB - mhealth_v12i1e48526_app1.docx ]

Multimedia Appendix 2
PRISMA Checklist.
[PDF File (Adobe PDF File), 112 KB - mhealth_v12i1e48526_app2.pdf ]

References


Abbreviations

AI: artificial intelligence
EbD: ethical by-design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SHT: smart home technology

©Jessica Percy Campbell, Jacob Buchan, Charlene H Chu, Andria Bianchi, Jesse Hoey, Shehroz S Khan. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 09.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Mobile Apps for COVID-19 Detection and Diagnosis for Future Pandemic Control: Multidimensional Systematic Review

Mehdi Gheisari¹,², PhD; Mustafa Ghaderzadeh³*, PhD; Huxiong Li¹*, PhD; Tania Taami⁴, PhD; Christian Fernández-Campusano⁵, PhD; Hamidreza Sadeghsalehi⁶, PhD; Aaqif Afzaal Abbasi⁷, PhD

¹Institute of Artificial Intelligence, Shaoxing University, Shaoxing, China
²Department of Cognitive Computing, Institute of Computer Science and Engineering, Saveetha School of Engineering, Chennai, India
³School of Nursing and Health Sciences of Boukan, Urmia University of Medical Sciences, Urmia, Iran
⁴Florida State University, Tallahassee, FL, United States
⁵Departamento de Ingeniería Eléctrica, Facultad de Ingeniería, Universidad de Santiago de Chile, Santiago, Chile
⁶Department of Neuroscience, Faculty of Advanced Technologies in Medicine, Tehran, Iran
⁷Department of Earth and Marine Sciences, University of Palermo, Palermo, Italy

*these authors contributed equally

Corresponding Author:
Mustafa Ghaderzadeh, PhD
School of Nursing and Health Sciences of Boukan
Urmia University of Medical Sciences
Kurdistan Blv Boukan
Urmia, 5951715161
Iran
Phone: 98 9129378390
Email: Mustaf.ghaderzadeh@sbmu.ac.ir

Abstract

Background: In the modern world, mobile apps are essential for human advancement, and pandemic control is no exception. The use of mobile apps and technology for the detection and diagnosis of COVID-19 has been the subject of numerous investigations, although no thorough analysis of COVID-19 pandemic prevention has been conducted using mobile apps, creating a gap.

Objective: With the intention of helping software companies and clinical researchers, this study provides comprehensive information regarding the different fields in which mobile apps were used to diagnose COVID-19 during the pandemic.

Methods: In this systematic review, 535 studies were found after searching 5 major research databases (ScienceDirect, Scopus, PubMed, Web of Science, and IEEE). Of these, only 42 (7.9%) studies concerned with diagnosing and detecting COVID-19 were chosen after applying inclusion and exclusion criteria using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) protocol.

Results: Mobile apps were categorized into 6 areas based on the content of these 42 studies: contact tracing, data gathering, data visualization, artificial intelligence (AI)-based diagnosis, rule- and guideline-based diagnosis, and data transformation. Patients with COVID-19 were identified via mobile apps using a variety of clinical, geographic, demographic, radiological, serological, and laboratory data. Most studies concentrated on using AI methods to identify people who might have COVID-19. Additionally, symptoms, cough sounds, and radiological images were used more frequently compared to other data types. Deep learning techniques, such as convolutional neural networks, performed comparatively better in the processing of health care data than other types of AI techniques, which improved the diagnosis of COVID-19.

Conclusions: Mobile apps could soon play a significant role as a powerful tool for data collection, epidemic health data analysis, and the early identification of suspected cases. These technologies can work with the internet of things, cloud storage, 5th-generation technology, and cloud computing. Processing pipelines can be moved to mobile device processing cores using new deep learning methods, such as lightweight neural networks. In the event of future pandemics, mobile apps will play a critical role in rapid diagnosis using various image data and clinical symptoms. Consequently, the rapid diagnosis of these diseases can improve the management of their effects and obtain excellent results in treating patients.
KEYWORDS
COVID-19; detection; diagnosis; internet of things; cloud computing; mobile applications; mobile app; mobile apps; artificial intelligence: AI; mobile phone; smartphone

Introduction

Following the widespread and rapid outbreak of COVID-19, the disease crossed geographical borders and had a devastating impact on the health, economy, and well-being of the worldwide population. According to the World Health Organization (WHO), high-severity COVID-19 was reported in 16%-21% of patients and almost 3% died. In the case of other variants, however, local statistics in many countries indicated a high mortality rate, with some studies estimating a mortality rate of 4% or higher. Due to the novelty of the disease, ways to deal with it were not known early on; still, researchers considered the screening and rapid diagnosis of patients and their separation from healthy people to be significant steps in fighting the disease [1,2].

Early and low-cost diagnosis of infections in any pandemic is essential for pandemic control. Therefore, if it is possible to diagnose and quarantine infected cases in the earliest phases of an outbreak, the outbreak can be managed in the epidemic phase and will not become a pandemic. In the COVID-19 pandemic, various diagnostic methods have been used, with the polymerase chain reaction (PCR) test being the primary diagnostic tool. Nevertheless, PCR is a time-consuming and costly method; until being diagnosed with this diagnostic test, patients might freely transmit COVID-19 and accelerate its conversion into a pandemic by increasing the transmission rate [3,4].

In the COVID-19 pandemic, a wide range of technologies came to aid in the faster diagnosis and screening of people with infection, many of which proved successful. Meanwhile, mobile phones, as highly abundant tools and an information gateway, helped people update their information and make more accurate decisions. In addition to being a platform for installing essential and valuable apps to detect people with infection, smartphones help track people and examine the keywords used by them for making a diagnosis [5-7].

Numerous technical and review studies have addressed smartphone apps for COVID-19 management. Many of these studies have dealt with various mobile apps used to estimate the prevalence, lessons, opportunities, and challenges of these devices and disease management. Still, none of them has adequately focused on detection and diagnosis. This study is the first to systematically review all the studies that have used smartphone technology to detect and diagnose COVID-19. Previous reviews have examined a limited number of studies. Herein, by covering the maximum number of databases, an effort was made to explore all published papers that used smartphones to diagnose COVID-19. In the area of mobile apps for the diagnosis of COVID-19, a thorough and complete study has not been conducted. Alnazi [8] examined apps related to COVID-19 released on Google Play. The 12 studies reviewed included mobile-based software for contact tracing, awareness building, appointment booking, and online consultation. The author only reviewed free apps, however, and although this study was conducted in 2021, it did not include many scientific or commercial apps [8].

Asadzadeh et al [9] determined the scope of mobile solutions in the COVID-19 pandemic and reviewed 16 mobile app studies on COVID-19–related data processing. The reviewed studies were classified into 4 categories: prevention, diagnosis, treatment, and protection. Despite noting a different range of mobile apps, this study did not mention the details and algorithms of these studies [9].

Aslani et al [10] studied mobile health apps in pandemics and epidemics. They examined 17 studies and explored common respiratory diseases and lung infections. Although this study was published during the COVID-19 pandemic, it did not mention mobile apps related to COVID-19 [10].

Kondylakis et al [11] examined 12 studies on mobile apps for COVID-19 data analysis in a more comprehensive investigation. These 12 studies covered the following domains: training, information sharing, risk assessment, self-management of symptoms, contact tracing, home monitoring, and decision-making. Still, this study did not include studies using machine learning (ML) methods to predict and diagnose COVID-19. It also did not deal with mobile apps for COVID-19 diagnosis and primarily focused on studies on COVID-19 education, care, and management using mobile apps [11].

Almalki et al [12] analyzed and discussed all the apps available on Google Play and the Apple Store and provided a brief explanation. Among its flaws, this study did not review scientific or academic studies and only examined mobile apps available in the market, many of which were developed without scientific or clinical supervision. Therefore, it is difficult to rely on these apps, as they lack scientific support and cannot be introduced to or proposed by the communities [12].

Table 1 lists some studies that have addressed mobile apps for COVID-19 data management, including the first author’s name, the country, the main topic, and the number of studies covered.
This study aimed to fill the gap left by previous reviews by conducting a comprehensive review of studies on smartphone apps for the diagnosis of COVID-19, providing solutions based on technological models, and answering research questions so that researchers and health systems can envision devices and their apps in preventing future pandemics.

**Methods**

**Search Criteria**

This systematic review, which was conducted for the first time using this method, aimed to identify relevant studies related to detecting and diagnosing COVID-19 using a variety of smartphone apps. The systematic search strategy was developed based on previous studies and the authors' knowledge. The main objective was to address the following analytical questions (AQs):

- AQ1: What are the uses of smartphones for COVID-19 detection and diagnosis?
- AQ2: What data do smartphones use to detect and diagnose COVID-19?
- AQ3: Which artificial intelligence (AI) methods and algorithms are used to process smartphone data?
- AQ4: How successful have smartphone apps been in COVID-19 detection and classification?
- AQ5: What suggestions can be made to improve the quality of mobile apps in disease diagnosis and pandemic control?

We reviewed electronic databases publishing papers on medicine and computer science. We concluded that PubMed, Web of Science (WoS), Scopus, IEEE, and ScienceDirect contain the most relevant papers. The search used the following keywords and logical expressions: ("COVID-19") AND (Detection OR Diagnosis) AND (Smartphone OR Mobile Application OR Mobile App)). The investigation was conducted from November 1, 2019, to late April 2022, and relevant published papers were extracted. The Embase database was eliminated from the examination due to the proximity of the publications.

**Data Extraction**

Relevant studies and the main elements of their methodology and results were recorded in data extraction forms in order to identify AI algorithms and techniques. Two researchers (authors AMR and MG) performed data extraction, and discrepancies between the researchers were resolved by discussion with an independent researcher (author AH). The extracted data elements included the first author’s name, country of origin, research population, data used, purpose, method, the role of the mobile app, and the evaluation method. The search in reputable databases was performed based on the search strategy, and 535 papers were extracted. After reviewing the papers’ abstracts and full texts, applying the inclusion and exclusion criteria, and selecting papers relevant to the title of this study, 42 (7.9%) full-text papers were finally selected. This process was performed based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart, as shown in Figure 1.

---

**Table 1.** Related studies on mobile apps for detecting and diagnosing COVID-19.

<table>
<thead>
<tr>
<th>Author and country</th>
<th>Main topic</th>
<th>Studies covered, n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanzi [8], Saudi Arabia</td>
<td>Mobile app used during COVID-19</td>
<td>12</td>
</tr>
<tr>
<td>Asadzadeh et al [9], Iran</td>
<td>Mobile health solutions</td>
<td>16</td>
</tr>
<tr>
<td>Aslani et al [10], Iran</td>
<td>Mobile health apps for epidemic and pandemic outbreaks</td>
<td>17</td>
</tr>
<tr>
<td>Kondylakis et al [11], Greece</td>
<td>Mobile app for COVID-19</td>
<td>12</td>
</tr>
<tr>
<td>Almalki et al [12], Saudi Arabia</td>
<td>Implemented an app to combat COVID-19</td>
<td>115</td>
</tr>
</tbody>
</table>
Results

Study Details
Due to the newly emergent status of COVID-19, the titles, abstracts, and keywords of all the papers published between 2020 and 2022 were reviewed, and 42 (7.9%) of 535 papers were ultimately selected as eligible. By comprehensively examining the mobile apps, we found that the role of smartphones was described in 6 areas with different types of COVID-19 data sets, including “Smartphones play the role of a platform for data collection,” “visualizing the input data,” “installing AI-based processing software,” “determining contact tracing,” and “COVID-19 data processing based on role-based and guideline-based methods,” to detect and diagnose COVID-19. A taxonomy was developed to better organize the content and concepts related to mobile apps for COVID-19 diagnosis and detection, as shown in Figure 2.

Figure 1. The review process and how to exclude papers according to the PRISMA flow diagram. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
The taxonomy presented in this research classified the studies conducted in line with the investigation into 2 branches: detection and diagnosis. In the field of detection, the AI method was used more, which included traditional ML algorithms, such as supervised, unsupervised, semisupervised, and reinforcement learning; however, deep learning (DL) dealt with images and sounds. In the field of diagnosis, rule- and guideline-based techniques were mainly used to diagnose COVID-19 with mobile apps.

**Detection vs Diagnosis of COVID-19 Using Mobile Apps**

The terms “detection” and “diagnosis” were used interchangeably in many of the studies, but they differed in meaning and usage. These 2 terms are used differently in different settings. In clinical settings and diagnostic laboratories, the term “diagnosis” is used, while in computer vision and pattern recognition, observing the first definite signs in determining the status of a disease is called “detection.” By examining the differences between these 2 terms in the dictionary and by inquiring from clinicians, a 2021 study found that detection identifies diseases in a set of patient and nonpatient cases. In detection, the disease is distinguished from other conditions, such that other cases may or may not have a disease-free status. However, the precise level and type of disease are completely specified during diagnosis. In the diagnosis concept, different cases might have a disease status or belong to other classes of abnormality or nonhealth, which can be determined [4,13-17].

Accordingly, detection was used for studies that distinguished cases of COVID-19 from healthy and normal cases. In contrast, diagnosis was used for studies that distinguished COVID-19 from other infectious pulmonary diseases (eg, different types of pneumonia). Detection makes sense in sets where other conditions (not infected with COVID-19) are specified, and COVID-19 can be distinguished with certainty from types of pneumonia or other coronaviruses [4,13,14]. By reviewing the 42 extracted papers, we found that 25 (59.5%) papers used DL to detect (identify) COVID-19, whereas 12 (28.6%) studies used it to diagnose and distinguish COVID-19 from other diseases. In addition, 3 (7.1%) studies did not precisely mention that diagnosis was their objective, while in 2 (4.8%) studies, the goal was to diagnose and detect. Figure 3 shows the amount of research performed to detect and diagnose COVID-19.
Overview of the Role of Smartphone Apps in the Diagnosis of COVID-19

The apps were classified into the following 6 categories by reviewing all the studies conducted on mobile technologies and apps in line with our research objectives: contact tracing, data gathering, data visualization, AI-based diagnosis, rule- and guideline-based diagnosis, and data transformation.

Mobile Apps for Contact-Tracing Analysis

One of the main uses of apps during the COVID-19 pandemic was for contact tracing to diagnose and classify COVID-19. Thus, smartphones were used to diagnose patients with COVID-19 who had been in contact with people with infection. Many of the studies used mobile apps to trace people via trackers, geofencing, and GPS.

Mobile Apps for Gathering Data From Users

The sole purpose of some mobile apps was to collect data for processing. These apps received data through a standard smartphone from mobile devices as input devices in perception layers and sent them to the edge and cloud layers, where processing took place. In this case, the mobile apps sent data from the output of biosensors to higher-level systems for processing.

Mobile Apps for Data Visualization

Mobile apps were used to visualize the data received from biosensors or other data collection tools in many of the studies. After data were received, they were displayed to users via statistical charts. Smartphones carried out this task as the first step in data processing, which means gathering data and visualizing these data for COVID-19 diagnosis. In the next stage, these data were transferred to higher levels of processing, such as cloud and fog spaces.

Mobile Apps Used to Analyze COVID-19 Data Based on AI Algorithms

One of the most frequently used mobile apps for COVID-19 diagnosis was an AI-based application in which learning occurred. Studies used different clinical data, sounds, and radiology images. Data were collected in 2 ways: by mobile app designers and developers and by clinical app users using mobile apps.

Since most mobile apps processed and predicted COVID-19 infection, they incorporated different AI algorithms that involved 2 types of methods: methods based on traditional ML algorithms and those based on DL algorithms.

Mobile Apps Used to Analyze COVID-19 Data Based on Guidelines and Rules

A notable study on COVID-19 diagnosis using mobile phones involved ontologies, clinical guidelines, and rules. Due to COVID-19 diagnostic guidelines in health care centers, guideline- and rule-based methods for COVID-19 diagnosis are expected to be popular among clinicians and physicians. By incorporating these guidelines and rules into mobile apps, clinicians and stakeholders can see the process as more tangible and acceptable.

Mobile Apps as a Platform to Transform Messages and Data

Another application of mobile phones for COVID-19 diagnosis was processing keywords related to COVID-19 in social media using smartphones or analog messaging methods. In this method of COVID-19 diagnosis, the main terms representing COVID-19 diagnosis were counted and the COVID-19 diagnosis was made based on the number of uses of these words and their relationship with pronouns and sentence components.
DataUsedinMobileAppsforCOVID-19DetectionandDiagnosis

A wide range of data were used in mobile apps, and according to specialty domains, different scientific disciplines made the initial diagnosis of COVID-19 differently. Mobile apps used a range of data for collection, classification, analysis, transfer, and visualization. These apps dealt with various data, including nasal swap samples, personal clinical data, signs and symptoms, voice and sounds, radiology images, and words and terms in media.

Several of the studies used a combination of data types. Some others used subjective clinical data and symptoms. Moreover, some apps used a variety of radiology images and patients’ voices (cough).

Table 2 lists 11 (26.2%) studies on mobile apps for diagnosing COVID-19. This table includes the study, data, the application method, the mobile role, and the evaluation output.

Table 2. Studies (n=11) evaluating mobile apps used for COVID-19 diagnosis (or a goal similar to diagnosis).

<table>
<thead>
<tr>
<th>Study</th>
<th>Data</th>
<th>Goal</th>
<th>Method</th>
<th>Mobile role</th>
<th>Evaluation output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bindra et al [18]</td>
<td>Symptoms, clinical and bibliography data</td>
<td>COVID-19 risk prediction</td>
<td>ML&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Platform for the applied model to calculate risk prediction</td>
<td>N/A&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Sharma et al [19]</td>
<td>Sound</td>
<td>Diagnosis</td>
<td>ML</td>
<td>Platform for analysis; speech and sound analysis</td>
<td>N/A</td>
</tr>
<tr>
<td>Nema et al [20]</td>
<td>Symptoms</td>
<td>Diagnosis</td>
<td>Rule-based reasoning</td>
<td>Gathering symptoms and receiving alerts (SMS)</td>
<td>N/A</td>
</tr>
<tr>
<td>Quer et al [21]</td>
<td>Smartwatch, activity tracker data, symptoms, testing results</td>
<td>Differentiating COVID-19–positive status</td>
<td>Single decision threshold</td>
<td>Data collection</td>
<td>Accuracy=83.3%</td>
</tr>
<tr>
<td>Elagan et al [22]</td>
<td>Heart rate, blood cell counts, temperature</td>
<td>Diagnosis</td>
<td>Sending patient data to a physician and receiving output from the physician</td>
<td>Estimating the heart rate, receiving data from wireless sensors used to measure white blood cells (WBCs) and red blood cells (RBCs), and estimating air temperature</td>
<td>N/A</td>
</tr>
<tr>
<td>Imran et al [23]</td>
<td>Cough sounds</td>
<td>Diagnosis</td>
<td>DL&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Receiving cough sounds and analyzing them using a designed app</td>
<td>N/A</td>
</tr>
<tr>
<td>Mukhtar et al [24]</td>
<td>Cough, SpO&lt;sub&gt;2&lt;/sub&gt;, temperature</td>
<td>Diagnosis</td>
<td>Rule-based reasoning</td>
<td>Collecting data, sending data, showing the assessment</td>
<td>N/A</td>
</tr>
<tr>
<td>Koshiti et al [25]</td>
<td>Symptoms</td>
<td>Diagnosis</td>
<td>ML</td>
<td>App platform</td>
<td>Accuracy=99%</td>
</tr>
<tr>
<td>Ertuğrul et al [26]</td>
<td>Personal data, observed symptoms (images, sounds)</td>
<td>Prediction</td>
<td>Ontology and rules</td>
<td>App platform</td>
<td>N/A</td>
</tr>
<tr>
<td>Maghded et al [27]</td>
<td>CT&lt;sup&gt;e&lt;/sup&gt; scan, cough, voice, breath sounds, fatigue</td>
<td>Diagnosis</td>
<td>DL</td>
<td>App platform</td>
<td>N/A</td>
</tr>
<tr>
<td>Rangarajan and Ramachandran [28]</td>
<td>CT images</td>
<td>Diagnosis</td>
<td>DL</td>
<td>App platform</td>
<td>N/A</td>
</tr>
</tbody>
</table>

<sup>a</sup>ML: machine learning.
<sup>b</sup>N/A: not applicable.
<sup>c</sup>DL: deep learning.
<sup>d</sup>SpO<sub>2</sub>: saturation of peripheral oxygen.
<sup>e</sup>CT: computed tomography.

There are several mobile apps available to detect COVID-19, as shown in Table 3. An overview of 31 (73.8%) studies is provided in the table, which includes the study, data, the application method, the mobile role, and evaluation results.
<table>
<thead>
<tr>
<th>Study</th>
<th>Data</th>
<th>Goal</th>
<th>Method</th>
<th>Mobile role</th>
<th>Evaluation output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Göökçen et al [29]</td>
<td>Cough</td>
<td>Detection</td>
<td>DL(^a)</td>
<td>Platform for applying a COVID-19 detector via cough sounds</td>
<td>Accuracy=79%; (F_1)-score=80</td>
</tr>
<tr>
<td>Mao et al [30]</td>
<td>Wastewater sample</td>
<td>Detection</td>
<td>Biosensor analysis</td>
<td>Interface to send data</td>
<td>N/A(^b)</td>
</tr>
<tr>
<td>Stasak et al [31]</td>
<td>Speech voice</td>
<td>Detection</td>
<td>ML(^c)</td>
<td>App platform</td>
<td>Accuracy&gt;82%–86%</td>
</tr>
<tr>
<td>Alkhodari and Khandoker [32]</td>
<td>Breath, cough, voice</td>
<td>Detection</td>
<td>DL</td>
<td>App platform</td>
<td>Accuracy=94.5% and 92.1%</td>
</tr>
<tr>
<td>Al-zubidi et al [33]</td>
<td>Blood index</td>
<td>Detection</td>
<td>ML</td>
<td>App platform</td>
<td>Accuracy=89%</td>
</tr>
<tr>
<td>Abdulrazzaq Al-shekhly et al [34]</td>
<td>Thermal images and location</td>
<td>Detection</td>
<td>Thermometer and Send location</td>
<td>API(^d) to calculate and send data</td>
<td>N/A</td>
</tr>
<tr>
<td>Berquedich et al [35]</td>
<td>Contact tracing</td>
<td>Detection and Management</td>
<td>Guideline based</td>
<td>App platform to prescribe drugs and send an alarm</td>
<td>N/A</td>
</tr>
<tr>
<td>Karataş et al [36]</td>
<td>Cough, voice</td>
<td>Detection</td>
<td>ML</td>
<td>App platform</td>
<td>Accuracy=96.5%</td>
</tr>
<tr>
<td>Awasthi et al [37]</td>
<td>Ultrasound images</td>
<td>Detection</td>
<td>DL</td>
<td>App platform</td>
<td>Accuracy=83%</td>
</tr>
<tr>
<td>Tawфик et al [38]</td>
<td>Cough sounds</td>
<td>Detection</td>
<td>ML and DL</td>
<td>App platform</td>
<td>Accuracy=98%</td>
</tr>
<tr>
<td>Kirsanik et al [39]</td>
<td>Symptoms</td>
<td>Detection</td>
<td>Rule-based reasoning</td>
<td>App platform</td>
<td>N/A</td>
</tr>
<tr>
<td>Ponomarchuk et al [40]</td>
<td>Breath and cough sounds</td>
<td>Detection</td>
<td>DL</td>
<td>App platform</td>
<td>N/A</td>
</tr>
<tr>
<td>Shreyas et al [41]</td>
<td>X-ray images</td>
<td>Detection</td>
<td>DL</td>
<td>App platform</td>
<td>Accuracy=98.4%</td>
</tr>
<tr>
<td>Mohsin et al [42]</td>
<td>Symptoms</td>
<td>Detection</td>
<td>Rule-based reasoning</td>
<td>App platform</td>
<td>N/A</td>
</tr>
<tr>
<td>Sanjeev et al [43]</td>
<td>Cough and clinical data ((\text{SpO}_2) level, body temperature, heart rate, symptoms)</td>
<td>Detection</td>
<td>ML</td>
<td>App platform</td>
<td>Accuracy=85%</td>
</tr>
<tr>
<td>Ponomarchuk [40]</td>
<td>Breath and cough sounds</td>
<td>Detection</td>
<td>DL</td>
<td>App platform</td>
<td>N/A</td>
</tr>
<tr>
<td>Bushra et al [44]</td>
<td>X-ray images</td>
<td>Detection</td>
<td>DL</td>
<td>Platform to trace and analyze keywords</td>
<td>Accuracy=98.6%</td>
</tr>
<tr>
<td>Verde et al [45]</td>
<td>Cough</td>
<td>Detection</td>
<td>ML</td>
<td>Platform to analyze cough sounds</td>
<td>Accuracy=82%</td>
</tr>
<tr>
<td>Stanciu et al [46]</td>
<td>Bluetooth data</td>
<td>Virus Detection</td>
<td>Contact management</td>
<td>Contact tracing</td>
<td>N/A</td>
</tr>
<tr>
<td>Han et al [47]</td>
<td>Nasal swab sample</td>
<td>Virus Detection</td>
<td>Fluorescent aptasensors</td>
<td>Data visualization</td>
<td>N/A</td>
</tr>
<tr>
<td>Fozouni et al [48]</td>
<td>Nasal swab sample</td>
<td>Detection</td>
<td>RNA analysis</td>
<td>Data visualization</td>
<td>Sensitivity in less than 30 minutes</td>
</tr>
<tr>
<td>Coppock et al [49]</td>
<td>Audio and sound</td>
<td>Detection</td>
<td>ML</td>
<td>Platform to install an app</td>
<td>N/A</td>
</tr>
<tr>
<td>Wong et al [50]</td>
<td>Symptoms</td>
<td>Detection</td>
<td>Medical protocol</td>
<td>Receiving data and analysis using a designed app</td>
<td>N/A</td>
</tr>
<tr>
<td>Hijazi et al [51]</td>
<td>Heart rate, feeling features, blood pressure</td>
<td>Detection</td>
<td>ML</td>
<td>Collecting data from users</td>
<td>Mean accuracy=83.3% (SD 1.6%)</td>
</tr>
<tr>
<td>Echeverría et al [52]</td>
<td>Sounds, symptoms</td>
<td>Early detection, management of close contacts</td>
<td>Guidelines</td>
<td>Gathering signs and symptoms</td>
<td>N/A</td>
</tr>
<tr>
<td>Verma et al [53]</td>
<td>CT(^f) scan</td>
<td>Detection</td>
<td>DL</td>
<td>Process unit and platform for applied model</td>
<td>Accuracy=99.6%; (F_1)-score=99.6</td>
</tr>
<tr>
<td>Chen et al [54]</td>
<td>Spike protein, nucleocapsid protein</td>
<td>Detection</td>
<td>Data transfer</td>
<td>Receiving, gathering, and transmitting data in edge layers</td>
<td>N/A</td>
</tr>
</tbody>
</table>

\(^a\) DL: Deep Learning
\(^b\) N/A: Not Available
\(^c\) ML: Machine Learning
\(^d\) API: Application Programming Interface
\(^e\) CT: Computed Tomography
\(^f\) SD: Standard Deviation

<table>
<thead>
<tr>
<th>Study</th>
<th>Data</th>
<th>Goal</th>
<th>Method</th>
<th>Mobile role</th>
<th>Evaluation output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen et al [55]</td>
<td>Temperature, heartbeat</td>
<td>Detection</td>
<td>Comparing received data with normal data</td>
<td>Receiving data from a sensor and sending it to the database</td>
<td>N/A</td>
</tr>
<tr>
<td>Wang et al [56]</td>
<td>Keywords on social media</td>
<td>Detecting the SARS-CoV-2 outbreak</td>
<td>Analysis of emerged keywords</td>
<td>Platform to install WeChat to trace and analyze keywords</td>
<td>N/A</td>
</tr>
<tr>
<td>Udhaya Sankar et al [57]</td>
<td>Speech, voice</td>
<td>Detection</td>
<td>Computational audit techniques</td>
<td>Data collection</td>
<td>N/A</td>
</tr>
<tr>
<td>Sun et al [58]</td>
<td>Horse nasal swab samples</td>
<td>Detection</td>
<td>ML</td>
<td>Smartphone-based collection and visualization of data</td>
<td>Results achieved in approx. 30 minutes</td>
</tr>
</tbody>
</table>

aDL: deep learning. 
bN/A: not applicable. 
cML: machine learning. 
dAPI: application programming interface. 
eSpO₂: saturation of peripheral oxygen. 
fCT: computed tomography.

Having thoroughly scrutinized the particulars of the investigations delineated in Tables 2 and 3, the methodologies implemented therein, the scholarly community involved, and the significance of mobile apps in the detection of COVID-19, we were equipped to provide answers to the AQs.

To answer AQ1, although different apps were used, the main apps belonged to 6 functional domains: contact tracing, data gathering, data visualization, AI-based platform to analyze data and signals, rule- and guideline-based methods for decision-making, and data transformation. Contact tracing is a critical application and a strength of this technology for the timely diagnosis of COVID-19 due to its high availability, serving as a module for accessing GPS satellites, determining people’s positions, and accessing higher data transaction layers, such as cloud and fog spaces. Berquedich et al [35] adopted a model for contact tracing and designed technology to reduce hospital visits and alert people who were in contact with patients to seek health care. Some domains, such as the processing of patients’ voice and audio data for COVID-19 detection, received more attention, whereas biosensor data visualization apps were less frequently designed and presented. Figure 4 depicts the 6 functional dimensions of smartphones for COVID-19 diagnosis.

**Figure 4.** Areas of application of mobile technology for the diagnosis of COVID-19. AI: artificial intelligence.

As can be seen in Figure 4, AI-based methods formed a large part (about 54.7% of the methodologies at the core of smartphones to diagnose COVID-19). It is believed that the researchers’ focus was on mobile phones as on-site processing tools for faster detection of COVID-19. However, the mobile apps that embedded guideline- and rule-based techniques in their processing core to analyze biomarkers are powerful tools for processing all kinds of numerical data and present them to users as a piece of point-of-care equipment. These aspects of mobile apps have been of great importance for researchers due to the ease of faster analysis of input data and the data input gates of mobile phones for more immediate identification of patients. More than 14% of the mobile apps analyzed all kinds of biomarkers using guidelines and rules to access a model in the field of COVID-19 diagnosis by classifying input data. A large number of mobile apps also focused on visualizing (11.9%) and transferring (9.5%) data.

One of the strengths of mobile technology in identifying and diagnosing diseases, especially infectious diseases, is the visualization of the data received from internet of things (IoT) technologies and biosensors. After analyzing the data with...
hardware chips and equipment or nucleic acid analysis, the signals are transferred to the mobile app for visualization through various means, such as IoT and Bluetooth. In these studies, blood, serology, and saliva data analysis results, after DNA and amino acid analysis, were sent by biosensors to mobile apps for visualization [27,28,47,52].

In some other studies, one of the most influential features of smartphone technology that had a significant impact on COVID-19 diagnosis and detection was the receiving and collection of data from people suspected of having COVID-19. These data included saliva samples, clinical signs and symptoms, vital signs, demographic information, cough sounds, and medical history. The data were obtained with mobile phones, the equipment connected to and embedded in them (eg, chemical sensors), and the recording of vital and clinical signs was stored in the mobile phones or moved to the cloud, allowing other computing devices to access this information. Some studies dealt with using smartphones for collecting patient data; for computation, they provided the data in a centralized space to specialists or decision support systems [21,22,25,26,32,53].

We observed many features to answer AQ2. The design and development of mobile apps for COVID-19 detection and diagnosis were commensurate with the frequency of COVID-19 diagnosis data. In other words, for frequent data, such as coughs, apps that used this type of data to diagnose COVID-19 were the most frequent. Figure 5 presents the usage of data types in mobile apps.

Figure 5. Data used in the detection and diagnosis of COVID-19.

Several laboratory biomarkers, diagnostic tests, biographical data, histories of diseases, and the voices and cough sounds of patients were used as input features for mobile apps based on methodology. According to Figure 5, it seems that the amount

![](image-url)
of research using these features has a linear relationship with the diagnostic methods of COVID-19, as most mobile apps used the features of signs and symptoms as well as the patients’ voices and cough sounds. In many apps designed to increase the accuracy of disease diagnosis, a combination of these features was used. Since mobile devices can receive and save sounds, these apps used this feature significantly.

For the analysis of nasal swab samples, apps that used biosensor technology incorporated multidisciplinary knowledge and used additional hardware. Following the sample analysis, the resulting data were transferred to the mobile apps via Wi-Fi or Bluetooth and were then used to automatically diagnose COVID-19. As a notable study, one can mention aptamers as a robust molecular tool for COVID-19 diagnosis [12]. The findings revealed that mobile technologies in pandemic control and prevention were a hot research topic and an exciting and trendy approach.

To answer AQ3, the notable use of smartphone apps in COVID-19 prevention assisted clinicians with the timely diagnosis of COVID-19. AI methods, such as ML and DL algorithms, achieved remarkable results in COVID-19 diagnosis. Many researchers [8,10,13,27,35] who used clinical data, such as serology data, vital signs, and symptoms, adopted ML algorithms. These features were received from users and applied to the ML model to diagnose COVID-19. Different ML algorithms were used for this purpose.

Since the main symptoms of COVID-19 diagnosis are cough sounds, the type of coughs, and respiratory sounds, many researchers adopted ML methods and algorithms, such as support vector machines, to identify patients’ sound patterns [10,20,40,48]. Sound feature extraction was carried out based on mathematical algorithms. Using mobile apps to classify these features led to the automatic diagnosis of COVID-19.

Due to the complexity of detecting the pattern of cough sounds, several of the studies required more efficient and voluminous methods, so they used DL algorithms. After training with more significant data and obtaining a more efficient model on systems with more powerful processing units, such as graphics processing units (GPUs), these algorithms were applied to smartphone operating systems with the cough sounds of people suspected of having COVID-19. The methods in this category differentiated infected from noninfected cases by receiving patients’ sounds with higher accuracy and precision [9,30]. Tawfik et al [38] used ML and DL methods simultaneously and achieved 98% accuracy.

Radiographic images offered another facility for rapid diagnosis of COVID-19 and, thus, pandemic control. DL methods and the convolutional neural network (CNN) algorithm were also used in studies using these images in mobile technologies. These models were optimally designed by training on radiology images using concepts such as transfer learning. At the point of care (PoC), they helped clinicians diagnose and detect new cases. These mobile apps demonstrated optimal performance and achieved an accuracy of >92%. By automatically extracting the features using manual feature selection, the bias in the results of diagnostic models was eliminated [18,29,45,46].

AI algorithms were used in various techniques in mobile phones, and the developers attempted to enhance their efficiency by adjusting the parameters of these algorithms [59-62]. Some of the studies used no performance evaluation index to determine the success rate of these apps (to answer AQ4). Regarding apps that used AI methods, the studies used the metrics in the confusion matrix. For different AI methods, different levels of accuracy were obtained. Upon evaluating the efficacy of ML and DL approaches in the identification and diagnosis of COVID-19, our analysis revealed that the accuracy of ML methods ranged from 83% to 99.5%, while DL algorithms exhibited accuracy rates within the range of 91% to 99.6%.

Studies that used all kinds of numerical variables as features, such as signs and symptoms, selected ML methods, while studies that used features such as sounds and images as recognition features, selected DL methods due to their high efficiency.

By calculating the approximate average accuracy of AI-based models in the detection and diagnosis of COVID-19, we concluded that ML algorithms used in mining COVID-19–related data achieved good results. In addition, in analyzing sound data, radiology, or their combination, DL methods achieved high accuracy; in many cases, this accuracy was reported to be >95%. As shown by the accuracy rates presented above, the findings demonstrate the acceptable results of AI methods. In addition, studies using apps to visualize RNA and DNA analysis tests detected cases of COVID-19 in a shorter time due to their sensitivity.

Discussion

Principal Findings

In this study, after reviewing all the research conducted on mobile apps for the diagnosis of COVID-19, several research questions that may have arisen for researchers and health care app development companies were answered. This section provides an answer to the remaining AQ by analyzing and evaluating the data entrance, technological method, and performance metrics of the methodologies outlined in the main tables.

To address AQ5, we recommend using mobile apps for gathering data from PoCs obtained from individuals, monitoring devices, or clinical data generators. In modern techniques, edge layer devices are used in the health care devices layer. Therefore, it is strongly advised to incorporate higher-level technologies, such as fog and cloud spaces, for computing processes or computational units in distributed data clusters. Figure 6 illustrates a model of this mobile app for diagnosis.
In the case of mobile apps that are used as data transmitters at the perception level, we suggest using 5th-generation (5G) instead of 4th-generation (4G) technology in centers and geographical areas that have access to this technology to transfer structured and unstructured data to a higher layer. Using this technology, and transferring health care data to the fog and cloud spaces, distributes computing and data processing. The model in Figure 7 is proposed for mobile apps with the purpose of data collection from sensors, biosensors, and monitoring equipment.
We compared the 42 studies in terms of the requirements for AI methods. In using AI algorithms, the volume of the data set, overfitting prevention methods, and method lightness (for use in mobile phones) are the parameters for analyzing the quality of studies incorporating AI methodology in mobile phones. Since the included studies did not mention these cases, future studies should adopt methods to prevent overfitting when using ML methods. In ML algorithms, it is advised to use 3 methods: early stopping, dropout, and cross-validation. In addition to the mentioned methods, data augmentation techniques should be used in DL methods so that the resulting model is free of any overfitting.

Studies using DL algorithms in smartphones should carefully consider the following points:

- Data preprocessing plays a vital role in model convergence and speed [54]. It is recommended that future studies use data preprocessing techniques, especially for images.
- Model parameters must be optimized to develop a robust and valid model. Future research should adopt the concept of transfer learning and pretrained networks, and these networks should be customized to the investigation.
- Mobile phones have a limited processing unit. As a result, it is suggested that the models be designed using cutting-edge techniques. It is preferred to use lightweight pretrained CNN networks, such as MobileNet, MobileNetV2, and Efficient, to obtain lightweight models that efficiently run on smartphones.
- In studies using mobile phones for data collection and visualization, we recommend using cloud storage and 5G technologies that significantly contribute to the comprehensiveness of the data and image visualization.

**Conclusion**

Mobile technology, including various apps, can help with COVID-19 diagnosis and detection and play a vital role in controlling COVID-19 outbreak. Contact tracing can prevent additional contacts during an epidemic or pandemic outbreak of any disease. Thus, on the front lines of outbreak control, healthy people can be separated from people with infection and be alerted through their mobile phones. In the second step, mobile app technology, biosensors (for rapid diagnosis), and AI methods (for diagnosis in the early and acute stages of the disease) can reduce high mortality rates and minimize the consumption of hospital resources. In the third step, mobile technology as a powerful tool can help clinicians form repositories of clinical data and signs and symptoms and collect data from individual smartphones to create such repositories and big data. This can shed more light on COVID-19, its symptoms, the prognosis, and treatment outcomes. In this and future pandemics, smartphones and their apps can be an integral part of controlling the disease and improving patients’ survival.
References


Abbreviations

5G: 5th generation
AI: artificial intelligence
API: application programming interface
AQ: analytical question
CNN: convolutional neural network
CT: computed tomography
IoT: internet of things
DL: deep learning
ML: machine learning
PCR: polymerase chain reaction
PoC: point of care
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SpO2: saturation of peripheral oxygen

©Mehdi Gheisari, Mustafa Ghaderzadeh, Huxiong Li, Tania Taami, Christian Fernández-Campusano, Hamidreza Sadeghsalehi, Aaqif Afzaal Abbasi. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 22.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Review

Lessons and Untapped Potential of Smartphone-Based Physical Activity Interventions for Mental Health: Narrative Review

Emily E Bernstein1,2, PhD; Emma C Wolfe3, BA; Brynn M Huguenet1,2, PhD; Sabine Wilhelm1,2, PhD

1Department of Psychiatry, Massachusetts General Hospital, Boston, MA, United States
2Department of Psychiatry, Harvard Medical School, Boston, MA, United States
3Department of Psychology, University of Virginia, Charlottesville, VA, United States

Corresponding Author:
Emma C Wolfe, BA
Department of Psychology
University of Virginia
Gilmer Hall
485 McCormick Road
Charlottesville, VA, 22903
United States
Phone: 1 3014665679
Email: rpu3zk@virginia.edu

Abstract

Background: Physical activity has well-known and broad health benefits, including antidepressive and anxiolytic effects. However, only approximately half of Americans meet even the minimum exercise recommendations. Individuals with anxiety, depression, or related conditions are even less likely to do so. With the advent of mobile sensors and phones, experts have quickly noted the utility of technology for the enhanced measurement of and intervention for physical activity. In addition to being more accessible than in-person approaches, technology-driven interventions may uniquely engage key mechanisms of behavior change such as self-awareness.

Objective: This study aims to provide a narrative overview and specific recommendations for future research on smartphone-based physical activity interventions for psychological disorders or concerns.

Methods: In this paper, we summarized early efforts to adapt and test smartphone-based or smartphone-supported physical activity interventions for mental health. The included articles described or reported smartphone-delivered or smartphone-supported interventions intended to increase physical activity or reduce sedentary behavior and included an emotional disorder, concern, or symptom as an outcome measure. We attempted to extract details regarding the intervention designs, trial designs, study populations, outcome measures, and inclusion of adaptations specifically for mental health. In taking a narrative lens, we drew attention to the type of work that has been done and used these exemplars to discuss key directions to build on.

Results: To date, most studies have examined mental health outcomes as secondary or exploratory variables largely in the context of managing medical concerns (eg, cancer and diabetes). Few trials have recruited psychiatric populations or explicitly aimed to target psychiatric concerns. Consequently, although there are encouraging signals that smartphone-based physical activity interventions could be feasible, acceptable, and efficacious for individuals with mental illnesses, this remains an underexplored area.

Conclusions: Promising avenues for tailoring validated smartphone-based interventions include adding psychoeducation (eg, the relationship between depression, physical activity, and inactivity), offering psychosocial treatment in parallel (eg, cognitive restructuring), and adding personalized coaching. To conclude, we offer specific recommendations for future research, treatment development, and implementation in this area, which remains open and promising for flexible, highly scalable support.

(JMIR Mhealth Uhealth 2024;12:e45860) doi:10.2196/45860

KEYWORDS

smartphone; digital health; exercise; physical activity; mental health; depression; anxiety; mobile phone
Introduction

Background

In the 21st century, anxiety and depression have been among the top 25 causes of global disease burden [1]. The COVID-19 pandemic has only intensified the rising prevalence as well as the personal and societal impacts of these disorders [2,3]. As there are simply not enough mental health professionals to meet this need [4,5], alternative interventions—both preventive and curative—are urgently needed. Targeting physical activity is a clear opportunity. Before the pandemic, more than 1 in 4 adults reported sitting for >8 hours per day [6], and this number rose to >40% during the pandemic [7]. Both of these statistics are likely underestimated [8]. Prolonged sedentary behavior, or extended time spent awake with minimal energy expenditure [9], is associated with more severe anxiety and depression as well as higher odds of developing related disorders [10-14]. In contrast, decades of research have demonstrated that regular physical activity is associated with numerous positive psychological outcomes [15-17]. Cross-sectionally, individuals who engage in regular exercise—a subset of physical activity that involves planned, structured, and repetitive bodily movement intended to improve or maintain fitness [18]—report fewer and less severe symptoms of anxiety and depression [19,20], greater positive affect and well-being [21], less stress [22], and lower rates of anxiety and depressive disorder diagnoses [23-25]. At the individual level, people report feeling better on days when they exercise [26-28].

Exercise as an Intervention for Depression and Anxiety

Prospective data support regular exercise as a potent population-level prevention tool, significantly lowering the risk of developing anxiety and depressive disorders [25,29,30]. Encouragingly, even small amounts of physical activity may have an enormous impact on mental health [31,32]. For example, an estimated 12% of new cases of depression could be prevented if the entire population exercised for just 1 hour per week [33]. In addition, among individuals presenting with diagnosable physical activity is associated with numerous positive psychological outcomes [15-17]. Cross-sectionally, individuals who engage in regular exercise—a subset of physical activity that involves planned, structured, and repetitive bodily movement intended to improve or maintain fitness [18]—report fewer and less severe symptoms of anxiety and depression [19,20], greater positive affect and well-being [21], less stress [22], and lower rates of anxiety and depressive disorder diagnoses [23-25]. At the individual level, people report feeling better on days when they exercise [26-28].

Prospective data support regular exercise as a potent population-level prevention tool, significantly lowering the risk of developing anxiety and depressive disorders [25,29,30]. Encouragingly, even small amounts of physical activity may have an enormous impact on mental health [31,32]. For example, an estimated 12% of new cases of depression could be prevented if the entire population exercised for just 1 hour per week [33]. In addition, among individuals presenting with diagnosable physical activity is associated with numerous positive psychological outcomes [15-17]. Cross-sectionally, individuals who engage in regular exercise—a subset of physical activity that involves planned, structured, and repetitive bodily movement intended to improve or maintain fitness [18]—report fewer and less severe symptoms of anxiety and depression [19,20], greater positive affect and well-being [21], less stress [22], and lower rates of anxiety and depressive disorder diagnoses [23-25]. At the individual level, people report feeling better on days when they exercise [26-28].

Exercise appears to enhance emotional flexibility, or a person’s ability to self-regulate under stress [37-40]. Physiologically, individuals who exercise more regularly show faster heart rate recovery following stressors than their peers who exercise less, and individual bouts of exercise can mitigate the hypothalamic-pituitary-adrenal axis, heart rate, and blood pressure reactivity to acute stress [39,41-43]. Physical activity can also increase the production of brain-derived neurotrophic factors, which are neurobiological changes that are understood to increase resilience [44]. These effects are mirrored in reports of exercise bolstering emotional recovery following stressors, enhancing coping self-efficacy, and mitigating the impact of rumination and other emotion regulation deficits on prolonging distress [34,39,42,45,46]. Furthermore, exercise benefits physical health targets that share bidirectional relationships with mental health, such as better sleep and cardiometabolic health [47-51]. Critically, positive treatment effects have been found for directly alleviating anxiety [52-55] and depressive [56-58] disorders as well as related and frequently comorbid conditions such as posttraumatic stress disorder [59] and obsessive-compulsive disorder [60,61]. Importantly, research has replicated the benefits of physical activity (ie, reducing psychiatric symptoms) in samples of people with severe mental illness, such as schizophrenia [62]. Similarly, physical activity and other health-related behaviors (eg, sleep hygiene) are considered to be an integral component of treatment for bipolar disorder [63]. Exercise has also been successfully used to augment the effects of other validated psychosocial treatments such as cognitive behavioral therapy (CBT) [64-67].

Despite the broad knowledge that regular physical activity is physically, cognitively, and emotionally beneficial, only approximately half of Americans meet even the minimum exercise recommendation of 150 minutes per week of moderate-intensity or equivalent physical activity [68]. Individuals with anxiety, depression, or related conditions are even less likely to do so [69,70]. They are also more likely than peers without mental health disorders to exhibit elevated sedentary behavior [71,72]. Thus, although acceptable and efficacious tools exist to help individuals meaningfully change their behavior and improve psychiatric symptoms, there is a large gap between the evidence and real-world implementation. Few clinicians include physical activity as an explicit treatment target or use it as an intervention tool [73-75]. Moreover, the larger barriers to treatment within our health care system remain, including the inaccessibility of treatment due to the acute shortage of qualified clinicians; stigma; and patients’ difficulty with travel, timing, and the cost of regular appointments [76].

Promise of Digital Platforms for Promoting Physical Activity

With the advent of mobile sensors and phones, experts have quickly noted the ability of technology to expand the reach of evidence-based psychiatric care; overcome the aforementioned barriers by providing treatment flexibly; and begin reducing long-standing disparities in treatment access, response, and dropout [77-79]. This could also be an efficient, scalable method of promoting increased physical activity among adults with or at risk of anxiety and depressive disorders.

Digital solutions show strong early benefits for activity measurement and intervention in nonpsychiatric populations. In fact, leveraging technology to measure and increase physical activity was an official recommendation from the National Heart, Lung, and Blood Institute and National Institute on Aging “Influences on sedentary behavior/Interventions to reduce sedentary behavior” joint workshop [80]. First, measuring behavior via mobile sensor is validated and widely used [81-84]. Similarly, people’s tendency to carry their phones with them throughout the day allows for more accurate monitoring of physical activity and related progress. The ease of use and, therefore, precision of such technologies (wearable and smartphone-based sensors) is an important boon for research and treatment as self-report measures of activity typically result in underreporting [81-84].

Second, delivering treatment in part or fully via mobile phones is effective in increasing physical activity in nonpsychiatric populations [85-87]. This parallels broader findings that apps
can effectively promote other health behaviors (eg, improved nutrition, smoking cessation, and medication adherence) [88,89]. Currently, 97% of Americans have a mobile phone, and an estimated 85% have a smartphone [90]. Although these numbers are lower in certain populations, such as those with serious mental illness (wherein an estimated 85% own a mobile phone and 60%-70% own a smartphone), the ubiquity of smartphones allows for the promotion of behavior change in real time and with a wider array of individuals [91-94]. Inactivity frequently occurs out of conscious awareness or choice due to people’s attention being fixated elsewhere (eg, watching television or taking the elevator at work). As such, personal devices can unobtrusively enhance awareness of behavior, which itself can promote increased activity [95]. Furthermore, technologies can deliver notifications in the moment to interrupt passive episodes while also providing tools to increase activity when individuals are most likely to take action [96]. Mobile app–based physical activity interventions can also gamify exercise to enhance enjoyment, which is a key mechanism for long-term engagement in physical activity [97,98]. In-the-moment enjoyment not only promotes regular exercise but is also in itself beneficial for mental health, contributing to the success of broader interventions such as behavioral activation. Overall, digital interventions are promising as they are low risk (ie, typically focusing on reducing sitting and increasing light activity), can be deployed without a clinician, and can be used in the context of a patient’s daily life. Ultimately, research conducted thus far in the general population supports mobile technologies as valid, accessible, and effective methods of promoting physical activity and reducing sedentary behavior.

**Current Objective**

Although it is reasonable to extrapolate that physical activity interventions could be implemented via smartphone in a similarly feasible, acceptable, and effective manner in psychiatric populations or for psychiatric targets, this remains an open question. High-quality trials of in-person exercise programs for mental health often unintentionally include components beyond the activity itself that are potentially therapeutic, such as regular, structured, and supervised sessions [65]. In other words, as part of most exercise programs, participants also receive regular social engagement or support, face-to-face time with a professional, instruction and demonstration of target behaviors, and guidance with behavioral scheduling or activation, all of which may enhance the therapeutic benefits of physical activity. However, the remote and asynchronous nature of technology-based interventions may provide a different experience from that of in-person programs, and thus, the impact may also differ. On the other hand, the aforementioned benefits of digital interventions, such as their ability to increase accessibility, lower logistical barriers to engagement, and enhance self-awareness while also promoting behavior change in real time, may boost response and, thus, lead to comparable—or even stronger—effects than face-to-face trials. As a result, it cannot be assumed that face-to-face physical activity programs or digital programs designed for other populations (eg, medical) will translate when delivered via smartphone or to a new population.

In this study, we explored the potential of physical activity interventions, as delivered (at least in part) via smartphone, to improve mental health in psychiatric populations. As this topic remains relatively new, we also considered available evidence on these tools to address mental health symptoms in nonpsychiatric populations. Specifically, we highlighted in which populations these tools have been tested; what outcomes have been evaluated (eg, acceptability, behavior change, and symptom change); and how strategies and tools have (or have not) been tailored to individuals with depression, anxiety, or related concerns. The goal was to provide a narrative overview and specific recommendations for future research on smartphone-based physical activity interventions for psychological disorders or concerns.

**Methods**

**Literature Search**

To provide a narrative overview of this emerging research area, we searched for articles that (1) described or reported an intervention intended to increase physical activity or reduce sedentary behavior; (2) included an emotional disorder, concern, or symptom as an outcome measure; (3) described or reported an intervention delivered entirely or in part via a smartphone app; and (4) were published in English and in peer-reviewed journals. Example search terms include “smartphone,” “smartphone application,” “mobile application,” “mobile app,” “digital mental health,” “app-based,” “app-assisted,” “mobile phone,” “ehealth,” “digital,” “mobile,” “exercise,” “physical activity,” “sedentary,” “sedentary behavior,” “physical inactivity,” “depression,” “dysthymia,” “mood,” “anxiety,” “phobia,” “trauma,” “posttraumatic stress,” “obsessive compulsive disorder,” “post-traumatic stress,” “obsessive-compulsive disorder,” “stress,” “emotional disorder,” “emotional problem,” “well-being,” “wellness,” “affective disorder,” “OCD,” “PTSD,” “MDD,” “GAD,” “mental health,” and “mental illness.” Web-based database (PubMed, Google Scholar, and Cochrane) searches and additional manual searches (eg, searching the reference sections of articles identified through database searches) were conducted up to March 2022. Records were initially reviewed by one coauthor; in cases of uncertainty about appropriateness for this review, records were reviewed in full by 2 additional coauthors and discussed until a consensus was reached.

**Data Review**

We attempted to extract the following information, where available, from each paper: sample size, inclusion criteria, demographics of the sample, primary aim, trial design (eg, randomized controlled trial), treatment duration, technology used, other interventions used (ie, in addition to physical activity), outcome measures (eg, primary and secondary measures of physical activity), results, treatment components or behavior change strategies, adaptations for mental health, and inclusion of coaching.
In taking a narrative lens, we drew attention to the type of work that has been done and used these exemplars to discuss key directions to build on. Table 1 provides a summary of the included articles.
<table>
<thead>
<tr>
<th>Study, year</th>
<th>Sample size, N</th>
<th>Population studied</th>
<th>Intervention used</th>
<th>Duration</th>
<th>Physical activity outcome</th>
<th>Primary psychiatric outcome</th>
<th>Primary medical outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aguilera et al [99], 2020</td>
<td>N/A&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Adults with diabetes and a score of &gt;5 on the PHQ&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Apps: DIA-MANTE to track data and deliver adaptive learning algorithm (active only) and HealthySMS to send messages (active and control)</td>
<td>6 months</td>
<td>Steps</td>
<td>Depressive symptoms (PHQ-8&lt;sup&gt;d&lt;/sup&gt;)</td>
<td>HbA&lt;sub&gt;1c&lt;/sub&gt; levels (blood glucose)</td>
</tr>
<tr>
<td>Broers et al [100], 2019</td>
<td>557</td>
<td>Adults diagnosed with hypertension, symptomatic heart failure, or coronary artery disease</td>
<td>Wearable (Fitbit) activity tracker, wearable (Beddit) sleep tracker, app (Moves) GPS tracker, app (Care-portal) home monitoring system</td>
<td>6 months</td>
<td>Steps; physical activity level (combined length of active periods and step count)</td>
<td>Anxiety (GAD-7&lt;sup&gt;f&lt;/sup&gt;) and depression (PHQ-9&lt;sup&gt;g&lt;/sup&gt;)</td>
<td>N/A</td>
</tr>
<tr>
<td>Damschroder et al [101], 2020</td>
<td>357</td>
<td>Web-based confirmation of veteran status</td>
<td>App (Stay Strong) and wearable (Fitbit Charge 2)</td>
<td>12 months</td>
<td>Active minutes per week; steps</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Edney et al [102], 2020</td>
<td>444</td>
<td>Adults currently completing &lt;150 minutes of MVPA&lt;sup&gt;b&lt;/sup&gt; per week</td>
<td>App (Active Team) and wearables (pedometer and Zencro TW64S)</td>
<td>3 months</td>
<td>Daily minutes of MVPA</td>
<td>Symptoms of anxiety, depression, and stress (DASS-D&lt;sup&gt;j&lt;/sup&gt;)</td>
<td>N/A</td>
</tr>
<tr>
<td>García-Estela et al [103], 2021&lt;sup&gt;a&lt;/sup&gt;</td>
<td>N/A</td>
<td>Spanish-speaking adults with MADRS&lt;sup&gt;i&lt;/sup&gt; score of &gt;1</td>
<td>App (IDEApp) and wearable (smartband)</td>
<td>8 months</td>
<td>SIMPAQ&lt;sup&gt;k&lt;/sup&gt;; functional exercise capacity (6MWT&lt;sup&gt;l&lt;/sup&gt; and 1-min sit-to-stand test); short Borg CR-10&lt;sup&gt;n&lt;/sup&gt; Scale</td>
<td>Depressive symptoms (PHQ-9&lt;sup&gt;o&lt;/sup&gt;); well-being (WHO-5&lt;sup&gt;p&lt;/sup&gt;)</td>
<td>Global functioning (SF-36v2&lt;sup&gt;q&lt;/sup&gt;)</td>
</tr>
<tr>
<td>Guo et al [104], 2020</td>
<td>300</td>
<td>Adults who are HIV-seropositive with elevated depressive symptoms</td>
<td>App (WeChat) and Run4Love program (adapted CB-SMP&lt;sup&gt;p&lt;/sup&gt; course and physical activity promotion) delivered through the WeChat app</td>
<td>3 months</td>
<td>Chinese version of the GPAQ&lt;sup&gt;q&lt;/sup&gt;</td>
<td>Depressive symptoms (CES-D&lt;sup&gt;q&lt;/sup&gt;)</td>
<td>N/A</td>
</tr>
<tr>
<td>Haufe et al [105], 2020</td>
<td>314</td>
<td>Adults with metabolic syndrome</td>
<td>Wearable (activity monitor; Forerunner 35; Garmin)</td>
<td>6 months</td>
<td>Freiburger Questionnaire on Physical Activity; steps</td>
<td>Anxiety severity and depression severity (HADS&lt;sup&gt;s&lt;/sup&gt;)</td>
<td>Change in metabolic syndrome severity; health-related quality of life (SF-36&lt;sup&gt;t&lt;/sup&gt;)</td>
</tr>
<tr>
<td>Kim et al [106], 2021</td>
<td>21</td>
<td>Adults aged &gt;46 years diagnosed with PD&lt;sup&gt;u&lt;/sup&gt; or atypical parkinsonism conditions and regular participation in a PD exercise program at least once a week</td>
<td>App (researcher-created physical activity app)</td>
<td>8 weeks</td>
<td>Total exercise calculated by multiplying the frequency and duration for all exercises; subjective exercise scale (Borg 6-20 scale); IPAQ&lt;sup&gt;v&lt;/sup&gt;</td>
<td>Depression (Geri-atric Depression Scale–Short Form)</td>
<td>PDQ-39&lt;sup&gt;w&lt;/sup&gt;</td>
</tr>
<tr>
<td>Lin et al [107], 2020&lt;sup&gt;a&lt;/sup&gt;</td>
<td>N/A</td>
<td>Physically inactive adults with a musculoskeletal diagnosis (ICD-10&lt;sup&gt;x&lt;/sup&gt;) and in rehabilitation following inpatient clinic treatment</td>
<td>App (MoVo)</td>
<td>12 months</td>
<td>BSA&lt;sup&gt;y&lt;/sup&gt;, sport activity and movement activity subscales</td>
<td>Depression (PHQ-9&lt;sup&gt;z&lt;/sup&gt;); anxiety (GAD-7)</td>
<td>Brief Pain Inventory</td>
</tr>
<tr>
<td>Study, year</td>
<td>Sample size, N</td>
<td>Population studied</td>
<td>Intervention used</td>
<td>Duration</td>
<td>Physical activity outcome</td>
<td>Primary psychiatric outcome</td>
<td>Primary medical outcome</td>
</tr>
<tr>
<td>-------------</td>
<td>---------------</td>
<td>------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------</td>
<td>----------</td>
<td>--------------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>Ma et al</td>
<td>N/A</td>
<td>Adult participants who were obese and experiencing depression</td>
<td>Wearable (Fitbit) and app or website (MyFitnessPal)</td>
<td>12 months</td>
<td>Minutes of physical activity logged on MyFitnessPal</td>
<td>Depression (SCL-20&lt;sup&gt;a&lt;/sup&gt;)</td>
<td>Changes in BMI</td>
</tr>
<tr>
<td>[108], 2015</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nadal et al</td>
<td>N/A</td>
<td>Adult users with mild to moderate depression who were assigned to iCBT&lt;sup&gt;aa&lt;/sup&gt; treatment for depression</td>
<td>Wearable (smartwatch; Mood Monitor watch app)</td>
<td>8 weeks</td>
<td>Smartwatch activity data</td>
<td>Depression (PHQ-9&lt;sup&gt;a&lt;/sup&gt;), anxiety (GAD-7&lt;sup&gt;a&lt;/sup&gt;), and functional impairment (WSAS&lt;sup&gt;ab&lt;/sup&gt;)</td>
<td>N/A</td>
</tr>
<tr>
<td>[109], 2021</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Park et al</td>
<td>60</td>
<td>Adults with a history of cardiovascular disease who were within 2 weeks of completing cardiac rehabilitation</td>
<td>Wearable (Fitbit Charge 2) and apps (Movi and Fitbit)</td>
<td>2 months</td>
<td>Steps; 6MWT; self-reported physical activity</td>
<td>Quality of life (QLESQ&lt;sup&gt;ac&lt;/sup&gt;); depression (PHQ-9&lt;sup&gt;b&lt;/sup&gt;)</td>
<td>N/A</td>
</tr>
<tr>
<td>[110], 2021</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Puszkielwicz et al [111], 2016</td>
<td>13</td>
<td>Adults with a diagnosis of breast, prostate, or colorectal cancer who had finished primary curative treatment</td>
<td>App (GAINFitness)</td>
<td>6 weeks</td>
<td>Physical activity (GLTEQ&lt;sup&gt;ad&lt;/sup&gt;)</td>
<td>Health and quality of life outcomes (EQ-5D&lt;sup&gt;a&lt;/sup&gt;); well-being (FACT-G&lt;sup&gt;ae&lt;/sup&gt;); anxiety and depression (HADS)</td>
<td>Cancer-related fatigue (FACT-G&lt;sup&gt;ae&lt;/sup&gt;); sleep quality (PSQI&lt;sup&gt;ag&lt;/sup&gt;)</td>
</tr>
<tr>
<td>Puterman et al [112], 2021</td>
<td>334</td>
<td>Adults with a score of 1-3 on the L-CAT&lt;sup&gt;ab&lt;/sup&gt; who were cleared to exercise</td>
<td>App (Down Dog suite of apps—HIIT&lt;sup&gt;ai&lt;/sup&gt; and yoga)</td>
<td>6 weeks</td>
<td>Sessions of yoga or HIIT completed; minutes of yoga or HIIT completed</td>
<td>Depressive symptoms (CESD)</td>
<td>N/A</td>
</tr>
<tr>
<td>Skrepnik et al [113], 2017</td>
<td>172</td>
<td>Adults with osteoarthritis eligible to receive the hylan G-F 20 injection</td>
<td>Wearable (Jawbone UP24) and app (OA GO)</td>
<td>90 days</td>
<td>Steps</td>
<td>VAMS&lt;sup&gt;aj&lt;/sup&gt;</td>
<td>Changes in sleep captured by the wearable activity monitor</td>
</tr>
<tr>
<td>Stephens et al [114], 2022</td>
<td>15</td>
<td>Youth (aged 11 to ≤18 years) with MS&lt;sup&gt;ak&lt;/sup&gt; and a disability rating of ≤4 on the EDSS&lt;sup&gt;al&lt;/sup&gt; and attending a pediatric MS and neurominflamatory disorder clinic</td>
<td>App (Atomic)</td>
<td>12 weeks</td>
<td>Physical activity measured via accelerometry; time spent in MVPA and sedentary activities; aerobic fitness, musculoskeletal strength, and walking endurance</td>
<td>Depression (CES-DC&lt;sup&gt;am&lt;/sup&gt;)</td>
<td>N/A</td>
</tr>
<tr>
<td>Teychenne et al [115], 2021</td>
<td>62</td>
<td>Mothers 3-9 months post partum, insufficiently active, and experiencing heightened depressive symptoms</td>
<td>App (smartphone app and web forum)</td>
<td>12 weeks</td>
<td>Self-reported physical activity; accelerometer-assessed physical activity and sedentary behavior</td>
<td>Depressive and anxiety symptoms (un-standardized questionnaires)</td>
<td>N/A</td>
</tr>
<tr>
<td>Wilczynska et al [116], 2020</td>
<td>N/A</td>
<td>Adults with or at risk of type 2 diabetes</td>
<td>App (eCoFIT)</td>
<td>20 weeks</td>
<td>N/A</td>
<td>Depressive and anxiety symptoms (PHQ-9 and GAD-7&lt;sup&gt;g&lt;/sup&gt;)</td>
<td>Social support, self-efficacy, nature relatedness, and perceived sleep quality</td>
</tr>
<tr>
<td>Study, year</td>
<td>Sample size, N</td>
<td>Population studied</td>
<td>Intervention used</td>
<td>Duration</td>
<td>Physical activity outcome</td>
<td>Primary psychiatric outcome</td>
<td>Primary medical outcome</td>
</tr>
<tr>
<td>------------------</td>
<td>----------------</td>
<td>--------------------------------------------</td>
<td>--------------------------------</td>
<td>----------</td>
<td>--------------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Wong et al [117], 2021</td>
<td>79</td>
<td>Adults with moderate depressive symptoms</td>
<td>App (Lifestyle Hubl)</td>
<td>8 weeks</td>
<td>Physical activity level (IPAQ)</td>
<td>Depressive and anxiety symptoms (PHQ-9 and GAD-7)</td>
<td>Insomnia (ISI&lt;sup&gt;a&lt;/sup&gt;); health-related quality of life; health-promoting behaviors (HPLP-II&lt;sup&gt;b&lt;/sup&gt;); functional impairment (SDS&lt;sup&gt;c&lt;/sup&gt;)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Published protocol.<br>
<sup>b</sup>N/A: not applicable.<br>
<sup>c</sup>PHQ: Patient Health Questionnaire.<br>
<sup>d</sup>PHQ-8: Patient Health Questionnaire-8.<br>
<sup>e</sup>HbA<sub>1c</sub>: hemoglobin A<sub>1c</sub>.<br>
<sup>f</sup>GAD-7: Generalized Anxiety Disorder-7.<br>
<sup>g</sup>PHQ-9: Patient Health Questionnaire-9.<br>
<sup>h</sup>MVPA: moderate to vigorous physical activity.<br>
<sup>i</sup>DASS-D: Depression Anxiety Stress Scales: Depression Subscale.<br>
<sup>j</sup>MADRS: Montgomery–Åsberg Depression Rating Scale.<br>
<sup>k</sup>SIMPAQ: Simple Physical Activity Questionnaire.<br>
<sup>l</sup>6MWT: 6-Minute Walk Test.<br>
<sup>m</sup>CR-10: Borg Category-Ratio scale.<br>
<sup>n</sup>WHO-5: World Health Organisation–Five Well-Being Index.<br>
<sup>o</sup>SF-36v2: 36-Item Short Form Health Survey version 2.<br>
<sup>p</sup>CBSM: cognitive behavioral stress management.<br>
<sup>q</sup>GPAQ: Global Physical Activity Questionnaire.<br>
<sup>r</sup>CES-D: Center for Epidemiologic Studies Depression Scale.<br>
<sup>s</sup>HADS: Hospital Anxiety and Depression Scale.<br>
<sup>t</sup>SF-36: 36-Item Short Form Health Survey.<br>
<sup>u</sup>PD: Parkinson disease.<br>
<sup>v</sup>IPAQ: International Physical Activity Questionnaire.<br>
<sup>w</sup>PDQ-39: Parkinson Disease Questionnaire-39.<br>
<sup>x</sup>CD-10: International Statistical Classification of Diseases, Tenth Revision.<br>
<sup>y</sup>BSA: Movement and Sport Activity Questionnaire.<br>
<sup>z</sup>SCL-20: Symptom Checklist Depression Scale.<br>
<sup>aa</sup>iCBT: internet-based cognitive behavioral therapy.<br>
<sup>ab</sup>WSAS: Work and Social Adjustment Scale.<br>
<sup>ac</sup>QLESQ: Quality of Life Enjoyment and Satisfaction Questionnaire.<br>
<sup>ad</sup>GLTEQ: Godin Leisure-Time Exercise Questionnaire.<br>
<sup>ae</sup>FACT-G: Functional Assessment of Cancer Therapy–General.<br>
<sup>af</sup>FACT-F: Functional Assessment of Chronic Illness Therapy.<br>
<sup>ag</sup>PSQI: Pittsburgh Sleep Quality Index.<br>
<sup>ah</sup>L-CAT: Stanford Leisure-Time Categorical Activity Item.<br>
<sup>ai</sup>HIIT: high-intensity interval training.<br>
<sup>aj</sup>VAMS: visual analog mood scale.<br>
<sup>ak</sup>MS: multiple sclerosis.<br>
<sup>al</sup>EDSS: Expanded Disability Status Scale.<br>
<sup>am</sup>CES-DC: Center for Epidemiologic Studies Depression Scale for Children.<br>
<sup>an</sup>ISI: Insomnia Severity Index.<br>
<sup>ao</sup>HPLP-II: Health-Promoting Lifestyle Profile-II.<br>
<sup>ap</sup>SDS: Sheehan Disability Scale.
Who Was Included in This Work?

Overall, a review of the literature revealed that little work has been done to test the impact of smartphone-based physical activity interventions on increasing physical activity or reducing mental health symptoms in psychiatric or at-risk populations. Most related trials with clinical populations have been conducted in the area of medicine, with studies investigating the effects of physical activity—encouraged through smartphone- and wearable-based interventions—on physical health conditions (e.g., diabetes [99,116], obesity [86,108], cancer [111,118], cardiovascular issues [100,110], and multiple sclerosis [114]). Physical activity is well established as a means of facilitating rehabilitation following serious illness or injury, as well as mitigating the progression of chronic health conditions [119-121]. For example, adults with Parkinson disease who used a minimally supported, customizable home-based exercise app for 8 weeks doubled their amount of weekly exercise (minutes) while also increasing the intensity of such exercise [106]. Similarly, engagement with smartphone-based physical activity interventions led to increased strenuous exercise among adults with cancer [111] as well as increased step count for those with cardiac issues [100,110] and for youths with multiple sclerosis [114].

In contrast, the practice of formally integrating exercise into mental health care is relatively new and not established in current standards of care. This is mirrored by the disproportion of extant research examining digital tools for increasing physical activity in medical versus psychiatric populations. When mental health targets were examined, they were largely included as secondary or exploratory outcomes and frequently framed in relation to coping with the medical concern of interest [99,100,104-108,110,111,113,114]. We only identified 16% (3/19) of the studies that specifically recruited individuals with psychiatric symptoms, and all (3/3, 100%) were focused on individuals with depressive symptoms [103,109,117]. The most common mental health outcomes were depression, anxiety, general quality of life, and emotional well-being [99,100,102-110,112,114-117]. Specifically, subclinical depressive concerns were the most frequently investigated psychiatric target, followed by subclinical anxiety [100-105,112,115-117]. The impact of smartphone-based physical activity interventions is yet to be investigated explicitly for individuals diagnosed with depressive or anxiety disorders, let alone other mental health conditions, including serious mental illness. Furthermore, although wide age brackets were represented across the studies, with average ages ranging from teenagers to older adults, men and non-White individuals were underrepresented. As research progresses in this space, it will be imperative to include the experiences and perspectives of adults with clinical levels of psychiatric concerns as well as diverse backgrounds and identities.

Do Smartphone-Based Physical Activity Interventions Benefit Mental Health?

Owing to the limited available data; small samples comprising mostly White Western, educated, industrialized, rich, and Democratic women with subclinical depression or anxiety; and heterogeneity of outcomes measured, it is difficult to conclude whether and to what extent existing smartphone-based physical activity interventions benefit mental health. However, with these caveats in mind, we aimed to synthesize the available evidence in the following sections.

Feasibility and Acceptability

There is encouraging evidence that smartphone-based physical activity interventions could be feasible and acceptable for psychiatric populations. However, supporting data were primarily collected in samples of individuals with elevated depressive symptoms or who were at risk of depression rather than in explicitly clinical samples or among individuals with other prominent mental health concerns. One study of postnatal women at risk of depression found low engagement with the digital aspects of a 12-week multicomponent physical activity intervention (home exercise equipment and a physical logbook combined with a motivational smartphone app and a web-based social support forum) [115]; however, other studies reported more positive participant response and engagement. For example, retention in smartphone interventions for physical activity tended to be high compared to that in other types of digital health interventions—one systematic review found that completion rates of digital mental health interventions ranged from 1% to 28% [124]. In a study of adults with diabetes, retention in a 20-week digital physical activity intervention (an app that allowed participants to use workout circuits, set goals, monitor progress, and learn cognitive and behavioral strategies) was of >70% [99]. Similarly, compliance with a suite of high-intensity interval training (HIIT) and yoga apps during a 6-week intervention was strong in a community sample with elevated depressive symptoms. More than half of the participants in the yoga and HIIT+yoga group and 40% in the HIIT group continued completing the recommended 4 sessions per week by the end of the trial [112]. Compliance and satisfaction ratings were comparably high when physical activity promotion was combined with cognitive behavioral stress management via a WeChat intervention in a group of adults with HIV and elevated depressive symptoms [104].

These findings are consistent with those of the larger literature showing that digital physical activity interventions tend to be well received by participants [85,125,126]. Public interest is already high, with physical activity and fitness apps dominating the mobile health space. Notably, a 2018 systematic review of the experience of adults who used mobile interventions to promote physical activity highlighted important themes to be considered for future design—self-reported engagement was most enhanced by the availability of social features, prompts, goal setting, personalization or customization, and gamification but was limited by low technological literacy, preference for coached apps, and a desire for social support [127].
**Change in Psychiatric Symptoms**

The impact of smartphone-based physical activity interventions on psychiatric symptoms was far more mixed. Some studies observed resultant improvements in symptoms. In one study that included patients specifically recruited for having elevated symptoms of depression and anxiety, patients experienced a reduction in depression and anxiety scores following a 6-month exercise intervention (150 minutes of moderate physical activity per week with individual recommendations given via a smartphone app) as compared to a waitlist control. However, it is notable that this intervention did not test a smartphone-based physical activity intervention in isolation but, instead, combined it with nutritional counseling and the option of receiving exercise recommendations through personal meetings or by phone instead of an app [105]. Among adults from the general population with low physical activity scores, using a publicly available suite of exercise apps for 6 weeks significantly improved depression symptoms compared to a waitlist control [112]. Psychiatric improvements were also observed in patients with medical comorbidities. For example, in a study of adults with obesity using a smartphone-based physical activity app (eCoFIT), depression symptom severity improved after 20 weeks [116]. Finally, in a study of older adults with Parkinson disease, using a mobile app to access and customize a home-based exercise program for 8 weeks led to reduced depression symptoms and improved quality of life [106].

However, other studies reported null effects. For example, among patients with a history of cardiovascular disease, a smartphone-based intervention including motivational prompts and educational messages did not yield significant changes in depressive symptoms from baseline to 2 months [110]. Furthermore, a trial of patients with cancer using a tailored physical activity smartphone app (which included workout videos, spoken instructions, and push notifications) did not observe changes in depression, anxiety, or quality of life after 6 weeks [111]. In a study testing an app-based physical activity program for youth with multiple sclerosis (including personalized coaching and promotion of aerobic fitness, musculoskeletal strength, and walking endurance), there was no change in depression levels over 12 weeks [114]. Furthermore, a meta-analysis of studies examining digital physical activity interventions in cancer survivors found that none of the included studies were successful in improving depression or anxiety [118]. In the general population, one study similarly did not find significant differences in depression, anxiety, stress, or well-being after 3 or 9 months of using an app (Active Team) and wearable pedometer [102].

These results should be interpreted cautiously for 2 reasons. First, there is the confound of potential floor effects—as none of these trials were designed to address questions about mental health, symptom levels at study start were typically already low, thus reducing investigators’ abilities to identify possible effects. In addition, studies varied widely in their evidence of behavior change, including outcome measures (eg, minutes of activity, number of sessions completed, exercise intensity, and fitness level; see Table 1 for detailed information on this variance) and the use of objective versus subjective reports, which are known to be discrepant [82-84]. In other words, if an intervention did not produce meaningful physical activity changes, it would be unlikely that downstream emotional changes would occur.

**How Have Smartphone-Based Physical Activity Interventions Been Tailored to Individuals With Depression, Anxiety, or Other Psychiatric Concerns?**

To date, interventions generally have not been tailored to the specific needs or presentations of individuals with mental health concerns. This is unsurprising as most technology-based physical activity trials have not been designed to target mental health. However, some studies have integrated components that specifically address psychological well-being. One adaptation that is low effort but high return is adding psychoeducation about the relationship between physical activity and mental health. For example, a recent trial for adults with mild to moderate depressive symptoms devoted the first in-person group session to discuss the relationship between depression and exercise to complement the personalized exercise program, smartphone app, and wearable device they received [103].

The second adaptation observed in the literature is offering concurrent psychotherapy-based tools. In some cases, psychotherapeutic content was interspersed with the physical activity intervention; for example, in a trial for adults living with HIV and depression in China, both an exercise promotion intervention and cognitive behavioral stress management course were delivered as multimedia messages through the WeChat app [104]. Relatedly, in a study of the eCoFit app for adults with or at risk of type 2 diabetes, short cognitive behavioral tasks (“FitMind Challenges”) were integrated throughout the program [116]. Examples of FitMind Challenges included motivational strategies, relaxation, cognitive restructuring, social support, and problem-solving. In other cases, the approaches were delivered in parallel. In one trial, adults who were overweight and experiencing depression received a 7-step problem-solving therapy via a workbook in addition to live lifestyle coaching, at-home video lessons, the MyFitnessPal app, and a Fitbit for monitoring [108]. In another study, SilverCloud’s guided internet-based CBT program for depression was the primary intervention, with smartwatch-based monitoring of sleep, steps, and mood added to promote greater awareness of the relationship between health behaviors and mood, thereby independently encouraging positive lifestyle changes [109].

A third but largely unexplored avenue is the inclusion of personalized or tailored messaging. This is an opportunity for coaches or other support persons to address barriers that may be specific to the experience of someone with mental health concerns (eg, navigating social anxiety to go to the gym and restructuring depressive thoughts). In one open trial of a physical activity app for youth with multiple sclerosis, coaches were trained in social cognitive theory for behavior change as well as motivational interviewing [114].

Although few digital physical activity interventions have been designed or modified to specifically affect mental health, many have been designed using evidence- and theory-based behavior change strategies that are ripe for implementation in psychiatric contexts. Indeed, the most successful interventions are based
on behavioral theory [73,127]—explicitly stated or not—such as the transtheoretical model [128], the theory of planned behavior [129], self-determination theory [130], and social cognitive theory [131]. Social cognitive theory is most often cited given its emphasis on internal, external, and social factors that reinforce learning and contribute to sustained change [131].

Targeting self-efficacy, self-regulation, and social support to engender meaningful, lasting behavior change aligns strongly with principles of psychotherapy as well. Digital physical activity interventions have also experimented with numerous evidence-based behavior change techniques, including goal setting and review, action planning, regular feedback, self-monitoring of behavior, instruction and demonstration of how to perform a new behavior, graded tasks, prompts and cues, and social rewards, to name a few [132-134]. Interventions integrating multiple behavior change strategies are more successful than those that rely on one (eg, self-monitoring or reminders alone [133]). Considering how such strategies could be adapted for individual presentations (eg, those with clinical levels of dysregulation) should be a priority for future iterations of these programs.

Furthermore, technology-driven techniques may uniquely (or at least more strongly than traditional treatments) engage key mechanisms of behavior change. For example, these tools can promote self-awareness. As people tend to keep their devices close to them throughout their daily lives, wearable and mobile platforms can provide objective, continuous monitoring and feedback related to behavioral patterns such as physical activity [102,108]. In addition, these approaches can enhance a person’s likelihood of changing their behavior by lowering the cognitive burden involved in initiating physical activity. Strategies include delivering content more flexibly (eg, when it is most convenient for a participant to engage or in doses of their choosing); modeling target behavior via written, image, or video instructions that can be reviewed on demand or infinite times; or tailoring activity suggestions to a person’s present context (eg, suggesting at-home activities on rainy days). This may be particularly meaningful for psychiatric audiences as depression and anxiety are associated with attention and memory deficits that can interfere with information processing and learning [135-138]. In addition, in-the-moment rewards and other gamification or reinforcement features could be particularly useful early on [139,140] as individuals with depression and anxiety may not experience initial sessions of exercise as intrinsically gratifying or mood boosting as others do; for example, depression is characterized by deficits in reward processing and motivation [141], and anxiety sensitivity and social anxiety can blunt positive responses or promote avoidance [142,143].

Finally, personal devices may allow for more consistent, flexible social support throughout an intervention. Social support is an established, evidence-based behavior change technique that promotes physical activity [126,144-146]. Smartphone-based physical activity interventions provide a range of avenues for social connection, such as texting with a coach [101,107], access to an web or app-based discussion forum [115], and creation of virtual “teams” [102]. Critically, although social media has been frequently incorporated as a means of facilitating connection, participant reactions have been mixed, and it may not be optimal for psychiatric populations [126]. In general, social support appears to boost engagement when it is perceived to facilitate emotional support, provide tips from peers, enhance motivation, foster social comparison or competition [126,147,148]. How to best leverage social support and social media for psychiatric populations requires nuanced future study.

Discussion

Principal Findings

The primary aim of this narrative synthesis was to examine the status of smartphone-based physical activity interventions for mental health and understand how they have and have not been tailored to or evaluated in psychiatric populations. Ultimately, the literature is limited and difficult to synthesize owing to the high heterogeneity across the studies in terms of sample selection; study design; outcomes of feasibility, acceptability, and efficacy; and degree of tailoring. To date, mental health outcomes have typically been secondary or exploratory within trials focused on medical outcomes (eg, diabetes management) and, when included, have had a narrow focus on measures of depression, anxiety, and general well-being in nonclinical populations. As a result, this review relied significantly on research focusing on medical populations to explore how smartphone-based physical activity interventions could be used to impact mental health outcomes and to infer how they may be used in psychiatric populations. Furthermore, although extant studies have included diversity of age, the samples in the included studies comprised mostly White and female individuals, thus reducing the generalizability of the already limited findings.

The feasibility and acceptability of these interventions for subclinical and at-risk populations are encouraging and suggest that digital physical activity programs may be similarly well received among individuals above diagnostic thresholds. The available data on psychiatric outcomes were mixed. However, it is difficult to draw meaningful conclusions given the limited data; high heterogeneity of intervention approach and target behavior; and lack of standardization in measurement and reporting of use, engagement, and behavior change, as well as the elevated risk of floor effects given the subclinical samples. These inconclusive psychiatric outcomes may also be related to a lack of tailoring of smartphone-based physical activity interventions to the specific needs of those presenting with mental health concerns. The existing tailoring included basic psychoeducation about physical activity as a treatment, adding concurrent psychotherapy-based tools, and including personalized or tailored messages. There was no standardization or evidence base for how this tailoring was applied. The upshot is that many of the papers included in this review presented interventions that were already built around established, evidence-based behavior change strategies, which suggests that psychotherapeutic tailoring could be efficiently integrated into existing smartphone-delivered physical activity interventions. In general, effective physical activity interventions use many of the same fundamental behavior change strategies commonly found in psychotherapy, such as education, goal setting,
self-monitoring, graded tasks, engaging social support, and motivational interviewing [149,150].

Taken together, the primary barrier to advancing the use of smartphone-based physical activity interventions in mental health care is the absence of evidence. The need for research in this area has been highlighted in other reviews as well [151,152]. To construct a more consistent, evidence-based foundation for intervention development, we outline several avenues for future research.

**Recommendations for Tailoring Physical Activity Interventions to Psychiatric Populations**

More research is needed to better understand how existing smartphone interventions can be tailored to fit the needs of psychiatric populations. The following are example adaptations rather than an exhaustive list. One likely critical step is to provide users with psychoeducation early on about the ways in which physical activity can be used to affect psychological health, such as improving mood and reducing anxiety. This should involve making explicit connections between health behaviors (eg, exercise), mental health symptoms, and emotion regulation so that users can better appreciate the bidirectional links between these areas of well-being. Including even a brief text summary of the literature or treatment rationale could likely augment the effects [153,154]. In fact, there is evidence with depression treatment that physical activity interventions lacking such psychoeducation or treatment rationale do not lead to robust clinical changes and can worsen dropout rates [155]. In contrast, attending to the mental and emotional benefits of exercise, particularly the acute or immediate impact on affect or resilience, can further enhance mood and motivation to continue exercising [35,156]. Highlighting these benefits and encouraging users to monitor such positive changes could improve sustained engagement and clinical response. Technology may be particularly helpful for this; apps, for example, can provide in-the-moment reminders through push notifications to attend to one’s affect or visual feedback of a user’s pretest-posttest change in self-reported mood with exercise. Furthermore, digital tools could provide information about the impact that mental health symptoms may have on program engagement. This can help users recognize that it is normal and expected for symptoms such as fatigue, anxiety sensitivity, or low motivation to serve as barriers to physical exercise and can proactively help users engage in related problem-solving.

Another compelling feature to test is the incorporation of modules or content that specifically address mental health concerns or symptoms. For example, technology-based physical activity interventions aimed at improving anxiety symptoms would benefit from including evidence-based skills such as cognitive restructuring and exposure practices. These approaches can be used to identify and challenge maladaptive beliefs that anxiety symptoms such as a racing heart are dangerous (known as anxiety sensitivity [157]), design a more graded exercise plan, and use activity as an interoceptive exposure by allowing patients to experience and tolerate those feared sensations [158]. Furthermore, CBFT skills can be incorporated to address exercise-related social anxiety, such as integrating exposures (eg, walking with a friend or going to a gym first at off-peak hours) and challenging associated negative expectations (eg, “I won’t be able to keep up with my friend and they will judge me”). Meanwhile, a digital physical activity intervention aimed at helping people with depression could include skills consistent with behavioral activation, such as tracking the relationship between mood and activities (including physical activity); intentionally adding new behaviors such as gardening, walking, or going to the gym to their weekly schedule; and generating more flexible approaches to regular movement. Furthermore, cognitive skills can be used for participants to identify and evaluate negative thoughts about themselves or the program (eg, “I can never stick to my goal of going for walks, so what’s the point”) in terms of their accuracy or utility. In this vein, equipping coaches with some knowledge of common mental health symptoms to look out for, destigmatize, and address could enhance outcomes.

In addition, coaches are understood to bolster digital interventions in general by providing further psychoeducation or resources, personalizing content or skill use, and answering questions. Given the known relationships between mental illness and both inactivity and chronic medical conditions, even with basic mental health knowledge, coaches could perform these roles better. It is also important to recognize that digital or physical activity interventions may not be the appropriate or most effective level or type of care for all individuals experiencing mental health concerns. As such, tools should include information for users on the signs or symptoms that may indicate that pursuing psychotherapy could be beneficial as well as resources for doing so.

Finally, physical activity promotion tools could be integrated to augment existing treatments. Currently, there are a number of well-established treatments for psychiatric disorders, such as CBFT and mindfulness. CBFT has extensive research support as the gold-standard treatment for a range of disorders, including depression and anxiety [159]. This treatment integrates both behavioral and cognitive skills, such as tracking and scheduling activities as well as evaluating and challenging maladaptive thoughts, to reduce the severity and impact of symptoms. Mindfulness—or the purposeful, nonjudgmental awareness of the present moment [160]—is increasingly included in “third wave” interventions to reduce psychiatric symptoms [161] as well as increase well-being, such as positive affect and quality of life [162,163]. Recently, a dominant focus of digital mental health innovation has been translating these gold-standard psychotherapies to digital platforms. There is now a strong foundation of evidence supporting the feasibility, acceptability, and efficacy of delivering these treatments through both face-to-face and digital means [164-166].

Thus far, the development of digital interventions for physical activity and for mental health has largely occurred separately. However, their concurrent delivery provides promising initial evidence [111,118]. This parallels in-person trials demonstrating that increasing physical activity strengthens psychotherapy outcomes [167,168]. The next step in this line of research is to more formally integrate the 2 or even develop technologies in which both sets of skills are delivered within a single coherent platform. For example, in a study conducted by Wilczynska et al [116], adults with diabetes used the eCoFit app, which
integrated guided workouts, goal setting, and cognitive behavioral skills. Some of the commercially available apps targeting mental health have already begun moving in this direction as well. The mindfulness-based app Headspace has recently incorporated a suite of video- and audio-guided exercises that help users engage in activities such as stretching, dancing, and yoga. Within these integrated platforms, it will be important to explicitly link the mental health and physical activity content rather than presenting them side by side as distinct intervention pathways.

**Recommendations for Developing and Testing Physical Activity Interventions for Populations With Psychiatric Disorders**

**Overview**

More research is also needed to understand to what degree smartphone interventions require tailoring and for whom. Given the wide range of possibilities, an important step in the development process is to have focus groups with the goal of hearing from individuals with lived experience about their wants and needs. Through pilot-testing, intervention design and refinement can be an iterative process wherein individuals of the target audience engage with the program, feedback is elicited, and changes are made in response to that feedback. This user-centered approach fits well within the larger preparation phase of a Multiphase Optimization Strategy. Following such development, digital tools should be scientifically tested and optimized, leading to a randomized controlled trial to examine their efficacy in achieving the outcomes of interest (eg, reduction in depression symptoms). This testing phase is necessary to establish a program as evidence-based, which would allow it to stand out in an otherwise large pool of digital applications that are not backed by research.

**Leveraging New Trial Designs**

New trial designs, such as sequential multiple-assignment randomized trials, microrandomized trials, and factorial designs, will be useful in intermediate stages to parse issues such as dosing, sequencing, and personalization. For example, research shows that at least 6 weeks are required for new physical activity habits to form [169]; thus, interventions that are, on average, 8 weeks long lead to more lasting changes than shorter ones [169]. Moreover, it remains unclear whether longer treatments, such as ≥24 weeks, have a greater impact on the general population [170]. It is unknown what duration would be sufficient for various clinical populations to observe changes in both the target behavior and in downstream symptoms and for whom extended support would be necessary. Individual components such as the aforementioned tailoring elements or the inclusion of coaching can also be efficiently tested using these new study designs [170].

**Understanding the Role of Human Support**

Previous work has shown that supervised exercise tends to have a larger impact on anxiety outcomes than unsupervised prescriptions [171]. This parallels guidance from experts in digital mental health that including human support (eg, a lay coach or therapist) alongside internet- or app-based cognitive behavioral and other therapies should enhance retention, engagement, and outcomes. Possible explanations could be greater accountability, the presence of social support, clearer guidelines, opportunities to ask for clarification, in-the-moment personalization, problem-solving, direct affirmation or reinforcement, and a more regular routine. However, guidance does not unilaterally improve outcomes for all digital interventions or all patients [164,172]. There are currently no evidence-based guidelines for implementing human support in digital interventions (ie, when, how often, how much, by whom, and for which users), let alone a nuanced understanding of how coaches can effect positive change in adherence or response [173]. Understanding the mechanisms of action would allow developers to maximize automation and most efficiently deploy human support when needed. As human support is the most expensive and scarce resource in digital health solutions, it will be critical to determine how to automate some of these supportive pathways and how to most efficiently identify who needs human support and at what dose.

**Considering Individual Factors**

Across intervention types, we must also consider the individual factors that may serve as facilitators of or barriers to engagement and success. This is doubly important for smartphone-based physical activity interventions as there are potential barriers inherent in both smartphone use and physical activity uptake.

**Barriers to Physical Activity**

Research examining barriers to physical activity in those with mental health conditions suggests that individuals with high symptom severity and low self-efficacy may be particularly disinclined to pursue physical activity–based interventions [174,175]. Lack of social support, lack of available time, and fear of injury were also frequently mentioned barriers in a sample of adults with anxiety and depression [174]. In addition, individuals with higher or lower levels of baseline physical activity or fitness may face different barriers and have different needs. A qualitative review suggested that those with lower baseline physical activity wanted an app that had more of a coaching role, whereas those with a higher baseline physical activity preferred an app that helped them intensify or optimize their current physical activity level [126]. In addition, developers should consider the accessibility of exercise suggestions; for example, exercise prescriptions that necessitate equipment, a gym membership, or access to a safe outdoor space may not be generalizable to many otherwise well-suited recipients.

**Barriers to Digital Mental Health Use**

A recent review by Borghouts et al [176] examining barriers to and facilitators of user engagement found that scoring high on neuroticism and agreeableness was associated with greater interest in using smartphone apps to reduce stress, whereas scoring high on extraversion was a predictor of preferring in-person services to web-based options [177]. The severity of baseline symptoms—both psychiatric and comorbid medical concerns—may also play a role in engagement and adherence. Most smartphone-based digital physical activity interventions have been investigated in those with mild to moderate symptoms, which can hamper engagement with apps [178,179].
Some studies suggest that those with mild depression may actually be at an even greater risk of dropout than those with moderate depression [178,180]. Researchers should also be cautious about potential iatrogenic app components. For example, the tracking components inherent in many smartphone-based physical activity interventions, particularly those related to physical health and tracking activity, run the risk of becoming compulsive or rigid. This could pose an issue for individuals with obsessive-compulsive, anxiety, and related disorders that are often characterized by perfectionism or inflexibility. Furthermore, peer support groups within apps, although often helpful, could also lead to negative social comparisons, thus exacerbating depression.

A possible mitigating approach to this would be to introduce smartphone-based physical activity interventions through a stratified care model in which individuals are allocated to different levels of an intervention depending on their clinical needs. In this model, providers could use patient-level data to decide whether an individual would benefit from the smartphone tool as a stand-alone intervention, as a coached version integrated with another level of care (eg, psychotherapy with a clinician), or delivered after progress with another intervention has been made (eg, medication stabilization).

**Barriers Related to Technology**

For many, significant barriers may include technology literacy and access. Multiple studies have identified technology literacy as an obstacle to digital physical activity intervention uptake [126,181] and for digital mental health use [178,180]. This, coupled with the lack of technical support provided by many apps, means that individuals who may be motivated to engage with smartphone-based physical activity interventions are stymied by the inability to navigate the app or seek help when issues arise. This necessitates a user interface and navigability features that can be understood or customized by a range of age groups and technological ability levels, and furthermore, it emphasizes the importance of accessible, embedded technical support tools for those who need them. Coached or guided tools may be helpful in mitigating this literacy issue but still require users to have the basic skills needed to contact their coaches or guides for help. In addition, there is still a subset (15%) of the US population that does not own a smartphone, many of whom represent communities that could benefit the most from flexible, low-cost, and accessible support options [182]. An even greater percentage of Americans lack a stable internet connection; this statistic is highly stratified along racial lines—8 in 10 White adults report having a broadband connection at home, whereas only 71% of Black adults and 65% of Hispanic adults report the same [183]. This suggests a need for digital physical activity interventions that can be accessible from communal settings, such as local community centers or publicly available fitness facilities.

It should be acknowledged that the vast majority of digital mental health interventions are designed—intentionally or not—with primarily White, Western, educated, industrialized, rich, and Democratic populations in mind. Research investigating the efficacy of both in-person and digital physical activity interventions also suffers from similarly nonrepresentative samples, calling into question which validated strategies are universally beneficial. For example, there is compelling research that physical activity interventions (both in person [184,185] and virtual [186]) are effective for health behavior change in Black Americans; however, other studies investigating the perspectives of Black Americans suggest that many in this community face unique social and structural barriers to physical activity that may not be considered by extant programs [187,188]. Thus, engagement in smartphone-based physical activity interventions by marginalized populations might be impacted by both the perception and reality that many of these apps are not designed with their community or culture in mind. Future research should engage with a range of underserved populations in qualitative research to understand the community and cultural values surrounding physical activity and technology use. This information should then be used to collaboratively design new programs or features or culturally tailor existing tools to meet the needs of a broader audience.

All these limitations support the importance of qualitative work as a future direction when building smartphone-based physical activity interventions for mental health. The research synthesized by Carter et al [126]; for example, presents valuable insights into individual-level concerns and emerging trends in patient preferences for components and design of apps. In particular, their identification of 2 key mechanisms through which mobile health use facilitates physical activity (strengthening motivation and changes in self-awareness and strategizing) is an important step in boosting engagement and exploring the mechanisms by which these apps function. Engaging qualitatively with a broad spectrum of stakeholders would also be foundational in improving the issue of representation, thereby supporting the goal that smartphone-based physical activity interventions for mental health are acceptable and efficacious for all.

**Practical Considerations**

The unfortunate reality of scientifically validated digital mental health products is that the vast majority do not move beyond their success in the laboratory [173], and those that do make it to market face fierce competition, flagging engagement rates, and a lack of financial means to scale the project. Thus, in developing and testing these promising digital mental health–focused physical activity tools, investigators should integrate elements essential for successful dissemination. One proposed pathway for improving the dissemination and ultimate success of digital mental health tools is to connect consumers through employers or public and private insurance companies, who have indicated a growing interest in expanding services to cover digital mental health. To illustrate, Blue Shield of California is now offering the mindfulness meditation app Headspace to subscribers [189]; Cigna offers the mental health app Ginger as part of its service package [190]; and Kaiser Permanente supports the use of Ginger, Calm, and MyStrength [191]. Physical activity technologies that track and manage exercise and step count are even more prevalently covered by insurance. Blue Cross Blue Shield, United, and others have partnered with Fitbit to offer low-cost wearable devices and use of their apps to promote health behavior change. Aetna and Cigna offer similar programs and occasional incentives to people who use health-tracking apps and devices. Taken together, the
enthusiasm for apps and devices promoting physical health, as well as the recent foray by insurers into the digital mental health space, suggests that smartphone-based physical activity interventions for mental health may be prime for scalable coverage. This also means that academics developing such tools should be mindful when designing research studies to collect outcome data relevant to insurers and other payers, such as outcomes related to health care costs (reduction in insurance claims and physician or therapist visits), disability-adjusted life years (reduction in overall illness burden), adoption and engagement rates, and user data such as acceptability and fidelity (whether people use the tools as intended). In addition, investigators and designers should carefully consider the costs inherent to their interventions, such as relying on “off-the-shelf” versus research-grade devices and other platforms, the extent to which an intervention relies on human support to be administered, and the broader infrastructure required for implementation and sustainment, all of which will alter accessibility and scale. By designing research studies on smartphone-based physical activity interventions with true scalability in mind, researchers will be better poised to expand their intervention beyond academia and better achieve the goal of connecting evidence-based interventions with those who need them.

Conclusions

Physical activity has well-known and broad mental health benefits. However, a minority of at-risk individuals or those with mental disorders meet even the minimum exercise recommendations. Smartphones may bridge this gap given their pervasiveness in daily life, capacity to help concurrently manage multiple dimensions of personal health, and ability to engage key mechanisms of behavior change. Although early data for smartphone-based physical activity interventions reducing psychological symptoms are encouraging, overall, surprisingly little work has been done in this area. Therefore, there is untapped potential for developing and disseminating accessible, beneficial tools that can have a great public health impact.

Acknowledgments

During the preparation of this manuscript, EEB and BMH were supported in part by the Hope Fund, an internal fellowship at Massachusetts General Hospital provided through the generosity of an individual donor to support early-career investigators.

Conflicts of Interest

EEB receives research support from Koa Health, is a presenter for the Massachusetts General Hospital Psychiatry Academy in educational programs supported by independent medical education grants from pharmaceutical companies, and has a consulting agreement with Otsuka Pharmaceutical Development & Commercialization, Inc. ECW has no competing interests to declare. BMH receives research support from Koa Health. SW is a presenter for the Massachusetts General Hospital Psychiatry Academy in educational programs supported by independent medical education grants from pharmaceutical companies and has received royalties from Elsevier Publications, Guilford Publications, New Harbinger Publications, Springer, and Oxford University Press. SW has also received speaking honoraria from various academic institutions and foundations, including the International Obsessive-Compulsive Disorder Foundation, Tourette Association of America, and Centers for Disease Control and Prevention. In addition, she received payment from the Association for Behavioral and Cognitive Therapies for her role as associate editor of the Behavior Therapy journal as well as from John Wiley & Sons, Inc, for her role as associate editor of the Depression & Anxiety journal. SW has also received honoraria from One Mind for her role in the PsyberGuide scientific advisory board. SW is also on the scientific advisory board for Koa Health, Inc, and Noom, Inc. SW has received research and salary support from Koa Health, Inc. In addition, SW has a consulting agreement with Noom, Inc.

References


132. Chase JD, Otmanowski J, Rowland S, Cooper PS. A systematic review and meta-analysis of interventions to reduce sedentary activity and sedentary behavior in youth: systematic review. JMIR Mhealth Uhealth 2018 Sep 17;6(9):e10799 [FREE Full text] [doi: 10.2196/jmir.33044538]


162. Remmers C, Topolinski S, Koole SL. Why being mindful may have more benefits than you realize: mindfulness improves both explicit and implicit mood regulation. Mindfulness 2016 Apr 5;7(4):829-837 [FREE Full text] [doi: 10.1007/s12671-016-0520-1] [Medline: 28269319]


Abbreviations

CBT: cognitive behavioral therapy
HIIT: high-intensity interval training

©Emily E Bernstein, Emma C Wolfe, Brynn M Huguenel, Sabine Wilhelm. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 15.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
The Goldilocks Dilemma on Balancing User Response and Reflection in mHealth Interventions: Observational Study

Lyndsay A Nelson¹,², PhD; Andrew J Spieker³, PhD; Lauren M LeStourgeon¹,², MPH; Robert A Greevy Jr³, PhD; Samuel Molli¹,², BS; McKenzie K Roddy¹,², PhD; Lindsay S Mayberry¹,²,⁴, MS, PhD

¹ 2 3 4

Corresponding Author:
Lyndsay A Nelson, PhD

Abstract

Background: Mobile health (mHealth) has the potential to radically improve health behaviors and quality of life; however, there are still key gaps in understanding how to optimize mHealth engagement. Most engagement research reports only on system use without consideration of whether the user is reflecting on the content cognitively. Although interactions with mHealth are critical, cognitive investment may also be important for meaningful behavior change. Notably, content that is designed to request too much reflection could result in users' disengagement. Understanding how to strike the balance between response burden and reflection burden has critical implications for achieving effective engagement to impact intended outcomes.

Objective: In this observational study, we sought to understand the interplay between response burden and reflection burden and how they impact mHealth engagement. Specifically, we explored how varying the response and reflection burdens of mHealth content would impact users' text message response rates in an mHealth intervention.

Methods: We recruited support persons of people with diabetes for a randomized controlled trial that evaluated an mHealth intervention for diabetes management. Support person participants assigned to the intervention (n=148) completed a survey and received text messages for 9 months. During the 2-year randomized controlled trial, we sent 4 versions of a weekly, two-way text message that varied in both reflection burden (level of cognitive reflection requested relative to that of other messages) and response burden (level of information requested for the response relative to that of other messages). We quantified engagement by using participant-level response rates. We compared the odds of responding to each text and used Poisson regression to estimate associations between participant characteristics and response rates.

Results: The texts requesting the most reflection had the lowest response rates regardless of response burden (high reflection and low response burdens: median 10%, IQR 0%-40%; high reflection and high response burdens: median 23%, IQR 0%-51%). The response rate was highest for the text requesting the least reflection (low reflection and low response burdens: median 90%, IQR 61%-100%) yet still relatively high for the text requesting medium reflection (medium reflection and low response burdens: median 75%, IQR 38%-96%). Lower odds of responding were associated with higher reflection burden (P<.001). Younger participants and participants who had a lower socioeconomic status had lower response rates to texts with more reflection burden, relative to those of their counterparts (all P values were <.05).

Conclusions: As reflection burden increased, engagement decreased, and we found more disparities in engagement across participants’ characteristics. Content encouraging moderate levels of reflection may be ideal for achieving both cognitive investment and system use. Our findings provide insights into mHealth design and the optimization of both engagement and effectiveness.

(JMIR Mhealth Uhealth 2024;12:e47632) doi:10.2196/47632

KEYWORDS
engagement; mobile phone; text messaging; messaging; SMS; diabetes; diabetic; mobile health; mHealth; technology; user response; users; quality of life; engagement; mHealth management; management; socioeconomic; effectiveness; support person; support worker; support persons; text message; text messages; reflection; behavior change
Introduction

Background

Mobile health (mHealth) is transforming health delivery as a highly convenient and effective approach for supporting individuals with chronic conditions [1-3]. Delivered via phones, tablets, and wearables, mHealth provides education, motivation, monitoring, and other forms of support to improve health behaviors. SMS text messaging is one form of mHealth that is uniquely poised to benefit everyone, including people who are older, are disadvantaged, and are from traditionally minoritized racial or ethnic backgrounds [4-6]. A critical factor influencing mHealth effectiveness is users’ engagement or interaction with the technology, which is typically measured via system use [7-9]. Across the mHealth literature, engagement tends to be highly variable [10,11], which has spurred a whole body of research that aims to understand predictors of engagement, including user characteristics and intervention features (eg, intervention duration and frequency of sending content) [10-14]. However, very little research has attended to the type of mHealth content that users are expected to engage with [15] and, more specifically, how the content may be requesting more or less cognitive reflection.

The primary goal in having users engage with mHealth content is health behavior change. With respect to mHealth interventions, there is a hyperfocus on wanting the user to interact with the technology (eg, responding to a text message), with less consideration of whether the user is reflecting on the content cognitively (eg, reflecting on past behavior and planning future behavior) [8]. Although interaction with the technology is a critical measure, there is a growing consensus that cognitive investment is also important for meaningful behavior change in many types of mHealth interventions [16-18]. Notably, content may be designed in a way that represents a low response burden, thereby easily eliciting a response (ie, producing high engagement), but such content may not evoke the necessary cognitive reflection required to change behavior [18]. Alternatively, content that is designed to encourage deeper reflection may overwhelm users, which risks them disengaging completely. Understanding how to strike the balance between response and reflection has critical implications for effective engagement (ie, engagement needed to impact outcomes) [19].

To our knowledge, no studies have explored the association between reflection demands and the degree of interaction with mHealth tool. Understanding the interplay between reflection burden and response burden will help guide the design of interventions seeking to strike this balance.

Objective

Our team previously developed an mHealth intervention (delivered via text messages and phone calls) called Family/Friend Activation to Motivate Self-care (FAMS) [20,21]. FAMS is a diabetes self-management intervention that targets persons with type 2 diabetes and provides the option for persons with diabetes to invite a support person to also receive text messages. We recently evaluated FAMS in a randomized controlled trial (RCT) [22], and during routine monitoring in the first few weeks of the trial, we observed a low response rate to text messages among support persons. Because support persons’ engagement with the text messages was an optional component of the intervention, we determined that this was an opportunity to explore how changing the content of these texts might impact response rates without compromising our ability to evaluate FAMS’ effects. Over the course of the RCT [22], we used a pragmatic approach to vary both the reflection burden and the response burden of the two-way text messages sent to support persons assigned to the intervention. In this observational study, our primary goal was to explore how these variations would impact users’ engagement with text messages, as measured via response rates. We also explored support persons’ characteristics that were associated with response rates for each type of text and described the different responses to each text.

Methods

Study Design and Eligibility

This study was conducted as part of the FAMS 2.0 RCT. The trial design, intervention details, and outcomes for persons with diabetes and support persons were published [22-24]. For the trial, dyads comprising a person with diabetes and their support person were randomized to FAMS or a control condition. We recruited persons with diabetes who were receiving care for type 2 diabetes at Vanderbilt University Medical Center primary care clinics. Enrolling persons with diabetes were asked to invite a support person to participate with them and receive text messages; however, support person invitation and enrollment were not required. We defined a support person as any family member or friend with whom the person with diabetes would feel comfortable talking about diabetes management and health goals. Eligible support persons were aged ≥18 years, could speak and read English, and had a mobile phone separate from that of the person with diabetes. The only exclusion criterion was the inability to receive and respond to a text after training. For this study, we analyzed data from support persons in dyads that were randomly assigned to the intervention group (FAMS).

Ethical Considerations

The Vanderbilt University Institutional Review Board approved all study procedures (institutional review board number: 200398; approved April 8, 2020), and the trial was registered on ClinicalTrials.gov (trial number: NCT04347291).

Procedure

While enrolling persons with diabetes into the trial, research assistants collected contact information for a potential support person. A research assistant then contacted potential support persons to verify interest and eligibility, obtain verbal informed consent, and ask support persons to complete a baseline survey. Surveys were completed by phone with a research assistant, on the web via an emailed link, or via a mailed paper copy, per participants’ preferences. All survey data were stored in REDCap (Research Electronic Data Capture; Vanderbilt University), a secure web-based platform that supports data capture for research studies [25,26]. In addition to collecting data on sociodemographic characteristics, surveys asked support persons to choose the time of day they wanted to receive text messages.
messages. Relevant survey responses were transferred from REDCap to our technology partner, PerfectServe, using an automated application programming interface. PerfectServe used participant information to tailor, schedule, and send text messages to support persons for 9 months. Support persons could earn a total of US $120 for completing all study surveys (through 15 mo for the larger RCT). There was no compensation for receiving or responding to text messages.

The Intervention

Persons with diabetes received daily text message support and monthly coaching sessions, during which they set behavioral diabetes self-management goals (as detailed by Mayberry et al [22]). Support persons received text messages that were designed to increase dialogue about and facilitate their involvement in the diabetes self-management of the persons with diabetes; a one-way message was sent 3 to 4 times per week, and a two-way message (also known as an interactive text message) was sent once per week. One-way messages were either a general text message about providing diabetes self-management support or a text message tailored to the identified diabetes goals of the persons with diabetes. Two-way messages asked support persons about how they supported the health of the persons with diabetes. Support persons who replied to the two-way text received an automated response, thanking them for their answer.

Although an individual support person’s intervention experience lasted 9 months (36 wk), intervention delivery for the trial lasted 2 years. Over the course of those 2 years, we varied both the response burden (the level of information requested for the response relative to that of other messages) and the reflection burden (the level of cognitive reflection requested relative to that of other messages) of the weekly two-way text messages, which were sent to support persons in 6 fixed periods (ie, waves). The waves coincided with the weeks of the trial; they did not coincide with the weeks of each individual support person’s intervention experience. Figure 1 includes the content for each version of the text message, the weeks of the trial when each text was sent (ie, calendar time), and the respective waves. We started the trial (wave 1) by sending a text message that was high in both reflection burden and response burden (high/high). In wave 2, we tested a text that was low in both reflection burden and response burden (low/low), and then in wave 3, we tested a text that involved medium reflection burden and low response burden (medium/low). In wave 4, we retested the high/high message to help determine if the point at which the text was sent during the trial impacted engagement. In wave 5, we sought to delineate the relative impacts of reflection burden and response burden; therefore, we tested a text message that was high in reflection burden and low in response burden (high/low). Finally, we closed out the trial by retesting the low/low text message (wave 6). The decisions about what messages to test were made iteratively based on response rates to the prior message, with the goal of learning how much reflection we could request while still achieving a relatively high response rate.
Of note, each support person only received the versions of the two-way text message that were sent during their 36-week trial participation, with most (120/148, 81.1%) receiving 2 or 3 different versions and no participants receiving the same message in 2 separate waves. Because this analysis used data from support persons only, we refer to them as participants henceforth.

Measures

Sociodemographic and Relationship Characteristics
We collected self-reported data on age, gender, race, ethnicity, socioeconomic status (measured based on education [ie, years in school] and annual household income), and health literacy (assessed via the Brief Health Literacy Screen [27]). In addition, we asked whether participants were cohabitating with the persons with diabetes and the frequency with which they provided diabetes-specific helpful involvement to the persons with diabetes at baseline, as assessed via the Family and Friend Involvement in Adults’ Diabetes (FIAD) helpful subscale, support person version [28].

Engagement
We operationalized engagement by using response rate (ie, two-way messages responded to divided by the two-way messages sent, for each participant).
Analyses

Statistical Analysis Overview

All statistical analyses were performed by using R version 4.2.1 (R Foundation for Statistical Computing). We described participant characteristics via means and SDs or via frequencies and percentages, as appropriate. Except for when examining temporality, message waves that included the same version of the two-way text message were grouped together. Because this study was exploratory, we did not perform sample size calculations.

Overall Engagement by Text Message Version

For each version of the text message, we determined the proportion of two-way text messages sent to support persons that received a response by study week (ie, calendar time). We also generated summary statistics (means, medians, and first and third quartiles) for response rates at the participant level; reporting both mean and median provides more detailed information on the distribution of data. If participants withdrew during their intervention experience, we calculated their response rates based on the data available prior to their withdrawal. To account for repeated measures within participants, we used generalized estimating equations with a working independence correlation structure and a logistic link function to compare the odds of responding to two-way text messages across the four versions.

Participant Characteristics and Engagement

We used Poisson regression to estimate associations (as incidence rate ratios) between participant characteristics and text message response rates for each version of the text message. We included the number of two-way messages sent to a participant as an offset term in order to account for variation in the number of messages sent to each participant in a given wave; therefore, the exponentiated coefficients from the Poisson regression model compared response rates on a per-message basis. Participant characteristics included age, race and ethnicity (non-Hispanic White vs minoritized race or ethnicity), gender, education (years), annual household income ($ \geq $50,000 per year), health literacy (Brief Health Literacy Screen), whether the persons with diabetes and support persons were cohabitating, and self-reported baseline helpful involvement (FIAD). Further, we multiplied participants’ age by 10 to allow for easier interpretation of the results. Especially in regression models, it can be difficult to interpret the association between age and an outcome when the change in the outcome is based on a single-year change in age (ie, the coefficients end up being too small). Scaling the age variable in this way allowed us to interpret the findings in a more meaningful way, that is, we compared groups that differed in age by 1 decade rather than 1 year.

For this analysis, we excluded 5 participants who were missing all baseline data. However, missing covariate values were otherwise addressed via multiple imputation by chained equations (M=500 iterations).

Types of Responses

We characterized the responses to each version of the two-way text message. For the low/low, medium/low, and high/low texts, we reported the frequency of responses based on what the respective text requested (eg, “Yes,” “No,” “1,” “2,” “3,” “4,” or “5”). For the high/high text messages, 2 team members reviewed responses and categorized each as being either high effort or low effort. High-effort responses included comments on what went well that week, comments on what could go better next week, or both, and they referred to a diabetes self-management behavior such as diet, exercise, stress management, or communication (eg, “[He] and I got out several times this week walking after work. Our biggest problem is watching portion size when we are eating. Always continue to work on that.”). A low-effort response consisted of only a brief phrase that did not reference a diabetes self-management behavior (eg, “Things went well” and “We were on vacation this week”) or did mention a behavior but was unclear as to what went well or what could go better next week (eg, “Walking”).

Results

Participant Characteristics

In the trial, of the 150 support person participants who were enrolled and randomized to receive the FAMS intervention, 2 withdrew before the intervention started. The remaining 148 were included in the analyses (Table 1). The mean age was 50.3 (SD 14.7) years; 28.4% (42/148) of participants were men, and 33.1% (49/148) reported a minoritized racial or ethnic background. The mean length of education was 14.9 (SD 2.5) years; 28.4% (42/148) of participants were men, and 33.1% (49/148) of participants had annual household incomes of $ \leq $50,000. Over half (84/148, 56.8%) were spouses or partners of the persons with diabetes, and 70.3% (104/148) were cohabitating with the persons with diabetes. Further, 9 participants withdrew at some point during the intervention; the analyses below reflect their engagement during the time they participated.
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age$^a$ (y), mean (SD)</td>
<td>50.3 (14.7)</td>
</tr>
<tr>
<td>Gender $^b$ , n (%)</td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>42 (28.4)</td>
</tr>
<tr>
<td>Women</td>
<td>101 (68.2)</td>
</tr>
<tr>
<td>Race and ethnicity, n (%)</td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>92 (62.2)</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>29 (19.6)</td>
</tr>
<tr>
<td>Other non-Hispanic races</td>
<td>12 (8.1)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>8 (5.4)</td>
</tr>
<tr>
<td>Missing</td>
<td>7 (4.7)</td>
</tr>
<tr>
<td>Socioeconomic status</td>
<td></td>
</tr>
<tr>
<td>Education$^c$ (y), mean (SD)</td>
<td>14.9 (2.5)</td>
</tr>
<tr>
<td>Annual household income (US $), n (%)</td>
<td></td>
</tr>
<tr>
<td>&lt;35,000</td>
<td>25 (16.9)</td>
</tr>
<tr>
<td>35,000-49,999</td>
<td>21 (14.2)</td>
</tr>
<tr>
<td>50,000-74,999</td>
<td>22 (14.9)</td>
</tr>
<tr>
<td>75,000-99,999</td>
<td>24 (16.2)</td>
</tr>
<tr>
<td>$\geq$100,000</td>
<td>40 (27)</td>
</tr>
<tr>
<td>Missing or unknown</td>
<td>16 (10.8)</td>
</tr>
<tr>
<td>Health literacy (BHLS$^{d,e}$), mean (SD)</td>
<td>13.7 (1.5)</td>
</tr>
<tr>
<td>Relationship variables</td>
<td></td>
</tr>
<tr>
<td>Relationship type, n (%)</td>
<td></td>
</tr>
<tr>
<td>Spouse or partner</td>
<td>84 (56.8)</td>
</tr>
<tr>
<td>Parent</td>
<td>15 (10.1)</td>
</tr>
<tr>
<td>Son or daughter</td>
<td>22 (14.9)</td>
</tr>
<tr>
<td>Grandchild</td>
<td>7 (4.7)</td>
</tr>
<tr>
<td>Friend</td>
<td>11 (7.4)</td>
</tr>
<tr>
<td>Other</td>
<td>4 (2.7)</td>
</tr>
<tr>
<td>Missing</td>
<td>5 (3.3)</td>
</tr>
<tr>
<td>Helpful involvement (FIAD$^{f,g}$), mean (SD)</td>
<td>2.7 (0.9)</td>
</tr>
<tr>
<td>Cohabitating with person with diabetes$^h$, n (%)</td>
<td>104 (70.3)</td>
</tr>
</tbody>
</table>

---

Overall Engagement by Text Message Type

Figure 2 presents the proportion of two-way text messages that received a response within each week of the trial (ie, by calendar time). Notably, text message response rates for waves 1 and 4 (both were high/high message waves) were comparable, as were those for waves 2 and 6 (both were low/low message waves). Table 2 includes descriptive statistics for the overall and
participant-level response rates for each version of the two-way text. The median response rates for the high/high, medium/low, and low/low messages were 23% (IQR 0%-51%), 75% (IQR 38%-96%), and 90% (IQR 61%-100%), respectively. When we kept reflection burden high but lowered response burden in the high/low message, the median response rate (10%, IQR 0%-40%) was closest to that for the high/high message, suggesting that reflection burden was responsible for the lower response rates seen with the high/high message.

Figure 2. Text message response rates by week across each wave. Response data were excluded for the first 4 weeks and the last 5 weeks of the trial when <5 individuals were receiving the intervention.
Table 1. Response rates for each version of the two-way text message.

<table>
<thead>
<tr>
<th>Reflection burden and response burden</th>
<th>Participants included in analysis, n</th>
<th>Participant-specific response rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>%, mean (SD)</td>
</tr>
<tr>
<td>High and high</td>
<td>127</td>
<td>33 (33)</td>
</tr>
<tr>
<td>Low and low</td>
<td>123</td>
<td>74 (32)</td>
</tr>
<tr>
<td>Medium and low</td>
<td>125</td>
<td>65 (35)</td>
</tr>
<tr>
<td>High and low</td>
<td>55</td>
<td>26 (32)</td>
</tr>
</tbody>
</table>

We also compared the odds of responding to the four versions of the text message (Table 3). When compared to the high/high message, the odds of responding to the low/low message was 53% (95% CI 42%-65%) higher, the odds of responding to the medium/low message was 40% (95% CI 31%-49%) higher, and the odds of responding to the high/low message was 7.5% (95% CI 0.7%-14%) lower. All other pairwise comparisons (Table 3) indicated decreasing odds of responding at increasing levels of reflection burden.

Table 4. Comparison of text message response rates by text message version. Included are odds ratios (ORs) and 95% CIs, along with P values.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>OR (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative to high reflection burden and high response burden</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low reflection burden and low response burden</td>
<td>1.53 (1.42-1.65)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Medium reflection burden and low response burden</td>
<td>1.40 (1.31-1.49)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>High reflection burden and low response burden</td>
<td>0.93 (0.86-0.99)</td>
<td>.03</td>
</tr>
<tr>
<td>Relative to low reflection burden and low response burden</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium reflection burden and low response burden</td>
<td>0.92 (0.86-0.98)</td>
<td>.01</td>
</tr>
<tr>
<td>High reflection burden and low response burden</td>
<td>0.60 (0.55-0.66)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Relative to medium reflection burden and low response burden</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High reflection burden and low response burden</td>
<td>0.66 (0.61-0.72)</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Participant Characteristics Associated With Odds of Responding

Table 4 presents estimated incident rate ratios, along with 95% CIs and P values, from multivariate Poisson regression models that were used to identify participant characteristics predictive of response rate. Younger participants, participants who were not cohabitating with the persons with diabetes, and participants who had a lower socioeconomic status had lower response rates to both the high/high message and the medium/low message compared to those of older participants, participants who were cohabitating, or participants who had a higher socioeconomic status, respectively. The only characteristic associated with response rates for the low/low message was gender, such that men had lower response rates. Further, the only characteristic associated with response rates for the high/low message was age, such that younger age was associated with lower response rates. Across message versions, younger participants had lower response rates to any message with more burden than the low/low message, and participants who were not cohabitating with the persons with diabetes had lower response rates to the higher-burden messages than those of participants who were cohabitating. Race, ethnicity, health literacy, and baseline helpful involvement provided to the person with diabetes did not show patterns indicating the prediction of response rates to any message version.
Despite the potential of mHealth to enhance self-management support and quality of life, there are still key gaps in understanding how to optimize mHealth engagement [16,17,19]. Most engagement research reports only on system use without consideration of the cognitive reflection done in the process of engaging with the content [16,29]. Ideally, we want to encourage reflection that results in meaningful behavior change, but it is unclear how much we can request, with respect to reflection, before users disengage. We varied the reflection and response burdens of two-way text messages to examine how these variations impacted users’ engagement, as assessed via response rates. We found, generally, that as the reflection burden of the message increased, participants’ engagement decreased. Importantly, when the same version of the text was sent at different points in the trial, participants’ engagement was consistent, suggesting that the message itself was key for response rates. The response rates for the high/low message were similar to those for the high/high message, and this supports reflection burden (vs response burden) being the primary driver of lower engagement. We also found evidence that as the reflection burden of the message increased, there were more disparities in engagement across participant characteristics. This finding helps inform who we may lose with content requiring more from users and thus how to ensure mHealth interventions do not widen existing health disparities.

Research focused on promoting mHealth engagement has proliferated in recent years, with the primary goal of increasing system use [30]. Across such studies, we see evidence that content that is more personalized; uses simpler, nontechnical language; and is empowering results in higher engagement [10,19,31-35]. Less research has compared specific types of language; and is empowering results in higher engagement [30]. Across such studies, we see evidence that content that is more personalized; uses simpler, nontechnical language; and is empowering results in higher engagement [10,19,31-35].

Table. Participant characteristics predicting text message response rates for each version of the text. Presented are estimated incident rate ratios (IRRs) and 95% CIs, along with P values.a

<table>
<thead>
<tr>
<th>Predictor</th>
<th>High reflection/high response</th>
<th>Low reflection/low response</th>
<th>Medium reflection/low response</th>
<th>High reflection/low response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IRR (95% CI)</td>
<td>P value</td>
<td>IRR (95% CI)</td>
<td>P value</td>
</tr>
<tr>
<td>Age (y × 10)</td>
<td>1.14 (1.06-1.23)</td>
<td>.004&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.99 (0.95-1.03)</td>
<td>.67</td>
</tr>
<tr>
<td>Race and ethnicity</td>
<td>1.07 (0.85-1.35)</td>
<td>.57</td>
<td>0.96 (0.84-1.09)</td>
<td>.51</td>
</tr>
<tr>
<td>Gender (men)</td>
<td>0.62 (0.49-0.78)</td>
<td>&lt;.001&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.78 (0.68-0.89)</td>
<td>&lt;.001&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Education</td>
<td>0.98 (0.94-1.02)</td>
<td>.34</td>
<td>1.00 (0.98-1.03)</td>
<td>.77</td>
</tr>
<tr>
<td>Income</td>
<td>0.71 (0.57-0.89)</td>
<td>.003&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1.05 (0.91-1.21)</td>
<td>.51</td>
</tr>
<tr>
<td>BHLS&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1.04 (0.97-1.12)</td>
<td>.30</td>
<td>1.04 (1.00-1.09)</td>
<td>.06</td>
</tr>
<tr>
<td>Cohabiting</td>
<td>1.37 (1.09-1.73)</td>
<td>.007&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1.03 (0.90-1.19)</td>
<td>.64</td>
</tr>
<tr>
<td>FIAD&lt;sup&gt;d&lt;/sup&gt;</td>
<td>1.05 (0.94-1.17)</td>
<td>.41</td>
<td>1.00 (0.93-1.07)</td>
<td>.99</td>
</tr>
</tbody>
</table>

aA total of 5 support persons without baseline characteristics were excluded from this analysis: 2 were excluded from the models for the high/high, low/low, and medium/low messages; 1 was excluded from the models for the high/high and medium/low messages; 1 was excluded from the models for the high/high, low/low, and high/low messages; and 1 was excluded from the models for the low/low and medium/low messages.

<sup>b</sup>P<.05.

<sup>c</sup>BHLS: Brief Health Literacy Screen.

<sup>d</sup>FIAD: Family and Friend Involvement in Adults’ Diabetes.

Types of Responses

In this section, we report on engagement at the text message level (vs the participant level). For the high/high messages, 1429 texts were sent, and 445 responses were received. Further, 13 responses were excluded from the analysis because the content was uninterpretable or was not relevant to the two-way text prompt. The reviewers categorized each response into the high- or low-effort response category, with 98.6% (responses: 426/432) agreement; of the 426 texts agreed upon, 350 (82.2%) supports reflection burden (vs response burden) being the consistent, suggesting that the message itself was key for response rates. The response rates for the high/high message, and this message, and this supports reflection burden (vs response burden) being the primary driver of lower engagement. We also found evidence that as the reflection burden of the message increased, there were more disparities in engagement across participant characteristics. This finding helps inform who we may lose with content requiring more from users and thus how to ensure mHealth interventions do not widen existing health disparities.

Research focused on promoting mHealth engagement has proliferated in recent years, with the primary goal of increasing system use [30]. Across such studies, we see evidence that content that is more personalized; uses simpler, nontechnical language; and is empowering results in higher engagement [10,19,31-35]. Less research has compared specific types of...
content and has rarely tested different types within the same study. An exception is a recent study by Klimis et al [15], wherein they used machine learning to demonstrate that text messages with informative (providing health facts or education) and instructional (providing tips or recommendations) message intents were associated with increased engagement, while notification messages that addressed noneducational matters (eg, welcome and exit messages) were associated with reduced engagement [15]. Our study targeted a two-way message and varied the levels of reflection and response burdens in that message. By adjusting the reflection level specifically, we gained unique insight into how engagement with each message variation may ultimately influence behavior change [17,18].

The main way in which our study differs from others in this area of research is that we looked beyond system use as the sole dimension of mHealth engagement. Our goal was not necessarily to see which message resulted in the highest response rate but rather to determine how much reflection we could request from users and achieve a level of interaction that suggested that they were still invested in the content. Although our results show generally that engagement decreases with more reflection, the nuances in our findings allow us to provide unique recommendations around mHealth design. For instance, it may be best to alternate through content with different levels of reflection burden. Although users were more likely to respond to content that was lower in reflection burden, nearly all (350/426, 82.2%) of the responses that we received to the high-reflection messages included a high-effort level of reflection. The act of asking people to reflect stimulates internal thoughts that are difficult to measure without a response [36] but may still occur among some persons who do not respond.

Alternating content may help promote periodic responding and reflecting throughout an intervention experience. Another option involves using an adaptive intervention to tailor the content based on each person’s responsiveness. That is, everyone could start receiving content with a high reflection burden, but if a person’s response rate starts to drop, they could then switch to content with a moderate reflection burden. Finally, especially in situations where there is limited flexibility with the mHealth functionality, researchers may consider sending the medium/low message to all participants, given that the content encouraged a moderate level of reflection (more than the low/low message) yet still yielded a high response rate.

Limitations
Our study has several limitations to acknowledge. For instance, our results are based on an SMS text messaging intervention, which is a specific form of mHealth. It is possible that users would have responded differently if the content was delivered via an app or wearable technology. Importantly, compared to apps and other internet-dependent technologies, SMS text messaging is both lower in cost and more easily accessed, and it tends to have higher rates of engagement [4,37]. In addition, this study recruited persons with diabetes and their support persons from a specific region in Middle Tennessee. We acknowledge that the findings may not be generalizable to other types of individuals who are living in other locations. Relatedly, the content asked about how the support persons supported the health of the persons with diabetes, and engagement may differ when asking about a user’s own health; however, the marked differences in engagement across message types support broader applications. Another limitation of our work is that we restricted our assessment of engagement to a behavioral measure (ie, responding to the text) and did not have a way to assess participants’ cognitive investment or experience with each version of the text. Based on our analysis of participants’ responses to the high/high message, it appeared that responders were cognitively engaged, but we were not able to compare cognitive engagement across the other messages. In addition, the sample size for the high/low message analysis (n=55) was considerably smaller compared to those for the other message analyses, which was due to testing the high/low message during only 1 wave near the end of the trial when fewer participants were enrolled. Relatedly, we did not have the time or a sufficient number of participants toward the end of the trial to test the medium/low and high/low text messages in a second wave, and we do not know for certain whether engagement with these texts could be impacted by temporality; however, as engagement with the high/high and low/low texts remained similar across multiple waves, it is unlikely. The ordering of the text messages was variable across participants, and due to the observational nature of this study, we cannot determine the extent that ordering may have impacted results; however, the average response rates and trends across message versions and waves provide general insights on how these variations may impact engagement. Finally, we did not assess the impact of engagement on outcomes, as this fell outside the scope of our study; however, other studies in digital health have examined this association [38,39].

Conclusions
In order for individuals to benefit from mHealth and achieve desired effects on outcomes, engagement with the mHealth tool is needed. Our results help elucidate how truly complex the nature of engagement is. Although our past work and that of others have demonstrated the importance of behaviorally interacting with mHealth interventions [9,31,40,41], this measure represents one piece of a larger puzzle. Engagement may be best conceptualized as including both a behavioral dimension and a cognitive dimension. Balancing these dimensions may be what is ultimately needed to achieve effective engagement for impacting intended outcomes. Our study contributes to a growing body of research that encourages a more nuanced approach to studying engagement that goes beyond measuring system use. We hope that our findings help advance the field of mHealth and inform intervention design, with the goal of optimizing both engagement and effectiveness.

Acknowledgments
This research was funded by the National Institutes of Health (NIH), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), through grant R01-DK119282. The content is solely the responsibility of the authors and does not necessarily
represent the official views of the NIH. MKR is funded by the National Center for Advancing Translational Sciences (NIH; grant KL2TR002245).

Authors’ Contributions
LAN, AJS, LSM, RAG Jr, and MKR significantly contributed to the conception of the study and planned analyses. LAN wrote the manuscript. AJS analyzed the quantitative data, and RAG Jr oversaw data analyses. LML and SM analyzed responses to the high/high message. LSM was the principal investigator of the Family/Friend Activation to Motivate Self-care (FAMS) study. All authors were involved in data interpretation and manuscript revision and approved the final version submitted for publication.

Conflicts of Interest
None declared.

References
17. Yeager CM, Benight CC. If we build it, will they come? issues of engagement with digital health interventions for trauma recovery. Mhealth 2018 Sep 11;4:37. [doi: 10.21037/mhealth.2018.08.04] [Medline: 30363749]


**Abbreviations**

- **FAMS**: Family/Friend Activation to Motivate Self-care
- **FIAD**: Family and Friend Involvement in Adults' Diabetes
- **mHealth**: mobile health
- **RCT**: randomized controlled trial
- **REDCap**: Research Electronic Data Capture

© Lyndsay A Nelson, Andrew J Spieker, Lauren M LeStourgeon, Robert A Greevy Jr, Samuel Molli, McKenzie K Roddy, Lindsay S Mayberry. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 19.1.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
A Smartphone Food Record App Developed for the Dutch National Food Consumption Survey: Relative Validity Study

Marga Ocké1*, PhD; Ceciel Simone Dinnissen1*, MSc; Coline van den Bogaard1, BSc; Marja Beukers1, BSc; José Drijvers1, BSc; Eline Sanderman-Nawijn1, MSc; Caroline van Rossum1, PhD; Ido Toxopeus1, PhD

National Institute for Public Health and the Environment, Bilthoven, Netherlands

*these authors contributed equally

Corresponding Author:
Ceciel Simone Dinnissen, MSc
National Institute for Public Health and the Environment
Antonie van Leeuwenhoeklaan 9
Bilthoven, 3721 MA
Netherlands
Phone: 31 088 689 8989
Email: ceciel.dinnissen@rivm.nl

Abstract

Background: In the Dutch National Food Consumption Survey, dietary intake has been assessed since 2003 through 24-hour dietary recalls using the GloboDiet software. A new self-administered smartphone food record app called DitEetIk! was developed for potential use in future surveys.

Objective: This study aims to evaluate the data collected using the DitEetIk! app and its relative validity for food group, energy, and nutrient intake compared with the previous dietary assessment method (GloboDiet 24-hour dietary recalls).

Methods: A total of 300 participants aged 18 to 79 years were recruited from a consumer panel. Participants were asked to keep a record of their consumption using the DitEetIk! app on 3 nonconsecutive days. Trained dietitians conducted a 24-hour dietary recall interview by telephone using the GloboDiet software (International Agency for Research on Cancer) regarding 1 of 3 DitEetIk! recording days. Nutrient intake was calculated using the NEVO database (version 2021/7.0). Relative validity was studied by comparing data from GloboDiet 24-hour dietary recalls and the DitEetIk app for the same day. Participants with implausible records, defined as days with energy intake of <0.6 or >3.0 basal metabolic rate, were excluded from the analyses. For 19 food groups and 29 nutrients, differences in median intake were assessed using the Wilcoxon signed rank test, and Spearman correlation coefficients were calculated. Bland-Altman plots with mean differences and 95% limits of agreement were created for energy intake and the contribution to energy intake from fat, carbohydrates, and protein.

Results: A total of 227 participants completed a combination of a DitEetIk! app recording day and a 24-hour dietary recall interview for the same day. Of this group, 211 participants (n=104, 49.3% men and n=107, 50.7% women) had plausible recording days. Of all recorded food items, 12.8% (114/894) were entered via food barcode scanning, and 18.9% (169/894) were searched at the brand level. For 31% (5/16) of the food groups, the median intake assessed using the DitEetIk! app was >10% lower than that assessed using 24-hour dietary recalls; this was the case for fruit (P=.005), added fats (P=.001), milk and milk products (P=.02), cereal products (P=.01), and sauces (P<.001). This was also the case for 14% (4/29) of the nutrients (all P<.001). Regarding mean intake, differences were generally smaller. Regarding energy intake, the mean difference and 95% limits of agreement were 14 kcal (−1096 to 1124). Spearman correlation coefficients between intake assessed using the DitEetIk! app and 24-hour dietary recalls ranged from 0.48 to 0.88 (median 0.78) for food groups and from 0.58 to 0.90 (median 0.72) for nutrients.

Conclusions: Compared with GloboDiet 24-hour dietary recalls, the DitEetIk! app assessed similar mean energy intake levels but somewhat lower median intake levels for several food groups and nutrients.

(JMIR Mhealth Uhealth 2024;12:e50196) doi:10.2196/50196

KEYWORDS
relative validity; smartphone food record; 24-hour dietary recall; mobile app; national food consumption surveys; smartphone; food; food consumption; app; diet; dietary intake; nutrients; survey; mobile phone
## Introduction

### Background

Many countries conduct national food consumption surveys as these are considered important instruments for prioritizing, developing, and evaluating food policies [1,2]. Food consumption survey data can be used to assess adherence to food-based dietary guidelines and obtain insight into the food consumption patterns of a population. After combination with other data sources such as food composition databases, food contamination occurrence data, and life cycle assessment data, the nutritional adequacy of the diet, dietary exposure assessment to potentially hazardous substances, and the environmental impact of dietary patterns can be assessed [3-5].

The use of national food consumption survey data for multiple purposes requires dietary assessment methods that allow all consumed food items to be reported with detailed characterizations. On the basis of European projects such as the European Food Consumption Survey Method [6] and European Food Consumption Validation [7], the 2014 guidance on the European Union Menu methodology by the European Food Safety Authority prescribes the use of food records for children and 24-hour dietary recalls for adults as dietary assessment methods in European national food consumption surveys [8].

It is advised that trained personnel is employed to administer the recall interviews or conduct a food record completion interview with the participants [8,9]. The requirements of trained personnel and detailed food descriptions, the large number of products, and continuously changing food items on the market, and the need to handle all possible reported food items make food consumption surveys costly [10]. Moreover, these requirements also pose a burden on the survey participants. It has been suggested that this burden has led to declining and possibly selective participation rates in national food consumption surveys [1].

In the past decades, various tools for self-administered 24-hour dietary recalls or food records have become available [11]. These digital tools have the potential to be less resource intensive. Many users prefer these applications over the traditional methods as they can be used where and when it is convenient [12]. Information and communications technology–based applications also enable the use of user-friendly support functionalities that were not feasible with interviewer-based methods. For example, food package barcode scanning using the camera function of a smartphone [11] combined with a comprehensive branded food database reduces the time burden of searching for a product through a long list of food items. However, without the help of a trained interviewer, it might be challenging for participants to report all food items consumed and describe and quantify them accurately. These developments warrant further exploration of whether interviewer-based dietary assessments in national food consumption surveys can be replaced with self-administered dietary assessments using digital food record applications.

In the Dutch National Food Consumption Survey (DNFCS), dietary intake has been assessed through 24-hour dietary recalls by trained interviewers using the GloboDiet software since 2003 [13-16]. A new food record app for dietary assessment called DitEetIk!, which uses self-administration, was developed for potential use in future national dietary surveys in the Netherlands. A smartphone food record was chosen over a self-administered digital 24-hour dietary recall because of the availability of a branded food database in the Netherlands [17]. Such a database allows for specific food identification and can be used most optimally when keeping a food diary throughout the day on a mobile phone with a camera function for barcode scanning. Moreover, the level of smartphone ownership and use in the Netherlands is high. In 2019, smartphones were present in 89% of Dutch households [18].

### Objectives

To assess the suitability of the DitEetIk! app for future surveys, it is important to evaluate its quality and comparative validity against the method currently used in the DNFCS. Therefore, the aim of this study was to evaluate the level of detail regarding the food description obtained in the reported consumption in the DitEetIk! app and determine how well the DitEetIk! app is able to assess the daily intake of food groups, energy, and nutrients in comparison with dietitian-administered 24-hour dietary recalls using the GloboDiet software (International Agency for Research on Cancer) in adults. This study focused on systematic differences at the food group and nutrient levels, with random error being of secondary interest. To study the potential effects of the study design on the 24-hour dietary recall data, we also compared the GloboDiet 24-hour dietary recall data in this study with those of a matched population of the DNFCS 2019 to 2021 [19].

### Methods

#### Recruitment

The intention was to collect data from 200 participants with sufficient variation in gender, age group, and educational level. To account for potential dropouts and invalid food recording data, 300 participants were recruited. Participants from a consumer panel of Kantar Netherlands were invited via email to take part in the study. Information regarding the privacy policy of the DitEetIk! app was provided. The sociodemographic characteristics of the panel members were known. Potential participants were eligible if they were aged between 18 and 79 years; were not institutionalized; did not participate in the DNFCS 2019 to 2021 [19] or the Eetmeter study [20]; did not use tube or parental feeding; and were able to use the DitEetIk! app on their smartphone, which had to run on the Android operating system version 7 or higher.

#### Ethical Considerations

The Medical Research Ethics Committee of Utrecht University evaluated that the study was not subject to the Medical Research Involving Human Subjects Act of the Netherlands (dossier 21/686). All study participants provided written informed consent. After completion of the study, participants received an incentive bonus (NIPOints to be exchanged for a gift card or coupon).
DitEetIk! App

Objective of the DitEetIk! App

In the Dutch language, “Dit eet ik” means “This is what I eat.” The DitEetIk! app was developed specifically for the objectives of the DNFCS. The food description had to be specific enough to provide insights into the intake of nutrients, the exposure to chemicals relevant from a food safety point of view, and the environmental impact of the diet of the Dutch general population aged 1 to 79 years.

Development Process

The DitEetIk! app was developed using an agile approach. Various usability tests and focus group sessions were part of this process. The development period lasted approximately 3 years, with a team of app builders, data managers, dietitians, and nutritionists. Technical development was conducted by Dienst ICT Uitvoering in collaboration with National Institute for Public Health and the Environment for functional development and formative evaluation. Both are Dutch governmental organizations. Safety and General Data Protection Regulation issues were considered in the app development. The DitEetIk! app does not collect information that makes the participant identifiable. The user interface of the app is in Dutch (B1 level).

App Availability, Registration, and Instruction

Version 1.0 of the DitEetIk! app was developed for Android smartphones and is available on the Google Play store. Using the DitEetIk! app, a person can keep a food record for specific days upon invitation (ie, the DitEetIk! app can be used only after entering a participant number with matching gender and age). An instruction movie can be viewed at any time after registration. Moreover, to support participants, context-specific information and relevant sections of the instruction movie are available on each screen. Participants were instructed to record all food and drinks consumed from getting up one day to getting up the next day.

Push Notifications and Feedback

At several moments—before, during, and after the specified day—push notification messages are sent via the DitEetIk! app to the participants to help remind them of food recording and submission of the food record. The DitEetIk! app does not provide instant feedback to the participants regarding their food consumption as, for dietary monitoring, it is important not to influence the participants.

Main Menu

Food recording is performed via a main menu where 4 eating occasions can be selected (ie, breakfast, midday meal, evening meal, and in between meals). If applicable, dietary supplements can be filled out separately (Figure 1A). After choosing breakfast, midday meal, or evening meal as eating occasions, the time and place of consumption need to be registered; for the occasion “in between meals,” time and place of consumption (eg, home or restaurant) are asked every time a food item is selected, whereas for dietary supplements, information regarding time and place of consumption is not asked for.

Figure 1. Screenshots of the DitEetIk! app depicting (A) eating occasions, (B) recently added products, and (C) food-specific follow-up questions.
Food Recording

The DitEetIk! app food list in this study included 140,781 food items, of which 3432 (2.44%) were generic and the rest were a selection of branded food items from the Dutch-branded food database Dutch-branded food database (Levensmiddelendatabank [LEDA]; download date: October 27, 2021) [17]. The selection consisted of 70.89% (137,349/193,742) of the branded food items in the LEDA database that were matched to a generic product. Generic food items can be found through text searching. Branded food items can be found by scanning the barcode or, for predefined food groups, by selecting the brand and specific branded product once a generic food item is chosen. If the scanned food items are not included in the LEDA database or are not matched to generic products, they are not recognized in the app, and the participant has to record their food via text search. Once a product is selected and added to an eating occasion, the food item can also be found via “recently added products” and can be selected again (Figure 1B). A recipe feature is available it allows participants to create mixed dishes indicating quantities of foods that were used for the prepared dish as a whole or save frequently consumed combinations of foods. Commonly used mixed dishes are also available as prepared generic food items.

Follow-Up Questions for Each Food Item

After a food item is chosen or scanned, food-specific follow-up questions are asked regarding the preparation method, consumption with or without skin or peel, and quantity consumed (Figure 1C). In case a preparation method with fat (eg, frying or deep-frying) is chosen, the participant is asked to specify the type of fat used. The follow-up questions are defined at the generic food item level. For this reason, each branded food item is linked to a comparable generic food item. Consumed amounts can be indicated via various options that always include the weight in grams or volume in milliliters and often the number or fraction of household measures, natural units, or commercial units. For user-defined recipes, the fraction of the total recipe can be indicated as the consumed portion.

Data Submission and Output

When all consumptions for a day have been recorded, the participant can submit the food record. Upon submission, questions regarding completeness are asked, as well as whether the day was special or not regarding consumption. The data collected using the DitEetIk! app can be downloaded as CSV files that include information on eating occasions with time, place, and registered food items and the answers to the follow-up questions for each consumption.

Study Design and Data Collection

Data were collected in the spring of 2022. Participants were asked to record all food, drinks, and dietary supplements consumed on 3 nonconsecutive days on the DitEetIk! app. The days were assigned by the DitEetIk! app in such a way that all days of the week were covered proportionally at the group level and that there were at least 7 days between each recording day. The day before each registration day, the participants received a push notification message on their smartphones. At the individual level, any combination of days could occur. In case the participant indicated that the day was not convenient for recording or did not record any food items, a replacement recording day was assigned automatically. This could be done maximally 3 times; thereafter, no new days were assigned, and the participant was excluded from further participation.

After 1 of the 3 DitEetIk! app recording days, participants were contacted by a trained dietician for a 24-hour dietary recall interview by telephone. For the 24-hour dietary recall interviews, it was allowed to make an appointment with the participant; for practical reasons, they were not unannounced. The 24-hour dietary recalls were administered using the Dutch version of the GloboDiet software (version 2021-09-24). This software, which was previously called EPIC-Soft, has been described in detail elsewhere [16]. Briefly, the interview started by composing a quick list in which the participant was asked to roughly list all consumed food and drinks for 7 potential eating occasions with time and place of consumption. Food items were recalled starting from getting up in the morning until getting up the following day. In the second step, the interviewer specified each food item on the quick list using a series of follow-up questions applicable to that food item, for example, asking about preparation methods and, if relevant, the type of fat used. Mixed dishes could be entered as new individual recipes or as (adjustments of) standard recipes, which the software disaggregated into ingredients. Dietary supplements were explicitly asked about. Consumed amounts of the food items could be quantified in several ways: by means of quantities as shown in photos in a picture booklet with a series of 61 food photographs, or in household measures, units, and standard portions; by weight or volume; and by the proportion of a total recipe. Bread shapes were used to estimate the quantity of spreads. At various points, quality control of the data was incorporated into the GloboDiet software, for example, checks on missing quantities, probing questions on often forgotten food items, and checks on total intake of energy and macronutrients.

After the 3 food records were completed, the perceived usability of the DitEetIk! app was evaluated by the participants through a web-based questionnaire, the System Usability Scale (SUS) [21]. This is a widely used questionnaire for the evaluation of electronic devices and systems, including smartphone apps. It consists of 10 statements with response options on a 5-point Likert scale ranging from strongly disagree (1) to strongly agree (5). There are 5 positive statements alternated with 5 negative statements. The originally English-worded items were translated into Dutch, and the word “system” was replaced with “application” to make the questions more specific to the device.

Data Handling

For the DitEetIk! app data, intake per food item per day was calculated by multiplying the weight of the chosen portion or serving by the number of portions. If applicable, food density (in the case of estimates in household measures or milliliters), an edible fraction (in the case of the inedible part), weight change because of food preparation (amount estimated as unprepared food), and percentage of fat absorption were applied to the calculation of the amount of food in grams per day in its consumed state. Similar calculations were performed using the GloboDiet software for the 24-hour dietary recalls [16].
food items reported in the GloboDiet 24-hour dietary recalls and the DitEetIk! app were categorized into the food groups mentioned in the Wheel of Five Dutch food-based dietary guidelines [22,23]. Subsequently, consumption of food groups per person per day was calculated.

For both GloboDiet and DitEetIk! data, intake of energy and nutrients per person per day was calculated by multiplying the consumed amount of food by the nutrient level per gram of food and adding the nutrient intake of all food items consumed in a day. Information on food composition was obtained from the NEVO database (version 2021/7.0) [24]. Dietary supplements were not considered in the calculation of nutrient intake.

All extremely high values in energy, nutrient, and food group intake in the 24-hour dietary recalls were evaluated using the same methodology as in the DNFCS [15]. The food items that contributed the most to the high intake values were checked for obvious errors. None of the extreme intake values were judged as unlikely.

The equations developed by Henry [25] were applied to calculate the estimated basal metabolic rate (BMR) using height and weight information provided by the participants in the DitEetIk! app. The average ratio of energy intake to BMR was calculated, as well as the percentage of extreme energy reporters (ie, those participants with a ratio of <0.6 or >3.0). Any day with such an implausible extreme of energy intake in the DitEetIk! app was excluded from further analyses.

Assessment of the Effect of Study Design on the GloboDiet 24-Hour Dietary Recall Results

The GloboDiet 24-hour dietary recall interview was always conducted after the recording in the DitEetIk! app. This might have influenced the results because of potential memory or learning effects [26]. The extent to which this occurred was estimated by comparing the results of the GloboDiet 24-hour dietary recalls in this study with the findings among participants in the DNFCS 2019 to 2021, in which 24-hour dietary recalls were collected using the same software. We only used the first GloboDiet 24-hour dietary recall interviews of these participants. Each participant in the DitEetIk! app evaluation study was matched with a participant in the DNFCS 2019 to 2021 based on characteristics associated with dietary intake, that is, age (5-year classes), gender, weight (10-kg classes), and educational level (3 classes). This provided matches for 86.3% (182/211) of the participants. For those without a match, the matching characteristics were relaxed, starting with educational level (28 matches) followed by age (1 match). For participants in the evaluation study with multiple possible matches in the DNFCS 2019 to 2021, a person with the same or the closest height was chosen (182 matches).

Statistical Analysis

For the analyses, only DitEetIk! data for the day that was recalled using the GloboDiet software were used. Frequency analyses were conducted to describe the population of DitEetIk! app users and the matched participants of the DNFCS 2019 to 2021 in terms of sociodemographic characteristics. For the items of the SUS and the total SUS score, means and SDs were calculated.

The medians, SDs, and IQRs of the food group, energy, and nutrient consumption assessed using both methods were calculated. Owing to skewed distributions, the nonparametric Wilcoxon signed rank test was used to test whether differences between the DitEetIk! app and the 24-hour dietary recalls for food group and nutrient intake were symmetrical around 0. Differences were considered relevant if the median intake was >10% of the 24-hour dietary recall value. Only food groups for which the 75th percentile was >0 were reported. The analyses were repeatedly stratified by educational level (3 classes) and BMI (3 classes).

In addition, the number of consumers and median intake for consumers only were calculated for each food group for each method. The McNemar test was used to test whether being a consumer per food group differed significantly between both methods. In addition, the Wilcoxon signed rank test was used to test whether the distribution of food group consumption only differed systematically by method. Spearman rank correlation coefficients were calculated for food group consumption and also energy and nutrient intake assessed using both methods.

Bland-Altman plots were constructed for the intake of energy and energy percentage derived from fat, protein, and carbohydrates, plotting the difference in intake assessed using both methods against the mean intake for each participant. The derived 95% limits of agreement [27] were presented to provide information on the variation in individual relative validity.

To assess any design effects, a nonparametric Wilcoxon signed rank test was used to test whether there was a systematic difference between the 24-hour dietary recalls in the relative validity study and the DNFCS 2019 to 2021 regarding the intake of energy, nutrients, and food groups.

All statistical analyses were conducted using SAS (version 9.4; SAS Institute Inc). P values of <.05 were considered statistically significant, and 2-sided statistical tests were performed.

Results

Study Population Characteristics

Of the 3418 contacted people, 443 (12.96%) were willing to participate (Figure 2). A total of 300 people were invited to start the study based on their ability to use an Android smartphone and their sociodemographic characteristics. Of the 300 invited individuals, 227 (75.7%) completed a combination of a DitEetIk! app recording day and a 24-hour dietary recall interview for the same day. In total, 7% (16/227) of the participants were excluded because of an implausible ratio of energy intake to BMR of <0.6 or >3.0, resulting in 211 participants included in the analyses.

Approximately half (107/211, 50.7%) of the participants were women (Table 1). The study population consisted of more participants with a middle and higher educational level (81/211, 38.4% and 93/211, 44.1%, respectively) than those with a lower educational level (37/211, 17.5%). Of the 211 participants, 129 (61.1%) had a BMI of ≥25 kg/m². Fewer people in the highest age category of 60 to 79 years participated in the study (58/211, 27.5%) than people in the 2 lower age categories (76/211, 36%...
and 77/211, 36.5%). The 211 participants included in the DitEetIk! app study were matched with 211 participants of the DNFCs 2019 to 2021. They had similar distributions in terms of gender, age, educational level, and BMI.

Of the 211 participants, 207 (98.1%) completed 3 DitEetIk! app recording days. Most GloboDiet 24-hour dietary recall interviews were for the first or second DitEetIk! app recording day (78/211, 37% and 84/211, 39.8%, respectively). Most of the participants started recording their food consumption on the DitEetIk! app in the morning (105/211, 49.8%) or afternoon (61/211, 28.9%) and ended their recording the same day in the evening (95/211, 45%) or the next day in the morning (82/211, 38.9%). The time between starting and ending the food recording using the DitEetIk! app was between 8 and 20 hours for 46% (97/211) of the participants, whereas it took >20 hours for 33.2% (70/211) of the participants and <8 hours for 20.9% (44/211) of the participants (Multimedia Appendix 1).

Food identification for recording was performed via text searching of the food items at the generic level for 23.6% (211/894) of the products and choosing previously selected food items (175/894, 19.6%). Branded food items were recorded via text searching (169/894, 18.9%) and barcode scanning (114/894, 12.8%). The option to make an individual recipe was used for 1.8% (16/894) of the products, associated food items (fats and oils for frying and milk or sugar in tea or coffee) were used for 16.4% (147/894) of the products, and supplements were used for 6.9% (62/894) of the products. Without considering previously selected food items, branded food items were recorded 39.4% (283/719) of times (Multimedia Appendix 2).

The SUS questionnaire was used to evaluate the system usability of the DitEetIk! app. The questionnaire was completed by 98.1% (207/211) of the included participants. The mean score per question ranged from 1.7 (SD 0.8) to 3.7 (SD 0.9). The mean overall score of the SUS was 66.6 (SD 15.1).

**Figure 2.** Participant recruitment and selection flow. A plausible recording day was a recording day with a ratio of energy intake over BMR between 0.6 and 3.0. BMR: basal metabolic rate; EI: energy intake.
Table 1. Characteristics of the 211 participants in the DitEetIk! app study and matched participants from the Dutch National Food Consumption Survey (DNFCS) 2019 to 2021.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Evaluation study (n=211), n (%)</th>
<th>DNFCS 2019-2021 (n=211), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>104 (49.3)</td>
<td>104 (49.3)</td>
</tr>
<tr>
<td>Women</td>
<td>107 (50.7)</td>
<td>107 (50.7)</td>
</tr>
<tr>
<td><strong>Age category (y)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-39</td>
<td>76 (36)</td>
<td>75 (35.5)</td>
</tr>
<tr>
<td>40-59</td>
<td>77 (36.5)</td>
<td>78 (37)</td>
</tr>
<tr>
<td>60-79</td>
<td>58 (27.5)</td>
<td>58 (27.5)</td>
</tr>
<tr>
<td><strong>Highest educational level attained</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lowa</td>
<td>37 (17.5)</td>
<td>31 (14.7)</td>
</tr>
<tr>
<td>Middleb</td>
<td>81 (38.4)</td>
<td>87 (41.2)</td>
</tr>
<tr>
<td>Highc</td>
<td>93 (44.1)</td>
<td>93 (44.1)</td>
</tr>
<tr>
<td><strong>BMI category (kg/m^2)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;18.5</td>
<td>1 (0.5)</td>
<td>2 (0.9)</td>
</tr>
<tr>
<td>≥18.5 to ≤25</td>
<td>81 (38.4)</td>
<td>81 (38.4)</td>
</tr>
<tr>
<td>≥25 to ≤30</td>
<td>73 (34.6)</td>
<td>73 (34.6)</td>
</tr>
<tr>
<td>≥30</td>
<td>56 (26.5)</td>
<td>55 (26.1)</td>
</tr>
</tbody>
</table>

*a*Low educational level: primary education, lower vocational education, or advanced elementary education.

*b*Middle educational level: intermediate vocational education or higher secondary education.

*c*High educational level: higher vocational education and university.

**Relative Validity for Food Groups**

For 44% (7/16) of the food groups, no statistically significant differences between the median intake of food groups assessed using the DitEetIk! app food record and the 24-hour dietary recall were observed (Table 2). No statistically significant differences were observed for the median consumption of vegetables ($P$=.13), meat ($P$=.10), eggs ($P$=.44), nuts ($P$=.73), bread ($P$=.95), potatoes ($P$=.96), and snacks ($P$=.41).

Statistically significant differences of $>$10% between the 2 methods were observed for the median consumption of fruit (83 vs 130 g/d; $P$=.005), added fats (12 vs 17 g/d; $P$=.001), milk and milk products (219 vs 252 g/d; $P$=.02), cereal products (6 vs 20 g/d; $P$=.01), and soups (2 vs 22 g/d; $P$<.001). For all food groups except “Other,” the DitEetIk! app assessed a lower median consumption and, therefore, relatively underestimated the consumption of food groups compared with the 24-hour dietary recalls. However, when looking at the mean rather than the median, for 7 food groups, higher values were observed in the DitEetIk! app. Spearman correlation coefficients ranged from 0.50 for the food group “Other” to 0.88 for the food groups Potatoes, Sandwich spreads, and Snacks. The median correlation coefficient was 0.78.

A significant difference in the number of consumers between the 2 methods was observed for the food groups Fruit ($P$=.002), Added fats ($P$=.02), Cereal products ($P$=.03), Sauces ($P$<.001), and Other products ($P$<.001; Table 3). This list includes 4 of the 5 food groups for which differences in median intake for all participants were statistically significant and of $>$10%. Focusing on the 5 food groups, the median intake of consumers differed significantly for Added fats and Sauces (median 15 and 24 grams per day in the DitEetIk! app vs 20 and 45 grams per day in the GloboDiet 24-hour dietary recalls; $P$=.06 and $P$=.004) but not for Fruit, Milk and milk products, and Cereal products, whereas for the food groups Fish (90 vs 115 g/d), Cheese (36 vs 45 g/d), Sandwich spreads (26 vs 20 g/d), and Soups (188 vs 50 g/d), significant differences of $>$10% were also observed in the median intake of consumers only. Within the strata of educational level (3 classes), a few significant differences that were $\geq$10% for the median were observed for food groups (Multimedia Appendix 1). For soups (low educational level $P$=.02; middle educational level $P$=.006; high educational level $P$<.001), the differences were consistent across the 3 educational levels. For milk and milk products, only those with a high educational level had a significantly lower median in the DitEetIk! app compared with the GloboDiet 24-hour dietary recalls ($P$=.01). Within the strata of BMI (3 levels), more differences were observed, but there was no consistent pattern in which differences were generally smaller or larger for persons in one of the BMI classes (Multimedia Appendix 2).
Table 2. The mean, SD, median, and IQR of consumption of food groups\(^a\) as assessed using the DitEetIk! app and 24-hour dietary recalls for the same day and their correlation for the 211 participants with plausible energy intakes.

<table>
<thead>
<tr>
<th>Food group</th>
<th>DitEetIk! app food record (g/d)</th>
<th>GloboDiet 24-hour dietary recall (g/d)</th>
<th>Wilcoxon signed rank test ( P ) value(^b)</th>
<th>Spearman correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Values, mean (SD)</td>
<td>Values, median (IQR)</td>
<td>Values, mean (SD)</td>
<td>Values, median (IQR)</td>
</tr>
<tr>
<td>Vegetables</td>
<td>163 (200)</td>
<td>117 (31-226)</td>
<td>160 (144)</td>
<td>130 (50-240)</td>
</tr>
<tr>
<td>Fruit</td>
<td>128 (186)</td>
<td>83 (0-188)</td>
<td>140 (146)</td>
<td>130 (0-217)</td>
</tr>
<tr>
<td>Added fats</td>
<td>16 (16)</td>
<td>12 (3-24)</td>
<td>19 (15)</td>
<td>17 (6-29)</td>
</tr>
<tr>
<td>Meat</td>
<td>12 (35)</td>
<td>0 (0-0)</td>
<td>17 (57)</td>
<td>0 (0-0)</td>
</tr>
<tr>
<td>Eggs</td>
<td>2 (18)</td>
<td>0 (0-0)</td>
<td>4 (20)</td>
<td>0 (0-0)</td>
</tr>
<tr>
<td>Nuts</td>
<td>103 (112)</td>
<td>73 (23-135)</td>
<td>92 (83)</td>
<td>75 (33-120)</td>
</tr>
<tr>
<td>Milk (products)</td>
<td>17 (37)</td>
<td>0 (0-0)</td>
<td>17 (34)</td>
<td>0 (0-13)</td>
</tr>
<tr>
<td>Cheese</td>
<td>15 (30)</td>
<td>0 (0-0)</td>
<td>15 (30)</td>
<td>0 (0-20)</td>
</tr>
<tr>
<td>Bread</td>
<td>264 (263)</td>
<td>219 (16-391)</td>
<td>288 (248)</td>
<td>252 (80-423)</td>
</tr>
<tr>
<td>Cereal products</td>
<td>33 (36)</td>
<td>30 (0-56)</td>
<td>39 (44)</td>
<td>31 (0-62)</td>
</tr>
<tr>
<td>Potatoes</td>
<td>146 (113)</td>
<td>126 (70-199)</td>
<td>138 (88)</td>
<td>132 (70-180)</td>
</tr>
<tr>
<td>Drinks</td>
<td>67 (133)</td>
<td>6 (0-88)</td>
<td>74 (106)</td>
<td>20 (0-119)</td>
</tr>
<tr>
<td>Sandwich spreads</td>
<td>72 (119)</td>
<td>0 (0-128)</td>
<td>66 (104)</td>
<td>0 (0-120)</td>
</tr>
<tr>
<td>Snacks</td>
<td>1888 (956)</td>
<td>1836 (1275-2311)</td>
<td>2097 (889)</td>
<td>1963 (1582-2539)</td>
</tr>
<tr>
<td>Sauces</td>
<td>15 (27)</td>
<td>0 (0-20)</td>
<td>12 (23)</td>
<td>0 (0-15)</td>
</tr>
<tr>
<td>Other</td>
<td>31 (91)</td>
<td>0 (0-0)</td>
<td>12 (46)</td>
<td>0 (0-0)</td>
</tr>
</tbody>
</table>

\(^a\)Food groups are Wheel of Five food groups—main groups [23]. The food groups Fish, Legumes, and Soups were excluded as the 75th percentile was 0 for both methods. Table 3 provides more information on these food groups.

\(^b\)Wilcoxon signed rank test (normal approximation) of the differences between intake assessed using the DitEetIk! app and the GloboDiet 24-hour dietary recalls for the same day.
Table 3. Number of consumers of a food group and median of consumed amount of a food group for consumers only as assessed using the DitEetIk! app and the 24-hour dietary recalls (n=211).

<table>
<thead>
<tr>
<th>Food group</th>
<th>DitEetIk! app food record (g/d)</th>
<th>GloboDiet 24-hour dietary recall (g/d)</th>
<th>McNemar test ( P ) value(^b)</th>
<th>Wilcoxon signed rank test ( P ) value(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Consumers (n=211), n (%)</td>
<td>Consumers, median</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vegetables</td>
<td>175 (82.9)</td>
<td>146</td>
<td>.16</td>
<td>.23</td>
</tr>
<tr>
<td>Fruit</td>
<td>134 (63.5)</td>
<td>156</td>
<td>.002</td>
<td>.21</td>
</tr>
<tr>
<td>Added fats</td>
<td>173 (82)</td>
<td>15</td>
<td>.02</td>
<td>.06</td>
</tr>
<tr>
<td>Fish</td>
<td>25 (11.8)</td>
<td>20</td>
<td>.66</td>
<td>.04</td>
</tr>
<tr>
<td>Legumes</td>
<td>5 (2.4)</td>
<td>10 (4.7)</td>
<td>.06</td>
<td>.38</td>
</tr>
<tr>
<td>Meat</td>
<td>178 (84.4)</td>
<td>179 (84.8)</td>
<td>.81</td>
<td>.06</td>
</tr>
<tr>
<td>Eggs</td>
<td>52 (24.6)</td>
<td>61 (28.9)</td>
<td>.08</td>
<td>.83</td>
</tr>
<tr>
<td>Nuts</td>
<td>66 (31.3)</td>
<td>70 (33.2)</td>
<td>.32</td>
<td>.29</td>
</tr>
<tr>
<td>Milk and milk products</td>
<td>165 (78.2)</td>
<td>173 (82)</td>
<td>.06</td>
<td>.13</td>
</tr>
<tr>
<td>Cheese</td>
<td>138 (65.4)</td>
<td>145 (68.7)</td>
<td>.11</td>
<td>.03</td>
</tr>
<tr>
<td>Bread</td>
<td>194 (91.9)</td>
<td>198 (93.8)</td>
<td>.10</td>
<td>.71</td>
</tr>
<tr>
<td>Cereal products</td>
<td>113 (53.6)</td>
<td>124 (58.8)</td>
<td>.03</td>
<td>.05</td>
</tr>
<tr>
<td>Potatoes</td>
<td>87 (41.2)</td>
<td>90 (42.7)</td>
<td>.44</td>
<td>.91</td>
</tr>
<tr>
<td>Drinks</td>
<td>206 (97.6)</td>
<td>211 (100)</td>
<td>N/A(^d)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Sandwich spreads</td>
<td>85 (40.3)</td>
<td>89 (42.2)</td>
<td>.25</td>
<td>.01</td>
</tr>
<tr>
<td>Soups</td>
<td>33 (15.6)</td>
<td>28 (13.3)</td>
<td>.10</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Snacks</td>
<td>173 (82)</td>
<td>172 (81.5)</td>
<td>.76</td>
<td>.69</td>
</tr>
<tr>
<td>Sauces</td>
<td>116 (55)</td>
<td>143 (67.8)</td>
<td>&lt;.001</td>
<td>.004</td>
</tr>
<tr>
<td>Other</td>
<td>126 (59.7)</td>
<td>59 (28)</td>
<td>&lt;.001</td>
<td>.11</td>
</tr>
</tbody>
</table>

\(^a\)Food groups are Wheel of Five food groups [23].

\(^b\)McNemar test of the differences between the number of consumers of the DitEetIk! app and the GloboDiet 24-hour dietary recalls for the same day.

\(^c\)Wilcoxon signed rank test (normal approximation) of the differences between intake assessed using the DitEetIk! app and the GloboDiet 24-hour dietary recalls for the same day for consumers of each food group.

\(^d\)N/A: not applicable.

### Relative Validity for Nutrients

For 41% (12/29) of the nutrients, the median intake assessed using the DitEetIk! app and the 24-hour dietary recalls did not differ significantly (Table 4). Statistically significant differences of >10% between the 2 methods were observed for the median intake of vitamin A (453 vs 515 \( \mu \)g retinol activity equivalents/d; \( P < .001 \)), folate (equivalents; 241 vs 269 \( \mu \)g/d; \( P < .001 \)), vitamin D (1.8 vs 2.1 \( \mu \)g/d; \( P < .001 \)), and vitamin E (9.7 vs 11.7 mg/d; \( P < .001 \)). For 8 nutrients, significant differences were between 5% and 10%, and for 5 nutrients, they were <5%. With the exceptions of mono- and disaccharides and magnesium, in all cases of statistically significant differences, the DitEetIk! app had lower values than the 24-hour dietary recalls. Spearman correlation coefficients between intake assessed using the DitEetIk! app and the 24-hour dietary recalls ranged from 0.55 for sodium to 0.9 for alcohol, with a median correlation coefficient of 0.72.

Compared with the expected energy intake, the mean underreporting using the DitEetIk! app was 19.9% (0.316/1.59; calculated as [measured-expected physical activity level]/[expected physical activity level]) versus 20.1% (0.321/1.59) using the 24-hour dietary recalls. At the individual level, 20.4% (43/211) of the participants could be considered to be underreporting and 0.5% (1/211) of the participants could be considered to be overreporting energy intake using the DitEetIk! app. For the 24-hour dietary recalls, 18% (38/211) of the participants could be considered to be underreporting, and none were overreporting.

Figure 3 shows the Bland-Altman plots for energy intake and for fat, carbohydrates, and protein expressed as a percentage of energy intake. The mean differences and 95% limits of agreement were 14 (−1096 to 1124) for energy in kilocalories, −2.6 (−16.2 to 11) for energy percentage derived from fat, 2.5 (−10.9 to 16) for energy percentage derived from carbohydrates, and −7.5 (−7.5 to 7.4) for energy percentage derived from protein.
Table 4. The mean, SD, median, and IQR of energy and nutrient intake per day as assessed using the DitEetIk! app and the 24-hour dietary recalls and their correlation (n=211).

<table>
<thead>
<tr>
<th>Nutrients</th>
<th>DitEetIk! app food record</th>
<th>GloboDiet 24-hour dietary recall</th>
<th>Wilcoxon signed rank test P value&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Spearman correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy (kcal)</td>
<td>Values, mean (SD) 2112 (795)</td>
<td>Values, median (IQR) 1994 (1534-2488)</td>
<td>Values, mean (SD) 2100 (675)</td>
<td>1994 (1614-2539)</td>
</tr>
<tr>
<td>Fat (g)</td>
<td>83 (42)</td>
<td>76 (51-105)</td>
<td>88 (39)</td>
<td>81 (61-110)</td>
</tr>
<tr>
<td>Saturated fatty acids (g)</td>
<td>29 (15)</td>
<td>27 (18-38)</td>
<td>31 (15)</td>
<td>28 (20-39)</td>
</tr>
<tr>
<td>Protein (g)</td>
<td>82 (34)</td>
<td>77 (59-100)</td>
<td>82 (30)</td>
<td>77 (62-101)</td>
</tr>
<tr>
<td>Vegetable protein (g)</td>
<td>33 (15)</td>
<td>30 (23-41)</td>
<td>33 (13)</td>
<td>30 (23-42)</td>
</tr>
<tr>
<td>Carbohydrates (g)</td>
<td>230 (99)</td>
<td>215 (166-280)</td>
<td>215 (77)</td>
<td>208 (160-259)</td>
</tr>
<tr>
<td>Mono- and disaccharides (g)</td>
<td>93 (53)</td>
<td>81 (56-120)</td>
<td>85 (43)</td>
<td>80 (56-111)</td>
</tr>
<tr>
<td>Fiber (g)</td>
<td>21 (11)</td>
<td>20 (14-26)</td>
<td>21 (9)</td>
<td>20 (14-26)</td>
</tr>
<tr>
<td>Alcohol (g)</td>
<td>10 (21)</td>
<td>0 (0-11)</td>
<td>9 (20)</td>
<td>0 (0-10)</td>
</tr>
<tr>
<td>Water (g)</td>
<td>2681 (1012)</td>
<td>2570 (2048-3229)</td>
<td>2907 (955)</td>
<td>2775 (2285-3349)</td>
</tr>
<tr>
<td>Vitamin A (µg RAE&lt;sup&gt;b&lt;/sup&gt;)</td>
<td>714 (1007)</td>
<td>453 (286-685)</td>
<td>804 (1013)</td>
<td>515 (334-806)</td>
</tr>
<tr>
<td>Vitamin B1 (mg)</td>
<td>1.0 (0.6)</td>
<td>0.9 (0.7-1.3)</td>
<td>1.0 (0.5)</td>
<td>0.9 (0.7-1.3)</td>
</tr>
<tr>
<td>Vitamin B2 (mg)</td>
<td>1.3 (0.6)</td>
<td>1.3 (0.9-1.6)</td>
<td>1.4 (0.6)</td>
<td>1.3 (1.0-1.8)</td>
</tr>
<tr>
<td>Vitamin B3 (mg)</td>
<td>18.8 (10.8)</td>
<td>16.5 (10.8-24.4)</td>
<td>19.0 (9.3)</td>
<td>17.3 (11.5-24.4)</td>
</tr>
<tr>
<td>Vitamin B6 (mg)</td>
<td>1.5 (0.7)</td>
<td>1.4 (1.0-1.9)</td>
<td>1.6 (0.7)</td>
<td>1.5 (1.1-1.9)</td>
</tr>
<tr>
<td>Folate (equivalents; µg)</td>
<td>266 (148)</td>
<td>241 (172-331)</td>
<td>286 (122)</td>
<td>269 (203-357)</td>
</tr>
<tr>
<td>Vitamin B12 (µg)</td>
<td>3.8 (3.2)</td>
<td>2.9 (2.0-4.8)</td>
<td>4.2 (3.4)</td>
<td>3.2 (2.2-5.1)</td>
</tr>
<tr>
<td>Vitamin C (mg)</td>
<td>94 (125)</td>
<td>65 (32-108)</td>
<td>88 (74)</td>
<td>67 (33-128)</td>
</tr>
<tr>
<td>Vitamin D (µg)</td>
<td>2.3 (2.2)</td>
<td>1.8 (0.9-2.9)</td>
<td>2.7 (2.5)</td>
<td>2.1 (1.2-3.5)</td>
</tr>
<tr>
<td>Vitamin E (mg)</td>
<td>11.4 (7.3)</td>
<td>9.7 (6.5-14.3)</td>
<td>13.1 (6.7)</td>
<td>11.7 (8.2-16.7)</td>
</tr>
<tr>
<td>Calcium (mg)</td>
<td>965 (451)</td>
<td>909 (658-1231)</td>
<td>1006 (452)</td>
<td>921 (664-1295)</td>
</tr>
<tr>
<td>Iodine (µg)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>173 (92)</td>
<td>159 (111-213)</td>
<td>172 (79)</td>
<td>162 (116-209)</td>
</tr>
<tr>
<td>Iron (mg)</td>
<td>10.3 (4.5)</td>
<td>9.3 (7.2-12.7)</td>
<td>10.2 (4.0)</td>
<td>9.4 (7.2-12.1)</td>
</tr>
<tr>
<td>Magnesium (mg)</td>
<td>338 (126)</td>
<td>334 (249-403)</td>
<td>346 (118)</td>
<td>328 (264-411)</td>
</tr>
<tr>
<td>Phosphorus (mg)</td>
<td>1504 (537)</td>
<td>1466 (1086-1864)</td>
<td>1506 (489)</td>
<td>1478 (1135-1790)</td>
</tr>
<tr>
<td>Potassium (mg)</td>
<td>3202 (1202)</td>
<td>3147 (2314-3860)</td>
<td>3206 (1059)</td>
<td>3106 (2475-3886)</td>
</tr>
<tr>
<td>Selenium (µg)</td>
<td>52 (44)</td>
<td>43 (31-58)</td>
<td>50 (32)</td>
<td>43 (31-63)</td>
</tr>
<tr>
<td>Sodium (mg)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2625 (1443)</td>
<td>2373 (1637-3096)</td>
<td>2410 (1040)</td>
<td>2283 (1694-2875)</td>
</tr>
<tr>
<td>Zinc (mg)</td>
<td>10.1 (4.5)</td>
<td>9.4 (7.0-12.5)</td>
<td>10.3 (3.9)</td>
<td>10.0 (7.5-12.3)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Wilcoxon signed rank test (normal approximation) of the differences between intake assessed using the DitEetIk! app and the GloboDiet 24-hour dietary recalls for the same day.

<sup>b</sup>RAE: retinol activity equivalents.

<sup>c</sup>Sodium and iodine from food only.
Evaluation of the GloboDiet 24-Hour Dietary Recall Data

For almost all food groups, the consumption of participants in the DitEetIk! app evaluation study, assessed using the GloboDiet 24-hour dietary recall, and the consumption of matched participants selected from the DNFCS 2019 to 2021 did not differ significantly (Table 5). Statistically significant differences of >10% in median intake between the 2 methods were only observed for the consumption of milk and milk products (median 252 grams per day in the DitEetIk! app vs 282 grams per day in the GloboDiet 24-hour dietary recalls; \( P=0.03 \)) and bread (median 132 grams per day vs 105 grams per day; \( P=0.03 \)).

For all nutrients except vitamin B3, the intake of participants in the evaluation study, assessed using the GloboDiet 24-hour dietary recalls, and the intake of the matched participants selected from the DNFCS 2019 to 2021 did not differ significantly (Table 6).
Table 5. Comparison of consumption of food groups assessed using the GloboDiet 24-hour dietary recalls in the DitEetIk! app evaluation study and the first interview in the Dutch National Food Consumption Survey (DNFCS) 2019 to 2021 for a matched group of participants (n=211).

<table>
<thead>
<tr>
<th>Food group</th>
<th>DitEetIk! app evaluation study (g/d)</th>
<th>DNFC 2019-2021 (g/d)</th>
<th>P value&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Values, mean (SD)</td>
<td>Values, median (IQR)</td>
<td>Values, mean (SD)</td>
</tr>
<tr>
<td>Vegetables</td>
<td>160 (144)</td>
<td>130 (50-240)</td>
<td>155 (140)</td>
</tr>
<tr>
<td>Fruit</td>
<td>140 (146)</td>
<td>130 (0-217)</td>
<td>124 (135)</td>
</tr>
<tr>
<td>Added fats</td>
<td>19 (15)</td>
<td>17 (6-29)</td>
<td>22 (20)</td>
</tr>
<tr>
<td>Fish</td>
<td>17 (57)</td>
<td>0 (0-0)</td>
<td>15 (44)</td>
</tr>
<tr>
<td>Legumes</td>
<td>4 (20)</td>
<td>0 (0-0)</td>
<td>8 (36)</td>
</tr>
<tr>
<td>Meat</td>
<td>17 (57)</td>
<td>0 (0-0)</td>
<td>15 (44)</td>
</tr>
<tr>
<td>Eggs</td>
<td>4 (20)</td>
<td>0 (0-0)</td>
<td>8 (36)</td>
</tr>
<tr>
<td>Nuts</td>
<td>92 (83)</td>
<td>75 (33-120)</td>
<td>88 (80)</td>
</tr>
<tr>
<td>Milk and milk products</td>
<td>17 (34)</td>
<td>0 (0-15)</td>
<td>16 (32)</td>
</tr>
<tr>
<td>Cheese</td>
<td>15 (30)</td>
<td>0 (0-20)</td>
<td>19 (45)</td>
</tr>
<tr>
<td>Bread</td>
<td>288 (248)</td>
<td>252 (80-423)</td>
<td>332 (267)</td>
</tr>
<tr>
<td>Cereal products</td>
<td>39 (44)</td>
<td>31 (0-62)</td>
<td>38 (39)</td>
</tr>
<tr>
<td>Potatoes</td>
<td>138 (88)</td>
<td>132 (70-180)</td>
<td>117 (80)</td>
</tr>
<tr>
<td>Drinks</td>
<td>74 (106)</td>
<td>20 (0-119)</td>
<td>79 (108)</td>
</tr>
<tr>
<td>Sandwich spreads</td>
<td>66 (104)</td>
<td>0 (0-120)</td>
<td>69 (93)</td>
</tr>
<tr>
<td>Soups</td>
<td>2097 (889)</td>
<td>1963 (1582-2539)</td>
<td>2132 (914)</td>
</tr>
<tr>
<td>Snacks</td>
<td>12 (23)</td>
<td>0 (0-15)</td>
<td>18 (29)</td>
</tr>
<tr>
<td>Sauces</td>
<td>12 (46)</td>
<td>0 (0-0)</td>
<td>17 (66)</td>
</tr>
<tr>
<td>Other</td>
<td>83 (89)</td>
<td>56 (14-126)</td>
<td>71 (79)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Food groups are Wheel of Five food groups [23].

<sup>b</sup>Wilcoxon signed rank test (normal approximation) of the differences between intake assessed using GloboDiet 24-hour dietary recalls in the DitEetIk! app evaluation study and the first interview with adults in the DNFCS 2019 to 2021.
Table 6. Comparison of intake of energy and selected nutrients assessed using the GloboDiet 24-hour dietary recalls in the DitEetIk! app evaluation study and the first interview with adults in the Dutch National Food Consumption Survey (DNFCS) 2019 to 2021 (n=211).

<table>
<thead>
<tr>
<th>Nutrients</th>
<th>DitEetIk! app food record</th>
<th>GloboDiet 24-hour dietary recall</th>
<th>P value&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Values, mean (SD)</td>
<td>Values, median (IQR)</td>
<td></td>
</tr>
<tr>
<td>Energy (kcal)</td>
<td>2100 (675)</td>
<td>1994 (1614-2539)</td>
<td>.86</td>
</tr>
<tr>
<td>Fat (g)</td>
<td>88 (39)</td>
<td>81 (61-110)</td>
<td>.86</td>
</tr>
<tr>
<td>Saturated fatty acids (g)</td>
<td>31 (15)</td>
<td>28 (20-39)</td>
<td>.48</td>
</tr>
<tr>
<td>Protein (g)</td>
<td>82 (30)</td>
<td>77 (62-101)</td>
<td>.29</td>
</tr>
<tr>
<td>Vegetable protein (g)</td>
<td>33 (13)</td>
<td>30 (23-42)</td>
<td>.31</td>
</tr>
<tr>
<td>Carbohydrates (g)</td>
<td>215 (77)</td>
<td>208 (160-259)</td>
<td>.71</td>
</tr>
<tr>
<td>Mono- and disaccharides (g)</td>
<td>85 (43)</td>
<td>80 (56-111)</td>
<td>.21</td>
</tr>
<tr>
<td>Fiber (g)</td>
<td>21 (9)</td>
<td>20 (14-26)</td>
<td>.18</td>
</tr>
<tr>
<td>Alcohol (g)</td>
<td>9 (20)</td>
<td>0 (0-10)</td>
<td>.64</td>
</tr>
<tr>
<td>Water (g)</td>
<td>2907 (955)</td>
<td>2775 (2285-3349)</td>
<td>.75</td>
</tr>
<tr>
<td>Vitamin A (µg RAE&lt;sup&gt;b&lt;/sup&gt;)</td>
<td>804 (1013)</td>
<td>515 (334-806)</td>
<td>.42</td>
</tr>
<tr>
<td>Vitamin B1 (mg)</td>
<td>1.0 (0.5)</td>
<td>0.9 (0.7-1.3)</td>
<td>.96</td>
</tr>
<tr>
<td>Vitamin B2 (mg)</td>
<td>1.4 (0.6)</td>
<td>1.3 (1.0-1.8)</td>
<td>.79</td>
</tr>
<tr>
<td>Vitamin B3 (mg)</td>
<td>19.0 (9.3)</td>
<td>17.3 (11.5-24.4)</td>
<td>.02</td>
</tr>
<tr>
<td>Vitamin B6 (mg)</td>
<td>1.6 (0.7)</td>
<td>1.5 (1.1-1.9)</td>
<td>.29</td>
</tr>
<tr>
<td>Folate (equivalents; µg)</td>
<td>286 (122)</td>
<td>269 (203-357)</td>
<td>.13</td>
</tr>
<tr>
<td>Vitamin B12 (µg)</td>
<td>4.2 (3.4)</td>
<td>3.2 (2.2-5.1)</td>
<td>.65</td>
</tr>
<tr>
<td>Vitamin C (mg)</td>
<td>88 (74)</td>
<td>67 (33-128)</td>
<td>.39</td>
</tr>
<tr>
<td>Vitamin D (µg)</td>
<td>2.7 (2.5)</td>
<td>2.1 (1.2-3.5)</td>
<td>.52</td>
</tr>
<tr>
<td>Vitamin E (mg)</td>
<td>13.1 (6.7)</td>
<td>11.7 (8.2-16.7)</td>
<td>.85</td>
</tr>
<tr>
<td>Calcium (mg)</td>
<td>1006 (452)</td>
<td>921 (664-1295)</td>
<td>.39</td>
</tr>
<tr>
<td>Iodine (µg)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>172 (79)</td>
<td>162 (116-209)</td>
<td>.42</td>
</tr>
<tr>
<td>Iron (mg)</td>
<td>10.2 (4.0)</td>
<td>9.4 (7.2-12.1)</td>
<td>.35</td>
</tr>
<tr>
<td>Magnesium (mg)</td>
<td>346 (118)</td>
<td>328 (264-411)</td>
<td>.86</td>
</tr>
<tr>
<td>Phosphorus (mg)</td>
<td>1506 (489)</td>
<td>1478 (1135-1790)</td>
<td>.67</td>
</tr>
<tr>
<td>Potassium (mg)</td>
<td>3206 (1059)</td>
<td>3106 (2475-3886)</td>
<td>.99</td>
</tr>
<tr>
<td>Selenium (µg)</td>
<td>50 (32)</td>
<td>43 (31-63)</td>
<td>.40</td>
</tr>
<tr>
<td>Sodium (mg)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2410 (1040)</td>
<td>2283 (1694-2875)</td>
<td>.42</td>
</tr>
<tr>
<td>Zinc (mg)</td>
<td>10.3 (3.9)</td>
<td>10.0 (7.5-12.3)</td>
<td>.53</td>
</tr>
</tbody>
</table>

<sup>a</sup> Wilcoxon signed rank test (normal approximation) of the differences between intake assessed using the GloboDiet 24-hour dietary recalls in the DitEetIk! app evaluation study and the first interview with adults in the DNFCS 2019 to 2021.

<sup>b</sup> RAE: retinol activity equivalents.

<sup>c</sup> Sodium and iodine from food only.

**Discussion**

Principal Findings

Compared with GloboDiet 24-hour dietary recalls, the DitEetIk! app assessed similar mean levels of energy intake but somewhat lower median levels of intake for several food groups and nutrients. Differences were of >10% for fruit; added fats; cereal products; sauces; and vitamins A, D, and E and folate. Of all logged food items and beverages, most were selected via text searching, whereas the scanning functionality was used for approximately one-seventh of the food products.

Incomplete recording of consumed food items in the DitEetIk! app seems to have occurred for various food groups, such as fruit, added fats, cereal products, and sauces. During the GloboDiet 24-hour dietary recall, the trained interviewer specifically probes for easily forgotten food items [28], which
may explain this difference. Similarly, other studies evaluating mobile food record apps based on text searching reported food omissions, particularly of condiment food items [29]. In a study using wearable camera images as a reference, it was observed that the most forgotten food groups in the Australian Eat and Track app were savory sauces and condiments, vegetables, confectionery, fruit and dairy, and alternatives [30]. There are various options to stimulate complete recording. According to a review of smartphone dietary assessment tools, the most common feature to do so was to allow participants to review the records and make adjustments if information was missing or false [31]. This feature was also built into the DitEetIk! app; before submitting the food recording for one day, participants were shown an overview of reported food items and were asked whether this was complete. In addition, incorporating (more) probing questions for frequently omitted food items into DitEetIk! app could be a way to remind a participant to report such food items. This could be probing questions either linked to other food items (eg, salad dressing in the case of salads) or linked to eating occasions (eg, fruit in between meals). Alternatively, sending prompts when an eating moment is expected or when the DitEetIk! app has not been used for a certain period or allowing participants to explicitly state that they did not consume anything at a given eating occasion are also approaches with the potential to improve completeness [32].

In contrast to the results for median intake, for some food groups and nutrients, the mean intake was higher in the DitEetIk! app than in the 24-hour dietary recall. For the food groups Vegetables, Meat, Bread, and Potatoes, this was caused by higher amounts recorded in the DitEetIk! app than those indicated in the 24-hour dietary recalls. Choosing unlikely high portion sizes in the DitEetIk! app was possible without a warning message in case the amount eaten was indicated in units rather than grams, whereas in the GloboDiet 24-hour dietary recalls, all indicated portions were converted into grams and checked against set improbable maximum values, and if needed, the interviewer was prompted to check with the participant whether the answer was correct [28]. A similar functionality could be considered for inclusion in the DitEetIk! app.

Of all logged food and beverages, most were selected via text searching. In the feedback given by participants in the remark field of the DitEetIk! app, they mentioned that finding the correct food item on the list was a challenging task. This disadvantage of extensive food item lists was also described in a systematic review [33]. This was probably the reason why the average SUS score for the DitEetIk! app was just below 70, the disadvantage of extensive food item lists was also described in a systematic review [33]. This was probably the reason why the average SUS score for the DitEetIk! app was just below 70, the threshold that is generally considered “good” [34]. Only 12.8% (114/894) of all logged food items and drinks were scanned. This percentage was lower than expected based on experiences in a project in which approximately 50% of the food items were scanned using a commercial smartphone food record (personal communication by MO). The availability of branded food items in the DitEetIk! app was still limited, and these did not include food items from some supermarket chains or that were not matched to the generic food composition database. This may have affected the use of the scanning option. One could understand that participants stopped scanning barcodes after some failed attempts. Therefore, including more branded food items in the DitEetIk! app is highly recommended. In addition, a crowdsourcing function could be incorporated whereby users can contribute information on missing products, such as that developed for the FoodSwitch app [35]. The collected food product information can then be added to the database to ensure that the DitEetIk! app is supported by actual and complete product information. If more food items are scanned, food recording will probably be perceived as easier.

As described previously, based on the main findings, several options for improvement via additional DitEetIk! app functionalities can be formulated. However, one should also be careful not to burden participants with too frequent notifications, reminders, and prompts [33]. More insight on the impact of different features used in smartphone-based dietary assessment tools and the characteristics of these features on the respondents’ willingness and ability to record intake reliably and on the validity of the recorded dietary data is needed [31].

Strengths and Limitations

This study is one of the few food record validation studies (Zhang et al [29]) that report results for a rather comprehensive list of food groups and nutrients in a group of >200 men and women of various ages. An important limitation is that relative rather than objective validity was studied. Therefore, lower and higher values compared with GloboDiet 24-hour dietary recall values cannot be interpreted as underestimation or overestimation. However, the results on energy misreporting were included, which are not dependent on the subjective reporting of dietary intake. In the future, follow-up validation with doubly labeled water and excretion of nitrogen, potassium, and sodium in 24-hour urine is recommended. Another limitation is the large number of statistical tests that were conducted, which may have led to chance findings. Moreover, the study did not follow a crossover design, which might have caused a potential memory or learning effect in the 24-hour dietary recall data. However, comparing these data with those from the DNFCS 2019 to 2021 gave no indication that this occurred. People with a lower education and of higher age were included in the study population but were underrepresented. We cannot conclude whether the relative validity is similar for these population subgroups. This study did not focus on the potential selection bias of including only participants who had an Android smartphone. Such an evaluation is also important for use in a national food consumption survey. Potentially, developing an iOS version of the DitEetIk! app and offering an interview option to persons without a smartphone needs to be considered. In this study, the DitEetIk! app was described according to the recommendations of Eldridge et al [11], and the validity study was reported according to the guidance provided by Kirkpatrick et al [26].

Comparison With Prior Work

Although many different smartphone-based dietary assessment tools exist, only a few validation studies have been conducted. Burrows et al [36] concluded that their validity seems to be comparable with that of more traditional dietary assessment methods and that energy intake is often underestimated. In a...
review from 2013 to 2019, Zhang et al [29] identified 14 smartphone-based food records that were not image based. In the meta-analyses based on 11 tools, all of them underestimated energy intake, with a pooled effect of approximately $-200 \text{ kcal}$ and limits of agreement of 1918 kcal. The results for the DitEetIk! app fit with this picture, with a below-average mean underestimation and similar limits of agreement for energy intake. The relative validity results were also in line with those of Eetmeter, another Dutch app, although the food groups for which underestimation occurred partly differed [20]. Unlike the DitEetIk! app, Eetmeter shows energy and nutrient values for logged food items, which might influence reporting.

Conclusions

Compared with GloboDiet 24-hour dietary recalls, the DitEetIk! app assessed somewhat lower levels of intake for several food groups and nutrients. Therefore, adding functionalities to the DitEetIk! app that stimulate more complete food recording is important before using the app in national food consumption surveys. In addition, it is advisable to develop a functionality to warn participants when entering extremely large consumption amounts. Less participant burden and more detailed information about consumed food items can be obtained by stimulating the use of barcode scanning.

Acknowledgments

This study as well as the development of the DitEetIk! app was funded by the Dutch Ministry of Health, Welfare, and Sport. The authors would like to thank Dr Maarten Schipper for providing statistical advice and Dr Steffen Bruns, Henny Brants, and Marjolein de Jong for critically reviewing the manuscript.

Data Availability

The data presented in this paper are available upon request from the corresponding author. The data are not publicly available to comply with the requirements of the General Data Protection Regulation (European Union) 2016/679.

Conflicts of Interest

National Institute for Public Health and the Environment (RIVM), where the authors work, is the owner of the software.

Multimedia Appendix 1

The mean, SD, median, and IQR of consumption of food groups (g/d) by educational level as assessed using the DitEetIk! app and 24-hour dietary recalls for the same day and their correlation for 211 participants with plausible energy intakes.

[DOCX File, 29 KB - mhealth_v12i1e50196_app1.docx]

Multimedia Appendix 2

The mean, SD, median, and IQR of consumption of food groups (g/d) by BMI as assessed using the DitEetIk! app and 24-hour dietary recalls for the same day and their correlation for 211 participants with plausible energy intakes.

[DOCX File, 29 KB - mhealth_v12i1e50196_app2.docx]

References


Abbreviations

- **BMR**: basal metabolic rate
- **DNFCS**: Dutch National Food Consumption Survey
- **LEDA**: Dutch-branded food database (Levensmiddelendatabank)
- **SUS**: System Usability Scale
Feasibility Study on Menstrual Cycles With Fitbit Device (FEMFIT): Prospective Observational Cohort Study

Anna-Lena Lang¹, MD; Rosa-Lotta Bruhn²; Maya Fehling¹, Dr med; Anouk Heidenreich¹; Jonathan Reisdorf¹; Ifrah Khanyaree¹, MSc; Maike Henningsen², Dr med, PD; Cornelius Remschmidt¹, Dr med, PD

¹Data4Life gGmbH, Potsdam, Germany
²Faculty of Health, University Witten Herdecke, Witten Herdecke, Germany

Corresponding Author:
Anna-Lena Lang, MD
Data4Life gGmbH
c/o Digital Health Cluster im Hasso-Plattner-Institut
Potsdam, 14482
Germany
Phone: 49 15756025551
Email: annalena.lang.26@gmail.com

Abstract

Background: Despite its importance to women’s reproductive health and its impact on women’s daily lives, the menstrual cycle, its regulation, and its impact on health remain poorly understood. As conventional clinical trials rely on infrequent in-person assessments, digital studies with wearable devices enable the collection of longitudinal subjective and objective measures.

Objective: The study aims to explore the technical feasibility of collecting combined wearable and digital questionnaire data and its potential for gaining biological insights into the menstrual cycle.

Methods: This prospective observational cohort study was conducted online over 12 weeks. A total of 42 cisgender women were recruited by their local gynecologist in Berlin, Germany, and given a Fitbit Inspire 2 device and access to a study app with digital questionnaires. Statistical analysis included descriptive statistics on user behavior and retention, as well as a comparative analysis of symptoms from the digital questionnaires with metrics from the sensor devices at different phases of the menstrual cycle.

Results: The average time spent in the study was 63.3 (SD 33.0) days with 9 of the 42 individuals dropping out within 2 weeks of the start of the study. We collected partial data from 114 ovulatory cycles, encompassing 33 participants, and obtained complete data from a total of 50 cycles. Participants reported a total of 2468 symptoms in the daily questionnaires administered during the luteal phase and menses. Despite difficulties with data completeness, the combined questionnaire and sensor data collection was technically feasible and provided interesting biological insights. We observed an increased heart rate in the mid and end luteal phase compared with menses and participants with severe premenstrual syndrome walked substantially fewer steps (average daily steps 10,283, SD 6277) during the luteal phase and menses compared with participants with no or low premenstrual syndrome (mean 11,694, SD 6458).

Conclusions: We demonstrate the feasibility of using an app-based approach to collect combined wearable and questionnaire data on menstrual cycles. Dropouts in the early weeks of the study indicated that engagement efforts would need to be improved for larger studies. Despite the challenges of collecting wearable data on consecutive days, the data collected provided valuable biological insights, suggesting that the use of questionnaires in conjunction with wearable data may provide a more complete understanding of the menstrual cycle and its impact on daily life. The biological findings should motivate further research into understanding the relationship between the menstrual cycle and objective physiological measurements from sensor devices.

(JMIR Mhealth Uhealth 2024;12:e50135) doi:10.2196/50135

KEYWORDS
women’s health; menstrual cycle; premenstrual syndrome; PMS; mobile app; wearable device; sensor data; digital health
**Introduction**

An estimated 56% to 93% of women across the world experience recurrent painful periods [1-5]. Menstrual pain is debilitating for many women and has a major impact on their health-related quality of life [6-11]. One condition that can cause painful menstruation is endometriosis, which affects 10% to 15% of women of reproductive age [12]. In addition to severe pain, intermenstrual bleeding, painful periods (dysmenorrhea), painful intercourse (dyspareunia), painful bowel movements (dyschezia), and painful urination (dysuria), endometriosis can cause infertility, increase psychological distress, and can affect sexuality and relationships [12,13]. Despite its debilitating nature, it still takes around 6 to 8 years from symptom onset to diagnosis [14,15]. The gold standard for diagnosis is still laparoscopy, which is not only risky but also expensive. With the rise of gender equality movements, clinical research into menstrual health has slowly gained more attention in recent years, with researchers emphasizing the importance of timely diagnosis and treatment of menstrual cycle-related concerns and disorders [16]. Each individual has a different baseline of subjective pain experience and the availability of sufficient baseline data can enable a shift toward precision medicine in menstrual health [17]. Frequent and continuous data collection can provide an understanding of symptom variability, which is likely to be an important contributor to variability in treatment response [17]. However, traditional clinical trials to date have relied on infrequent in-person assessments and subjective retrospective data, failing to capture the daily changes in physical and mental well-being that occur over the course of the menstrual cycle [18,19]. In this regard, commercially available menstrual tracker apps offer new opportunities for research. Women can continuously track their menstrual health in digital diaries, with self-learning algorithms continuously improving predictions and educational content empowering users to increase their knowledge about menstrual health [20-22]. As a complement to subjective digital questionnaires, commercially available wearable technology can provide an easy way to collect continuous and objective real-world health data for women’s health research purposes. In the example of endometriosis, the combined collection of subjective data from digital diaries and objective wearable devices could help distinguish between “normal” menstrual pain and pain associated with endometriosis, potentially speeding up the diagnosis process and avoiding unnecessary invasive tests.

Previous research, albeit with small sample sizes, has already shown that the collection of sensor data such as step count, heart rate, and sleep duration combined with self-reported menstrual cycle data can uncover interesting correlations and advance knowledge about menstrual health [23-26]. For example, wearables are already increasingly being used to evaluate alternative contraceptive methods and to predict the fertility window [26-29]. Interesting findings from previous research using wearables include an observed lower distal skin temperature (as measured with an Oura ring) during ovulation, as well as higher heart rate in the ovulatory, mid, and late luteal phases [24,25]. While previous studies show a potential effect of physical activity on the menstrual cycle and vice versa [30-32], none of them used wearable devices for daily activity tracking during the menstrual cycle. Similarly, there is a lack of studies using wearable data to analyze sleep behavior during the menstrual cycle. Current scientific knowledge on changes in sleep behavior across the menstrual cycle is conflicting, with some studies finding a decrease in subjective and objective sleep quality during the premenstrual phase and menses [33-36], while other studies did not find such correlations [24,37]. The ongoing Apple Women’s Health Study [27], a mobile app–based longitudinal cohort study that includes both survey and sensor-based data, has not yet published results related to sensor data.

Combining subjective data from women’s health questionnaires with objective sensor data from wearables can not only facilitate cycle tracking for the everyday consumer but also allow researchers and participants alike, to gain a deeper insight into the clinical changes during the menstrual cycle. Commercially available wrist-worn sensor devices, such as the Fitbit Inspire 2 device, are thereby not only much more affordable than research-grade sensor devices, but can be nearly as accurate [38,39]. However, real-world data collection studies conducted exclusively in the home can present difficulties in terms of retention and adherence to the study protocol [19,40]. To explore the feasibility of consistently collecting wearable and questionnaire data across multiple menstrual cycles, we conducted a 12-week feasibility study with 42 participants. Using the Fitbit Inspire 2 device and digital questionnaires within the Data4Life study app, our primary objective was to refine methods for collecting authentic menstrual cycle data in a real-world setting. This included assessing participant retention rates and gathering usability feedback on their engagement with wearable devices and digital questionnaires. In addition, our study explored potential clinical correlations between wearable and questionnaire data, with the aim of uncovering potential correlations with key parameters of the menstrual cycle.

**Methods**

**Study Design**

This digital prospective cohort study was conducted mainly online with enrollment between December 2021 and April 2022. This study was called the Feasibility Study of Menstrual Cycles With Fitbit Device (FEMFIT). Participants were recruited by their local gynecologist at a practice in Berlin, Germany. The owner of the practice, who is also a coauthor of this study (MH), received financial incentives for recruiting participants. During recruitment, participants received a token to access the FEMFIT study in the study app. During account creation, participants provided digital informed consent to share their data for research. All participants received a free Fitbit Inspire 2 (Fitbit International Limited) as an incentive after completing the study. Participants were asked to wear the device at all times throughout the study period of 3 menstrual cycles or 12 weeks. Dropout was defined as individuals who did not provide data for more than 14 days. This cut-off was deliberately set to manage participant dropout while ensuring the inclusion of data from individuals with shorter menstrual cycles.
The primary outcome variables of this study centered on assessing the feasibility of continuous menstrual cycle data collection using the Fitbit Inspire 2 device and digital questionnaires. Specifically, we focused on participant retention rates and usability feedback regarding engagement with these tools. In addition, secondary outcome variables included exploring potential clinical correlations between the collected wearable and questionnaire data and key menstrual cycle parameters.

**Data Collection: Digital Questionnaires and Wearable Data**

Prior to enrollment, participants completed a paper questionnaire to assess their digital literacy [41] (Multimedia Appendix 1).

Digital questionnaires (Multimedia Appendix 1; Figure S2 in Multimedia Appendix 2) were accessible within the app at varying intervals, with email notifications reminding participants of new questions. Email notifications reminded participants of newly available questions. At enrollment, participants provided demographic information and were asked for clinical information with a focus on women’s health, including preexisting conditions, use of hormonal contraception, cycle regularity, and the first day of the last menses (Multimedia Appendix 1). After enrollment, weekly questionnaires focused on monitoring the duration of wearable device use duration and assessing mental and physical well-being (Multimedia Appendix 1). Notably, the specific results of the mental and physical well-being assessments are not presented in this publication. Triggered by the first day of their last menstrual cycle, from day 13 to day 5 of their menstrual cycle, participants were asked about the 12 typical symptoms of premenstrual syndrome (PMS): seclusion, irritability, swelling, anger, weight gain, joint pain, headache, confusion, dejection, bloating, anxiety, and tenderness [42]: participants were also asked about the severity of their bleeding. If a new menstrual cycle had begun, the first day of bleeding triggered the new cadence for all future questionnaires. Data from these daily questionnaires were used for correlation with the wearable data collected simultaneously.

In addition to the digital questionnaires, the Fitbit device recorded 3 parameters daily: average resting heart rate, total steps, and total sleep time. For this study, average resting heart rate was rounded to the nearest 5 beats per minute, sleep time was rounded to the nearest 10 minutes, and step count was rounded to the nearest 100 steps. This generalization was made to protect the anonymity of participants when analyzing deidentified donation records in this small feasibility study.

**Definitions**

The relevant phases of the menstrual cycle were defined as follows: the menses phase (MP; days 1 to 5), the ovulation day (OD; day 1 of the next MP minus 14) [43], the follicular phase (day 1 of MP to OD minus 1), the luteal phase (OD plus 1 to the menses), and the midluteal phase (OD plus 3 to OD plus 9).

In our analysis of PMS, only PMS symptoms from the days after OD to the end of MP were considered. Severe PMS was defined as a symptom severity rating of 3, equivalent to severe, on the visual analog scale. Participants were included in the severe PMS group if they reported severe PMS symptoms in at least 2 cycles. Participants were included in the no or low PMS group if they never reported PMS symptom severity higher than 1 on a severity scale from 0 to 3.

**Data Storage and Wearable Data Integration**

The nonprofit organization Data4Life provided the research infrastructure (study application, data storage, and analysis platform) for this study. All research data was stored in Data4Life’s secure research environment on servers in Germany. Data4Life is certified by the German Federal Office for Information Security (BSI). For participants who allowed the use of cookies, we were able to analyze user behavior using the General Data Protection Regulation–compliant business analytics tool Matomo (Matomo; data stored in Europe only).

In the web app, after entering their study token and providing digital study consent, participants were asked to connect to Fitbit to allow access to their wearable data via the Fitbit web application programming interface (API) [44]. After completing the OAuth 2.0 Authorization Code Grant Flow [45], the API refresh token was stored in end-to-end encrypted form for subsequent use for the duration of their participation in the study or until they withdrew their consent. This allowed the study’s wearable parameters to be retrieved client-side via the respective activity and sleep endpoints, and stored in end-to-end encrypted form as Fast Healthcare Interoperability Resources Standard for Trial Use, version 3. [46] observation records each time participants logged into the web app to complete their questionnaires. As the API returns time series data over a few days, data gaps between donations were filled by comparing previously stored records with the data points returned by the API.

**Statistical Analysis**

All analyses were performed on the Data4Life analytics platform on a jupyterhub notebook running Python (version 3.10.4), using pandas (version 1.4.2), matplotlib (version 3.5.2), seaborn (version 0.11.2), and numpy (version 1.22.4). Descriptive statistics were used to report details of the study cohort, retention, and adherence measures. Continuous values were reported as mean with SD or median with IQR; categorical values were reported as numbers with percentages. For comparative analyses of clinical outcomes, we used the Mann-Whitney test [47] for nonnormally distributed data and the 2-tailed t test [48] for normally distributed data.

**Ethical Considerations**

The study was approved by the ethics committee of the Berlin Chamber of Physicians (Eth-11/22). Registration was open for people aged 18 years and older. All participants provided digital informed consent for study participation. Participation was voluntary. Participants were allowed to keep the Fitbit Inspire 2 device after the study ended. Email and password were required to log in to the study app as well as two-factor authentication via phone. Participants could access all study content through a web app on both desktop and mobile devices. All study data was end-to-end encrypted and pseudonymized. Only authorized researchers were provided access to the data on the Data4Life analytics platform. Data was stored exclusively on Data4Life data centers in Germany. Based on IT-Grundschutz.
Results

Study Setting and Participants

This feasibility study of combined wearable and questionnaire data collection on the menstrual cycle was conducted online with self-reporting of questionnaire data in a study app accompanied by the collection of wrist-worn sensor data from Fitbit Inspire 2 devices. A total of 42 cisgender women were recruited from December 2021 to April 2022 at a gynecological practice in Berlin and agreed to provide menstrual cycle data for 12 weeks (Figure 1).

One participant never requested a wearable device and therefore dropped out of the study before creating an account in the study app. Three participants never donated any questionnaire or wearable data (Figure S1 in Multimedia Appendix 2), leaving data from 38 participants who started to donate data in the study app. Of these, a total of 5 participants withdrew from the study within the first 2 weeks of study entry, 3 of whom never donated any wearable data, leaving valuable data from 33 participants for further analysis. Active participants (n=33) were on average 24.2 (SD 3.0) years old and predominantly identified as female, with 1 individual identifying as diverse (Table 1). Five participants (15.1%) reported preexisting mental health problems (Table 1). In terms of women’s health issues, 2 (6.1%) participants reported a history of chlamydia infection and 1 person (3%) reported a history of human papillomavirus infection. At enrollment, the median time since the last menstruation was 17 (IQR 12-26) days.

Figure 1. Feasibility Study of Menstrual Cycles With Fitbit Device (FEMFIT) study concept. The FEMFIT study on collecting combined wearable and questionnaire data on the menstrual cycle was carried out online. Self-reporting of questionnaire data in a study app was accompanied by the collection of wrist-worn sensor data from Fitbit Inspire 2 devices. Weekly questionnaires asked about physical and mental well-being. From day 13 to day 5 of the menstrual cycle, participants answered daily questionnaires on bleeding severity and PMS symptoms. PMS: premenstrual syndrome.
Table 1. Details on study participants (n=33).

<table>
<thead>
<tr>
<th>Demographics of active study participants</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>24.2 (3.0)</td>
</tr>
<tr>
<td><strong>Gender, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>32 (97)</td>
</tr>
<tr>
<td>Diverse</td>
<td>1 (3)</td>
</tr>
<tr>
<td><strong>Preexisting health condition, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Chlamydia</td>
<td>2 (6.1)</td>
</tr>
<tr>
<td>HPV</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Mental health</td>
<td>5 (15.1)</td>
</tr>
<tr>
<td>Hypothyroidism</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Hyperthyroidism</td>
<td>1 (3)</td>
</tr>
<tr>
<td><strong>Cycle regularity, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Irregular</td>
<td>2 (6.1)</td>
</tr>
<tr>
<td><strong>Hormonal contraception, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>4 (12.1)</td>
</tr>
<tr>
<td>Days since last period, median (IQR)</td>
<td>17 (12-26)</td>
</tr>
</tbody>
</table>

aHPV: human papillomavirus.

High Degree of Digital Literacy of Study Participants

Of the 42 people initially recruited, 25 (59.5%) people reported that they had never worn a wearable device before. Of those who had experience wearing a sensor device, 5 had worn an Apple Watch, and other devices worn included Fitbit (n=2), Garmin (n=1), Samsung (n=1), and Denver (n=1). The majority of participants (26/42, 61.9%) regularly used apps on their smartphones to track their food intake, physical activity, or menstrual cycle. The most commonly used menstrual cycle tracking apps were Flo (9/26, 34.6%) and Clue (4/26, 15.3%). At the start of the study, a standardized digital literacy questionnaire [41] showed that the participants were highly digitally literate with an average score of 37.1 (SD 3.1) out of 48.

Increased User Engagement After Early Dropouts

Across all 41 participants who enrolled in the study and logged into the study app, the mean time spent in the study (measured from enrollment to last weekly or daily questionnaire delivery) was 63.3 (SD 33.0) days. The reported reasons for not completing the study were too great a time commitment (n=1), a high technical barrier (n=2), and technical difficulties that made it impossible to complete the study (n=3).

Among participants who remained in the study for more than 14 days (n=33), questionnaire completion rates were high. The weekly physical and mental well-being questionnaires were completed on average 10.2 times (SD 2.6) out of 12, while 69.7% (n=23) of active participants completed at least 11 of the 12 weekly questionnaires. An average of 65.8% (SD 24.1%) of all questionnaires on menstrual bleeding and PMS symptoms, which were asked daily from day 13 to day 5 of the menstrual cycle, were completed (Figure 2A). Overall, 66.7% (n=22) of participants reported starting dates of at least 3 menstrual cycles, with a median cycle length of 28.0 days (IQR 24-31), and we collected data from a total of 50 complete ovulatory cycles. In general, logins to our study app occurred mostly after email notifications (Figure S3 in Multimedia Appendix 2), suggesting that frequent study reminders may increase retention.

The average number of days wearable data were provided was 65.2 (SD 17.0) days for step count, 64.2 (SD 17.1) days for heart rate, and 60.5 (SD 17.8) days for sleep time (Figure 2B). There were difficulties in collecting continuous wearable data over the entire study period, particularly for sleep time. None of the 33 active participants provided consecutive wearable data over the entire study period. The average number of consecutive days of data donation was 36.4 (SD 25.6) days for sleep time, 38.6 (SD 31.4) days for heart rate, and 52.6 (SD 28.3) days for step count. Wearable data was provided consecutively for at least 50 days by 11 participants and for more than 70 days by 3 participants. In the weekly study questionnaire, participants reported that they were wearing the device continuously for an average of 9 (SD 2.9) weeks out of 12. When participants reported that they did not wear the device continuously, off-times were stated to be longer than 3 hours per day.
Figure 2. Participant engagement. (A) Questionnaire completion rates of weekly and daily questionnaires. Questionnaires on physical and mental well-being were filled out once per week, whereas questions on premenstrual symptoms and bleeding severity were filled out daily from day 13 to day 5 of the menstrual cycle. The box plots show the median and IQR, with whiskers reflecting minimum and maximum values for outliers. Dots represent all data points. (B) Bar plot visualizing the total days of wearable data donation during the study period per participant. Bars are color-coded by the parameter assessed using the Fitbit Inspire 2 wearable device.

Fitbit Device and Study App Supported Subjective Health Monitoring

At the end of the study, 22 participants completed questions about their study experience. Most participants reported that the Fitbit device was very easy to use (n=22; mean rating 1.6, SD 0.6; 1=very easy, 5=very challenging) and did not interfere with their daily life (n=22; mean rating 1.2, SD 2; 1=very comfortable, 5=very disturbing). One person found wearing the Fitbit device at night very uncomfortable, whereas overall it was not perceived to be bothersome at night (n=22; mean rating 2.2, SD 1).

The majority of participants completing the end-of-study questionnaire (13/22, 59%) stated that wearing the device increased their physical activity over the study period. Step count data from all participants over the study period thereby showed an average step count of 9987 (SD 3856) steps per day in the last 2 weeks of study participation compared with an average of 9301 (SD 4076) steps per day in the first 2 weeks of study participation, but this observation was not significant.

Overall, participants did not feel that the sensor tracking changed their sleep behavior (13/22, 59%). Five participants reported that wearing the device and answering the questionnaires made them feel stressed, while 9 participants reported that wearing the Fitbit device at night very uncomfortable, whereas overall it was not perceived to be bothersome at night (n=22; mean rating 2.2, SD 1).

Digital Symptom Questionnaires Revealed Patterns of PMS Symptoms Across the Menstrual Cycle

PMS-related symptoms and their severity on a scale of 0 to 3 were only assessed on potential PMS days during the luteal phase and menses. The majority of participants (22/33, 66.7%) reported severe PMS symptoms on at least 1 day of the menstrual cycle for at least 2 cycles. Severe PMS was reported by 18.2% (n=6) of participants in at least one of the cycles recorded in our study, and only 15.2% (n=5) did not experience any severe PMS symptoms during the study period.

In terms of mental health, dejection (average score 0.5, SD 0.9) and seclusion (average score 0.3, SD 0.7) were among the most commonly reported symptoms during the menses. The data suggested a notable impact on participants’ mental health during the menses, with higher severity of seclusion (Mann-Whitney P=.03), dejection (Mann-Whitney P=.02), and irritability (Mann-Whitney P=.01) reported during the menses compared with the luteal phase (Figure S4 in Multimedia Appendix 2). Reported levels of anxiety, confusion, anger, and tenderness did not differ substantially between the luteal phase and menses (Figure S4 in Multimedia Appendix 2). In contrast to mental health symptoms, reported physical symptoms during menses were more severe. Participants experienced a significantly higher severity of bloating during the menses (mean 0.7, SD 1.1) compared with the luteal phase (mean 0.3, SD 0.7; Mann-Whitney P<.001; Figure 3). Similarly, the swelling was reported to be more severe during the menses (mean 0.3, SD 0.8; luteal phase: mean 0.1, SD 0.5; Mann-Whitney P<.001). In addition, participants reported more severe headaches during the menses compared with the luteal phase (mean 0.38, SD 0.7; luteal phase: mean 0.2, SD 0.67; Mann-Whitney P=.005) and more bothersome joint pain during the menses (mean 0.3, SD 0.8; luteal phase: mean 0.2, SD 0.5; Mann-Whitney P=.001).

In terms of reported weight gain, there was a trend toward increased weight gain during the menses, although this increase did not reach statistical significance (Figure S4 in Multimedia Appendix 2).
Combined Questionnaire and Wearable Data Delivered Interesting Insights Into the Menstrual Cycle

Participants provided a total of 23,424 symptom reports, with a total of 2468 symptoms reported in the daily questionnaires, which were administered during the luteal phase and menses. Heart rate sensor data was collected on a total of 2118 days, step count on 2153 days, and sleep time on 1996 days.

The average resting heart rate over the entire study period was 66.7 (SD 6.7). Participants slept an average of 7.3 (SD 1.6) hours while wearing the Fitbit device. With an average of 10,184 (SD 6120) steps per day, the FEMFIT cohort was quite physically active, with some participants walking >20,000 steps per day (153 days from 21 participants recorded with >20,000 steps; Figure S5 in Multimedia Appendix 2). When examining correlations between sensor data and questionnaire data on the menstrual cycle, we observed that mean resting heart rate appeared to be highest in the mid and late luteal phases relative to the time between menses and ovulation (Mann-Whitney $P<.001$; Figure 4, Figure S6 in Multimedia Appendix 2). We did not observe any significant differences in the change in step count (Mann-Whitney $P=.72$) or sleep time (Mann-Whitney $P=.58$) between the follicular and luteal phases (Figure 4).

In an exploratory analysis of the wearable data across participants with different PMS severity, participants with severe PMS ($n=22$) walked significantly fewer steps (mean daily steps 10,283, SD 6277) during the luteal phase and menses compared with participants with no or low PMS ($n=5$; mean 11,694, SD 6458; Mann-Whitney $P<.001$). There was no significant difference in sleep duration (Mann-Whitney $P=.18$) or heart rate with PMS severity (Mann-Whitney $P=.72$).

To investigate whether hormonal contraception had an effect on step count, sleep time, or heart rate, we compared 4 participants using hormonal contraception with 29 participants not using hormonal contraception. Participants on hormonal contraception appeared to be less active (mean step count 7400, SD 4433; no hormonal contraception: mean 9800, SD 6203; Mann-Whitney $P<.001$). The average daily resting heart rate was significantly higher in the hormonal contraceptive group (mean 68.2, SD 7.7 vs mean 66.6, SD 6.6; Mann-Whitney $P<.001$). There was no difference in sleep behavior between the 2 groups.
**Discussion**

**Principal Findings**

To date, only a few studies have assessed the menstrual cycle using both objective wearable and subjective questionnaire data [23,27,29,49,50]. Previous research has demonstrated potential effects of physical activity on the menstrual cycle and vice versa [30-32]. However, none of these studies used wearable devices to monitor daily activity levels, and there is a gap in studies using wearable data to examine sleep patterns across the menstrual cycle. Recognizing that participant retention is a significant challenge in digital cohort studies [19,40], our feasibility study was designed to primarily assess the effectiveness of study participation and adherence to the study protocol, highlighting challenges and opportunities for future large-scale studies in women’s health. Secondarily, we focused on potential biological outcomes.

Our cohort was a group of young, mostly digitally literate, cisgender women with regular cycle lengths. Overall, retention was good among participants who remained in the study beyond the second week. The observed dropout rate of 21.4% (n=9) within the first 2 weeks of the study is consistent with rates reported in other digital cohort studies [51,52], and the overall median retention of 81.0 days was remarkably high [52]. Nevertheless, these dropouts may have introduced selection bias. Among the 5 dropouts who provided demographic data, we did not identify any discernible demographic patterns in relation to retention. The mean age of the dropouts was 24.8 (SD 3.2) years, which was not significantly different from the rest of the cohort. None of the dropouts had preexisting medical conditions and only 1 reported menstrual irregularity. In future large-scale studies, it would be advisable to collect demographic information directly at recruitment (in our case when the wearable device was distributed) to ensure comprehensive data on dropouts. This may contribute to a better understanding of the potential retention patterns and help to address selection bias. Among those who dropped out early, technical barriers were the most commonly reported problem. Although our study team provided access to technical support, future studies could potentially reduce dropout rates by further reducing barriers to accessing technical support or by conducting active outreach.

The young age of participants, such as our cohort, is also known to be associated with lower retention [19,52]. Targeted engagement strategies such as increased outreach and communication with younger participants could counteract this [19]. For large-scale studies, continuous monitoring of data donations in an ongoing study can facilitate real-time strategies such as increased outreach to engage inactive participants, reduce dropouts, and ensure data completeness [19]. In 2 similar studies using sensor devices to investigate physiology surrounding the menstrual cycle [23,25], retention was generally not an issue, but retention efforts were not further highlighted in the publications.

In addition to the potential bias introduced by study dropouts, ensuring high data completeness is critical for investigating temporal patterns in menstrual cycles. Similar to our study, Maijala et al [23] reported issues with continuous data contribution across consecutive days. In particular, issues related to participant retention and data completeness could potentially undermine the integrity of the results. In our study, this was particularly challenging for sleep data. To mitigate these concerns, several strategies can be considered for future research. The comfort of wearable devices should be a priority in study design, as some participants in our study found the devices too uncomfortable. Data completeness may be further increased if a clear purpose for data collection is visible to participants, for example, having the clinical condition of interest in the study [52,53]. As this was a feasibility study without a specific disease focus, we expect retention and data completeness to be higher when applying a similar study setting.

![Average resting heart rate](image1.png)
![Step count](image2.png)
![Sleep time](image3.png)
to a clinical condition with a high disease burden, such as endometriosis [54]. Recruitment strategies should also be tailored to improve data completeness. Although participants in our study were highly digitally literate, the vast majority had never worn a wearable device before. Recruiting participants who already have a sensor device and are accustomed to wearing it on a daily basis may increase data completeness. However, such targeted recruitment would introduce additional bias, as sensor devices are mostly owned by individuals from high-income households [55]. Although some participants in our study felt stressed by wearing the sensor devices, positive prior experience with wearing sensor devices as a selection criterion for study participation may further increase data completeness. In our study, participants received the wearable device as a free gift at the start of the study, and the incentive was not dependent on their questionnaire completion rate or the duration of wearing the device. Future studies may benefit from offering incentives upon successful completion of the entire study, thereby encouraging full participation [52,56]. Increasing compensation beyond just the Fitbit device could further improve retention [52]. In terms of questionnaire data completeness, it may be beneficial to reduce the number of daily questions. In our study, completion rates for daily questionnaires were lower than for weekly questionnaires, and 1 of the 9 people who dropped out within the first 2 weeks reported that the study was too time-consuming. These strategies can potentially help to improve participant retention and ensure more comprehensive and reliable data.

Despite challenges related to data completeness, our study successfully demonstrated the technical feasibility and value of combining wearable and questionnaire data. The information derived from the digital questionnaires helped to map the sensor data to the respective day of the menstrual cycle and to correlate the objective physiological data with the subjectively reported symptom data from the study app, leading to interesting biological insights. The average resting heart rate measured by wrist-worn wearable sensors changed over the course of the menstrual cycle, with an increase in the mid and late luteal phases compared with the menses. This observation supports the findings of a larger study by Shilaih et al [25], where the pulse rate measured using laboratory standard photoplethysmography technology showed a peak in the midluteal phase. Another study with a smaller sample size similarly found increased heart rate (measured during sleep) in the mid and late luteal phases compared with menses and ovulation [24]. In our feasibility study, we also found that people with more severe PMS symptoms had lower daily step counts. Due to the observational nature of our study, we cannot establish causality. This finding may suggest that individuals with more severe PMS symptoms may exercise less due to the impact of the symptoms on their physical activity. Alternatively, this finding may be consistent with the existing literature, where increased physical activity has been found to reduce PMS symptoms [57-62]. In attempting to fill the research gap regarding sleep patterns across the menstrual cycle, we did not make any notable discoveries when comparing sleep duration between different phases of the menstrual cycle, across PMS severity, or when comparing individuals using hormonal contraception with those who did not. It is worth noting that there were more issues with data completeness for sleep data compared with daytime data, highlighting the need for improved retention strategies to motivate participants to wear the devices consistently at night. Although of very small sample size, the observed differences in heart rate and physical activity in participants using hormonal contraception (n=4) also provided an interesting starting point for future investigation. These observations show that sensor devices may offer a great opportunity to study physical changes over the menstrual cycle and suggest that further investigation on a larger scale may be valuable. Overall, it is important to interpret the biological findings of this feasibility study with caution, given the limitations of the small sample size and issues with data completeness.

Limitations

Apart from the interesting results of this feasibility study, this study has several limitations that should be taken into account when interpreting the findings.

First, the study was conducted as a feasibility study with a small sample size. Any biological interpretations made from such a small sample size must be taken with caution and merely represent an interesting starting point for future large-scale research. Our cohort was also not representative, recruited from only 1 recruitment site, and lacked ethnic diversity.

We further had limitations in data accuracy. The amount of wearable data we collected was limited to 3 aggregated parameters, which severely limited our ability for quality filtering. As we did not collect continuous wearable data throughout the day, but rather one aggregated data point for each measurement of interest, we had no objective insight into how long the device was worn each day and relied only on subjective reporting of wear time in the weekly questionnaires. Not wearing the device consistently or correctly could have resulted in inaccurate data.

Furthermore, the study relied on self-reported questionnaire data, which may introduce reporting bias and errors in data collection. In addition, the wrist-worn wearable device used in our study may not be as accurate as research-grade devices in tracking changes in physiological parameters related to the menstrual cycle. One must consider potential confounding factors that may affect the measurement accuracy of the sensor devices such as body sweat and unusual movements [63].

Regarding the medical accuracy of possible correlations between wearable data and specific phases of the menstrual cycle, the design of our study only allowed the day of ovulation to be estimated based on the total cycle length. In future studies, additional at-home ovulation tests could ensure greater accuracy in assessing the different phases of the menstrual cycle [25].

Conclusions

In conclusion, this study demonstrated that an app-based approach to collecting combined wearable and questionnaire data on the menstrual cycle is technically feasible and provides interesting biological insights. Study participants had a high level of digital literacy, which may have supported adherence to the study protocol. With 9 of 42 participants dropping out
within the first 2 weeks, engagement efforts at baseline would need to be improved in future, larger studies to ensure long-term adherence to the study protocol. To further improve the quality and applicability of such large-scale studies, challenges with data completeness need to be addressed. This could include using more frequent reminders for study participation, using more user-friendly wearable technology, or providing incentives to encourage participants to maintain consistent data contributions over the entire study. Although the sample size was small, we discovered an increase in average resting heart rate over the menstrual cycle and found differences in step count by PMS symptom severity in our cohort. Ultimately, using digital studies to research the menstrual cycle can be feasible and has the potential to improve our understanding of women's health and inform the development of personalized health care approaches.

Acknowledgments
The authors want to thank all feasibility study of menstrual cycles with Fitbit device (FEMFIT) participants for their contribution to women’s health research. The authors further want to acknowledge the work and dedication of all Data4Life employees working on this project. The authors especially want to thank the business intelligence (BI) and analytics team for helping with all inquiries concerning BI data and the development team for the technical implementation of the study. This study was fully funded by the Data4Life gGmbH.

Data Availability
The data sets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions
MF and CR initiated the study and wrote the study protocol. ALL conducted all data analysis on wearable and questionnaire data. IK supported data analysis. JR set up wearable data collection from Fitbit devices. RLB and MH recruited participants and handed out pre- and poststudy questionnaires. RLB evaluated the questionnaire responses to surveys conducted before and after the study. ALL wrote the first version of the manuscript. All authors have contributed to writing the manuscript.

Conflicts of Interest
ALL, IK, JR, CR, and MF worked for Data4Life, the company that developed the study and study app as well as the underlying research infrastructure. RLB declares no competing interests. MH received financial incentives from Data4Life for recruiting participants for this study.

Multimedia Appendix 1
Study questionnaires.
[XLSX File (Microsoft Excel File), 22 KB - mhealth_v12i1e50135_app1.xlsx ]

Multimedia Appendix 2
Supplementary figures.
[DOCX File , 618 KB - mhealth_v12i1e50135_app2.docx ]

References


22. Lang et al JMIR MHEALTH AND UHEALTH 2024 | vol. 12 | e50135 | p.287 https://mhealth.jmir.org/2024/1/e50135 (page number not for citation purposes)


Abbreviations

API: application programming interface
FEMFIT: Feasibility Study of Menstrual Cycles With Fitbit Device
MP: menses phase
OD: ovulation day
PMS: premenstrual syndrome

©Anna-Lena Lang, Rosa-Lotta Bruhn, Maya Fehling, Anouk Heidenreich, Jonathan Reisdorf, Ifrah Khanyaree, Maike Henningsen, Cornelius Remschmidt. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 12.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Investigating Citizens’ Acceptance of Contact Tracing Apps: Quantitative Study of the Role of Trust and Privacy

Grace Fox¹, PhD; Lisa van der Werff¹, PhD; Pierangelo Rosati², PhD; Theo Lynn¹, PhD

¹Irish Institute of Digital Business, Dublin City University, Dublin, Ireland
²J.E. Cairnes School of Business & Economics, University of Galway, Galway, Ireland

Corresponding Author:
Theo Lynn, PhD
Irish Institute of Digital Business
Dublin City University
Collins Ave, Dublin 9
Dublin
Ireland
Phone: 353 1 700 6873
Email: theo.lynn@dcu.ie

Abstract

Background: The COVID-19 pandemic accelerated the need to understand citizen acceptance of health surveillance technologies such as contact tracing (CT) apps. Indeed, the success of these apps required widespread public acceptance and the alleviation of concerns about privacy, surveillance, and trust.

Objective: This study aims to examine the factors that foster a sense of trust and a perception of privacy in CT apps. Our study also investigates how trust and perceived privacy influence citizens’ willingness to adopt, disclose personal data, and continue to use these apps.

Methods: Drawing on privacy calculus and procedural fairness theories, we developed a model of the antecedents and behavioral intentions related to trust and privacy perceptions. We used structural equation modeling to test our hypotheses on a data set collected at 2 time points (before and after the launch of a national CT app). The sample consisted of 405 Irish residents.

Results: Trust in CT apps was positively influenced by propensity to trust technology (β=.074; P=.006), perceived need for surveillance (β=.119; P<.001), and perceptions of government motives (β=.671; P<.001) and negatively influenced by perceived invasion (β=−.224; P<.001). Perceived privacy was positively influenced by trust (β=.466; P<.001) and perceived control (β=.451; P<.001) and negatively influenced by perceived invasion (β=−.165; P<.001). Prelaunch intentions toward adoption were influenced by trust (β=.590; P<.001) and perceived privacy (β=.247; P<.001). Prelaunch intentions to disclose personal data to the app were also influenced by trust (β=.215; P<.001) and perceived privacy (β=.208; P<.001) as well as adoption intentions before the launch (β=.550; P<.001). However, postlaunch intentions to use the app were directly influenced by prelaunch intentions (β=.530; P<.001), but trust and perceived privacy only had an indirect influence. Finally, with regard to intentions to disclose after the launch, use intentions after the launch (β=.665; P<.001) and trust (β=.215; P<.001) had a direct influence, but perceived privacy only had an indirect influence. The proposed model explained 74.4% of variance in trust, 91% of variance in perceived privacy, 66.6% of variance in prelaunch adoption intentions, 45.9% of variance in postlaunch use intentions, and 83.9% and 79.4% of variance in willingness to disclose before the launch and after the launch, respectively.

Conclusions: Positive perceptions of trust and privacy can be fostered through clear communication regarding the need and motives for CT apps, the level of control citizens maintain, and measures to limit invasive data practice. By engendering these positive beliefs before launch and reinforcing them after launch, citizens may be more likely to accept and use CT apps. These insights are important for the launch of future apps and technologies that require mass acceptance and information disclosure.

(JMIR Mhealth Uhealth 2024;12:e48700) doi:10.2196/48700

KEYWORDS

privacy; trust; public health surveillance; contact tracing; mobile apps; adoption; information disclosure
Introduction

Background

The outbreak of COVID-19 and the ensuing global pandemic resulted in many governments undertaking digital government transformation [1] through the introduction of public health surveillance technologies including contact tracing (CT) apps [2,3]. As a result, and unlike previous health emergencies, governments had access to an unprecedented volume, variety, and velocity of location and health data [4]. The use of such data for epidemiological surveillance can aid in decision support, accelerate case identification, interrupt community transmission, and enable public health communication [5]. Notwithstanding these benefits, the pace at which these apps have been implemented and the level of surveillance they enable have raised ethical concerns [6] and fears around privacy and public trust [7].

The success of CT apps is dependent on uptake by large populations [8], and privacy-related concerns have been positioned as a critical barrier facing government-introduced CT apps [9]. Government-introduced CT apps differ from surveillance technologies and mobile health (mHealth) apps, as they combine both location-based data and electronic personal health information (ePHI). Both these contexts, in themselves, raise significant privacy concerns, particularly with respect to potential secondary use and government intrusion [10]. Thus, the COVID-19 pandemic presents a unique empirical context to explore citizens’ perceptions of health surveillance using mobile apps that capture both location-based data and ePHI. Government-introduced CT apps constitute a new public health context. Although there is an established literature based on traditional CT, digital CT is an innovation that can only be fully explored during pandemics, and thus research opportunities are limited. Given the novel context for individuals and public health agencies, where the former engage with new or unfamiliar trust referents, it is particularly critical to explore how individuals marry competing beliefs about surveillance, trust, and government motives for introducing these technologies and how these beliefs influence their behavioral responses. Extant literature has demonstrated the importance of trust in the government in influencing CT adoption [11,12] and perceptions of CT technologies [13] and shown that privacy concerns represent a barrier to adoption [9].

Prior Work

This study builds upon important extant research focusing on the acceptance of CT apps to delve further into the role of privacy and trust and addresses 3 gaps in the literature. First, existing studies support the importance of trust in driving the acceptance of CT apps, but the approach to measuring trust and the trust referent under examination varies. For example, studies have found that high trust in the national government, the health care system, and science positively impacted willingness to use CT apps in Switzerland [14]. A US-based study found that trust in COVID-19 information positively influenced citizens’ comfort with and acceptance of CT [8]. In this study, our emphasis is on technology-related trust, as opposed to trust in an individual or organization. Specifically, we examined trust perceptions regarding a specific technology, namely a CT app. In addition, we investigated the perceptual factors that shape trust, namely perceptions of government motivations, need for surveillance, and propensity to trust technology (PTTT).

Second, studies have investigated the influence of privacy on CT adoption, with many studies finding that privacy concerns reduce intentions toward adopt CT apps [11,14]. Several studies conclude that privacy represents a barrier to the success of CT apps, with respondents in several countries citing privacy concerns as a reason for not installing apps [9] or conversely, those with low privacy concerns are more likely to use CT apps [12]. However, the influence of privacy concerns on CT adoption intentions was weak in another study [15]. Although privacy concerns are the most common proxy for measuring privacy across many contexts, the negative connotation and failure to directly capture privacy suggest the need for more precise operationalization [16]. Thus, we focus on perceived privacy defined as “an individual’s self-assessed state in which external agents have limited access to information about him or her” [16]. In this study, perceived privacy refers to a citizen’s belief in the level of privacy afforded by the CT app. Perceived privacy influenced intentions toward CT apps in a recent Brazilian study, thus supporting its use [17]. To further our understanding of perceived privacy, this study investigated the role of trust, perceived control, and perceived intrusion in shaping citizens’ perceptions of privacy.

Third, many CT studies are cross-sectional in nature, with the exception of a small number of longitudinal studies [12]. In addition, the literature focuses largely on whether citizens adopt CT apps or engage in behaviors recommended by CT apps such as staying at home [14,18]. This study broadens our understanding of citizens’ acceptance of CT apps by examining 2 variables related to acceptance, namely intention to adopt or continue using the app and willingness to disclose personal information. These acceptance variables are measured before and after the app launch, thereby deepening our understanding of how privacy and trust influence intentions toward and use of CT apps.

We argue that understanding the determinants of success of CT apps is critical not only for future digital CT but also for other contexts that require rapid digital technology adoption by the population [19]. This paper proceeds with an overview of the hypothesized relationships and our research context. Our methodology, data analysis, and results are presented in the following sections. The Discussion section outlines the implications of this study. The paper concludes with the limitations and avenues for future research.

Theory

Overview

Privacy Calculus Theory (PCT) posits that before engaging in a behavior such as adopting a new technology or disclosing personal information, individuals will conduct a cognitive comparison of the costs and benefits associated with this behavior [20]. Individuals are likely to engage in the behavior for as long as the benefits outweigh the costs [20]. Thus, PCT has direct comparisons with the concept of calculus-based trust,
which underpins trust decisions when engaging with new or unfamiliar trust referents [21]. PCT has been operationalized in many contexts using a variety of belief-based variables that can be grouped into confidence beliefs and risk or privacy beliefs. An extension of PCT that holds considerable promise for understanding privacy and trust in the context of CT apps is the inclusion of the procedural fairness theory. Procedural fairness refers to an individual’s perception that a particular activity is conducted fairly [22]. In the context of information privacy, fairness refers to the perception that personal data are collected and used fairly. Culnan and Armstrong [22] proposed that perceptions of procedural fairness can help citizens to “strike a balance between the competing forces of privacy and information use.” Individuals’ perceptions of the fairness of an organization’s data collection and use practices can influence their decision-making related to technology use and information disclosure [23]. In this study, we investigated the drivers of trust and privacy through the procedural fairness lens, which suggests the importance of factors related to the legitimacy of data collection (ie, the need for government surveillance and perceptions of the government’s motive for the app), the costs to the citizen (ie, perceived intrusion), and the level of autonomy and input citizens are afforded (ie, perceived control). Furthermore, the wider literature on fairness and trust suggests that individual differences in citizens’ PTTT are likely to play an important role alongside procedural fairness perceptions in trust [24], particularly given the unfamiliar referent of the CT app. In addition, the theory of reasoned action (TRA) allows us to consider the influence of these perceptions on behavioral outcomes. The TRA argues that individuals’ behaviors are determined by their beliefs, attitudes, and intentions [25]. We propose that individuals will express positive intentions toward downloading the app and disclosing information if they believe that the app demonstrates fairness.

**Hypotheses**

Trust is an important factor in the success of CT apps as demonstrated in recent studies [14]. Indeed, trust allows individuals to overcome concerns about uncertainty and fosters a willingness to engage in trust-related behaviors, such as disclosing information and engaging with technology [26]. In this study, our emphasis is on technology-related trust, which refers to individuals’ beliefs that the technology in question will perform as expected [27]. Trust perceptions refer to the extent to which the CT app will consistently deliver the proposed services and act in citizens’ best interest.

A significant body of theoretical work suggests that variables related to trust propensity are important drivers of trust perceptions [28], particularly in new and unfamiliar trust referents [29]. PTTT refers to a general tendency that is not specific to one trustee or situation but focuses on individuals’ willingness to depend on technology across different contexts and technologies [27]. We propose that, in the current context, PTTT will positively influence trust in CT apps.

- **Hypothesis 1a:** PTTT will have a positive association with trust in the app.

Surveillance programs are often introduced following large-scale events such as terrorist attacks [30]. Given the public health emergency caused by the COVID-19 pandemic, the importance of surveillance technologies, such as CT apps, is clear. The link between surveillance and trust has long attracted discussion with Dutton et al [31], highlighting the existence of trust tension between the government’s need to collect surveillance data and citizens’ concerns about the excessive use of this information. They assert that developing trust is imperative for resolving this tension. As CT apps require the surveillance of large groups of people [32], citizens must understand the need for government surveillance in the general sense to build trust in a CT app. Need for government surveillance refers to individuals’ perceptions that the government requires authority to access personal information using web-based means [33].

- **Hypothesis 1b:** need for surveillance will have a positive association with trust in the app.

Procedural fairness theory suggests that if individuals believe that the government’s motivations to introduce the app are rooted in good intentions, such as reducing virus transmission, they will express higher trust in the app to perform consistently and with their best interests in mind. Indeed, a perception of benevolent motives is at the heart of theories regarding trustworthiness [28], and perceptions of trustworthiness at the government level are likely to trickle down to influence trust in related referents [34]. Accordingly, we posit that government motive will engender trust in the app.

- **Hypothesis 1c:** government motive will have a positive association with trust in the app.

Government surveillance technologies can garner negative opinions, which may lead individuals to alter their behaviors. Perceived intrusion is described as a “harmful incursion into the personal information space” [35]. This relates to procedural justice and individuals’ perceptions of whether data are collected fairly in the CT app. Government surveillance technologies can be viewed as intrusive, but not all intrusions are considered harmful [35]. For example, the CT app may be viewed as intrusive, but as data are collected to reduce virus transmission, some individuals may not view this as harmful. Thus, only if individuals believe that the intrusiveness of the app is harmful to them, then their trust in the app is likely to be reduced.

- **Hypothesis 1d:** perceived intrusion will have a negative association with trust in the app.

Studies have asserted that privacy concerns represent a barrier to the success of CT apps [14,36]. However, there is a lack of research examining if citizens are likely to accept CT apps when they perceive that they provide some level of privacy. As perceived privacy refers to a perception that access to personal information by external agents is limited, the relevance of perceived intrusion as a privacy cost resulting from the use of an app is apparent. Indeed, the potential of CT apps to violate citizens’ privacy has been raised [36]. Thus, if individuals believe that the app is intrusive in a harmful manner in their informational space, they are less likely to believe that the app affords them privacy.

- **Hypothesis 2a:** perceived intrusion will have a negative association with perceived privacy.
Perceived control is described as individuals’ perceptions of their ability to control their personal information [37]. If individuals perceive that they maintain control over their information when transacting with a technology, they are more likely to feel comfortable that the technology will not act in a harmful manner [37], strengthening their perception of privacy in that context. A recent study supports the positive association between perceived control and perceived privacy of CT apps in Brazil [17]. We proposed that if individuals believe that they maintain control in the app, they will express higher levels of perceived privacy.

- **Hypothesis 2b:** perceived control will have a positive association with perceived privacy.

Finally, we argue that from a theoretical perspective, trust in the app will act as an uncertainty-reducing mechanism [38,39] and a heuristic that allows citizens to form privacy perceptions. Specifically, if citizens believe that the app will perform consistently and with their best interests in mind when using personal information, they will believe that the app provides some degree of privacy. This uncertainty reduction provides a foundation for facilitating other judgments of a technological artifact. Thus, trust in the app may influence perceptions of privacy.

- **Hypothesis 2c:** trust in the app will have a positive association with perceived privacy.

The success of CT apps is largely dependent on a critical mass of people downloading [40] and disclosing personal information. Thus, we draw on the TRA to consider 2 context-critical dependent variables: adoption intention and willingness to disclose personal information. Before the app launch, these variables were behavioral intentions. Adoption intention is described as an individual’s internal subjective judgment of the probability that they will perform the behavior in question [25]. The willingness to disclose information is based on an individual’s willingness to provide personal information when using the app [33]. Trust has repeatedly been identified as a driver of behavioral intentions across a range of contexts, including the acceptance of e-government technologies [41] and surveillance [30]. Indeed, behavioral operationalizations of trust often use behavioral intentions related to disclosure and reliance [42]. Empirical evidence suggests that perceived privacy influences intentions toward the use and data disclosure in general CT apps [17]. Although our study focuses on the privacy perception of a government-led CT app as opposed to general CT apps, we argue for similar effects. Finally, if individuals express high intentions toward downloading the app, we argue that they will be more willing to disclose personal information, as it is crucial to the app’s functionality.

- **H3a-b:** trust in the app before the launch (a) and perceived privacy (b) will have a positive association with adoption intentions before the launch.
- **H4a-c:** trust in the app before the launch (a), perceived privacy (b), and adoption intentions before the launch (c) will have a positive association with disclosure intentions before the launch.

There have been calls for research to understand the perceptions of a technology before and after the launch [43]. In the context of CT apps, it is important to explore how both perceptions of privacy and trust influence individuals’ behavioral intentions after launch. Thus, we examined both intentions after the launch. As some individuals may have already downloaded the app, adoption intentions are represented by future use intentions, which encompasses intentions to continue use among app users and intentions to adopt in the future among nonusers. TRA asserts that intentions will lead to behavior [44]. In other words, individuals’ intentions to download the app before the launch will be positively related to their use intentions after the launch. We draw on the TRA to posit effects similar to those hypothesized for before the launch. We argue that trust perceptions regarding the app and perceived privacy will positively impact use intentions after the launch.

- **H5a-b:** trust in the app (a) and perceived privacy (b) will have a positive association with intentions to use after the launch.
- **H5c:** adoption intentions before the launch will have a positive association with intentions to use after the launch.

The influence of trust and privacy on the willingness to disclose a specific surveillance technology after the launch has not been explored. Again, we leverage the TRA intention-behavior link and our hypotheses before the launch and argue that trust and perceived privacy will positively impact willingness to disclose personal information after the launch. Finally, as was the case before the launch, we argue that if individuals express high intentions toward using the app after the launch, they will be more willing to disclose personal information.

- **H6a-b:** trust in the app (a) and perceived privacy (b) will have a positive association with disclosure intentions after the launch.
- **H6c:** intentions to use after the launch will have a positive association with disclosure intentions after the launch.

The hypotheses are depicted in **Figure 1**.
Methods

Study Context

On July 6, 2020, the COVID Tracker app was launched in Ireland. COVID Tracker has 3 main features. CT uses Bluetooth and anonymous ID numbers to log phones within close contact for more than 15 minutes. It downloads the anonymous ID numbers of people who have tested positive and provides an alert if the user has been in close contact with those ID numbers. Check-in allows users to check for symptoms and seek health advice. The updates provide an overview of the daily COVID-19 figures. Within 48 hours of launch, 1 million people had downloaded the app [45]. By November 2021, the COVID Tracker app had over 1.7 million active users, representing 67% to 76% of the total possible Irish users [46].

Instrument Development

We used existing scales when developing our instrument with minor wording amendments to adapt the items to the context. We provide the full list of items in Table S1 in Multimedia Appendix 1. The survey at T1 included general constructs related to PTTT, the need for government surveillance, and situationally framed constructs related to the proposed app, namely government motive for introducing the app, trust in the app, perceived surveillance in the app, perceived control in the app, and perceived privacy in the app. The dependent variables included intention to download the app on launch and willingness to disclose personal information to use the app. At T2, the emphasis was on future behavioral intentions, namely the use of the app and information disclosure. At T2, we asked participants if they had downloaded the app, and app users’ intentions to continue to use the app were examined and nonusers’ intentions to download the app in the future were examined. These intentions were combined as use intentions for analysis. Willingness to disclose personal information to the app was examined across both users and nonusers of the app. In addition, gender, age, and education were used as the control variables. Respondents were asked if they had any health condition that left them particularly susceptible to COVID-19. This was also a control variable. Both surveys were piloted and tested among a small panel of survey design experiments, and several wording amendments were made. Respondents were asked to answer demographic and health questions first, followed by general perceptual constructs and control variables, the order of which was randomized. In the third section, participants were presented with a neutrally framed description of the proposed national CT app at T1, and a description of the launched app was presented at T2. The final section examined perceptions of the app, behavioral intentions, and willingness to disclose personal data, the order of which was randomized.
Ethical Considerations
Ethics approval was obtained from the university’s research ethics committee before the launch of the survey (DCUREC/2020/096).

Recruitment
Qualtrics (Qualtrics International Inc) was used to host and administer the survey using their panel services. An attention check was included to screen for unengaged responses. A total of 1109 complete responses were received at T1 and were recontacted at T2. After 2 follow-up invitations, 405 responses were received at T2, achieving a response rate of 36.5%. Responses at T1 and T2 were, on average, 77 days apart. Incomplete responses and responses failing the attention check were removed using Qualtrics. The sample characteristics are illustrated in Table 1, along with the population characteristics as per the latest census at the time of data collection. Overall, the sample demographics were similar to the population characteristics of Ireland, as reported in the last census, and included respondents from the 26 counties within the country. Of the 405 respondents in T2, 202 had downloaded the app and 203 had not yet downloaded the app.

Table 1. Sample and population characteristics (N=405).

<table>
<thead>
<tr>
<th>Sample, n (%)</th>
<th>Population (%)(^\text{a,b})</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
</tr>
<tr>
<td>Man</td>
<td>180 (44.4)</td>
</tr>
<tr>
<td>Woman</td>
<td>225 (55.6)</td>
</tr>
<tr>
<td>Rather not say</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>Age group (years)</strong></td>
<td></td>
</tr>
<tr>
<td>18-24</td>
<td>13 (3.2)</td>
</tr>
<tr>
<td>25-44</td>
<td>124 (30.6)</td>
</tr>
<tr>
<td>45-64</td>
<td>173 (42.7)</td>
</tr>
<tr>
<td>≥65</td>
<td>95 (23.5)</td>
</tr>
<tr>
<td><strong>Employment</strong></td>
<td></td>
</tr>
<tr>
<td>Employed</td>
<td>186 (45.9)</td>
</tr>
<tr>
<td>Self-employed</td>
<td>26 (6.4)</td>
</tr>
<tr>
<td>Unemployed</td>
<td>36 (8.9)</td>
</tr>
<tr>
<td>Student</td>
<td>11 (2.7)</td>
</tr>
<tr>
<td>Unavailable for work</td>
<td>42 (10.4)</td>
</tr>
<tr>
<td>Retired</td>
<td>104 (25.7)</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
</tr>
<tr>
<td>Secondary school</td>
<td>157 (38.8)</td>
</tr>
<tr>
<td>Trade</td>
<td>5 (1.2)</td>
</tr>
<tr>
<td>Diploma</td>
<td>32 (7.9)</td>
</tr>
<tr>
<td>Bachelor degree</td>
<td>133 (32.8)</td>
</tr>
<tr>
<td>Other qualification</td>
<td>64 (15.8)</td>
</tr>
<tr>
<td>Doctorate degree</td>
<td>14 (3.5)</td>
</tr>
</tbody>
</table>

\(^{a}\)Population figures are based on data provided by the Irish Central Statistics Office in the latest population census at the time of data collection (ie, 2016).
\(^{b}\)Employment and education figures include all people aged ≥15 years living in Ireland in 2016, whereas our sample only includes people aged ≥18 years.

\(^{c}\)N/A: not applicable.

Results
Reliability and Validity Testing
Data analysis was performed using IBM AMOS (version 25.0). The proposed model comprising 11 constructs was examined using Confirmatory Factor Analysis with further detail provided in Table S2 in Multimedia Appendix 1. In total, 3 items were dropped from the PTTT because of their low loadings. The model indicated a good fit: \(c_{\text{min}}/df=1.805\), comparative fit index=0.980, root mean square error of approximation=0.045, and standardized root mean squared residual=0.034. A test of equal specific bias was conducted to examine potential common method bias among the data [47]. This test demonstrated an unevenly distributed bias; thus, the specific bias construct was
retained for causal analysis to control for any effects because of method [48]. The validity and reliability of all the constructs were explored. Convergent validity was assessed by calculating the average variance extracted (AVE). As all the variables had AVE scores above 0.500, convergent validity was achieved [49]. Discriminant validity was tested by comparing the square root of the AVE with the interconstruct correlations. As the square root of AVE was higher than the interconstruct correlations, discriminant validity was achieved, as shown by the italicized diagonal values in Table 2. Reliability was assessed by calculating composite reliability for each construct. With composite reliability scores above 0.700, all constructs were reliable [50]. Further details on the validity testing are provided in the supplementary appendices available on the web.

### Table 2. Validity and reliability statistics.

<table>
<thead>
<tr>
<th>Construct</th>
<th>Composite reliability</th>
<th>Average variance extracted</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Need for Surveillance</td>
<td>0.868</td>
<td>0.526</td>
<td>0.725&lt;sup&gt;a&lt;/sup&gt;</td>
<td><em>b</em></td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Propensity to trust technology</td>
<td>0.870</td>
<td>0.626</td>
<td>0.183&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.791</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Perceived control</td>
<td>0.967</td>
<td>0.879</td>
<td>0.320&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.210&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.938</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Willingness to disclose information (T1)</td>
<td>0.983</td>
<td>0.966</td>
<td>0.378&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.271&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.682&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.983</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Intention to adopt (T1)</td>
<td>0.990</td>
<td>0.970</td>
<td>0.342&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.270&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.666&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.872&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.985</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Perceived intrusion</td>
<td>0.932</td>
<td>0.820</td>
<td>-0.181&lt;sup&gt;d&lt;/sup&gt;</td>
<td>-0.109&lt;sup&gt;d&lt;/sup&gt;</td>
<td>-0.471&lt;sup&gt;d&lt;/sup&gt;</td>
<td>-0.508&lt;sup&gt;d&lt;/sup&gt;</td>
<td>-0.309&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.906</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Trust in App</td>
<td>0.916</td>
<td>0.786</td>
<td>0.365&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.336&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.702&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.814&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.784&lt;sup&gt;d&lt;/sup&gt;</td>
<td>-0.491&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.886</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Perceived Privacy in App</td>
<td>0.962</td>
<td>0.895</td>
<td>0.371&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.259&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.852&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.801&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.753&lt;sup&gt;d&lt;/sup&gt;</td>
<td>-0.603&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.852&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.946</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Intention to adopt or use (T2)</td>
<td>0.991</td>
<td>0.973</td>
<td>0.234&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.167&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.500&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.612&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.658&lt;sup&gt;d&lt;/sup&gt;</td>
<td>-0.034&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.574&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.548&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.986</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Willingness to disclose information (T2)</td>
<td>0.985</td>
<td>0.970</td>
<td>0.309&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.241&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.537&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.678&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.666&lt;sup&gt;d&lt;/sup&gt;</td>
<td>-0.401&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.686&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.646&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.844&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.985</td>
<td>---</td>
</tr>
<tr>
<td>Government motive</td>
<td>0.971</td>
<td>0.894</td>
<td>0.246&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.332&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.550&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.682&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.654&lt;sup&gt;d&lt;/sup&gt;</td>
<td>-0.378&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.799&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.658&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.494&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.575&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.946</td>
</tr>
</tbody>
</table>

<sup>a</sup>As the square root of AVE was higher than the interconstruct correlations, discriminant validity was achieved, as shown by the italicized values.
<sup>b</sup>Not available.
<sup>c</sup>Significance at 10% level.
<sup>d</sup>Significance at 1% level.
<sup>e</sup>Significance at 5% level.

### Hypotheses Testing

The causal model was tested using Structural Equation Modeling in AMOS. The model indicated a good fit $c_{min}/df=2.835$, comparative fit index=0.985, root mean square error of approximation=0.067, and standardized root mean squared residual=0.021. H1a-d focused on the antecedents of trust in the app. H1a proposed a positive relationship between PTTT and trust. The data revealed a positive, significant relationship supporting hypothesis 1a ($\beta=0.074$; $P=0.006$). Hypothesis 1b posited that the perceived need for government surveillance would positively influence trust. This was also supported (hypothesis 1b: $\beta=0.119$; $P<0.001$). H1c posited that government motive would be positively related to trust. The data supported the hypothesis (hypothesis 1c: $\beta=0.671$; $P<0.001$). The negative relationship between perceived intrusion and trust was supported (hypothesis 1d: $\beta=-0.224$; $P<0.001$). The following set of hypotheses examined the antecedents of perceived privacy. Hypothesis 2a proposed a negative association between perceived intrusion and perceived privacy. These data supported hypothesis 2a ($\beta=-0.165$; $P<0.001$). We hypothesized that perceived control and trust would be positively related to perceived privacy. Both relationships were supported (hypothesis 2b: $\beta=0.451$; $P<0.001$; H2c: $\beta=0.466$; $P<0.001$).

In terms of T1 behavioral intentions, it was posited that perceived privacy and trust would positively influence the intention to adopt. Both relationships were supported (hypothesis 3a: $\beta=0.247$; $P<0.001$; hypothesis 3b: $\beta=0.590$; $P<0.001$). H4a-c proposed that trust, perceived privacy, and adoption intentions would positively influence willingness to disclose personal information. The data revealed that trust ($\beta=0.215$; $P<0.001$), perceived privacy ($\beta=0.208$; $P<0.001$), and adoption intentions ($\beta=0.550$; $P<0.001$) each positively influenced willingness to disclose. In terms of T2 behavioral intentions, hypothesis 5a to 5c proposed that trust, perceived privacy, and adoption intentions (T1) would all influence use intentions (T2). Both trust ($\beta=0.124$; $P=0.15$) and perceived privacy ($\beta=0.042$; $P=0.60$) had a positive but nonsignificant influence on intentions. T1 adoption intentions significantly influenced use intentions.
supporting hypothesis 5c (β=.530; P<.001). Finally, trust, perceived privacy, and use intentions at T2 were proposed to positively influence the willingness to disclose at T2. Perceived privacy (β=.042; P=.40) had a nonsignificant influence, whereas trust and use intentions had significant relationships supporting hypothesis 6a and hypothesis 6c (hypothesis 6a: β=.250; P<.001; H6c: β=.655; P<.001). In terms of control variables, COVID-19 vulnerable illness had a significant negative effect on individuals’ willingness to disclose at T1 (β=.131; P=.001) and willingness to disclose at T2 (β=.127; P=.04). Similarly, trust had a significant influence on intention to download (β=.394; P<.001) and willingness to disclose at T2 (β=.386; P<.001). Further detail is provided in Table S3 in Multimedia Appendix 1.

Discussion

Principal Findings

This study focuses on understanding how citizens’ beliefs shape their perceptions of privacy and trust to influence their acceptance of a CT app for COVID-19. Our study found that trust in the app was positively influenced by the PTTT, perceived government motive, and perceived need for government surveillance, whereas perceived intrusion had a negative influence. Perceived privacy was positively shaped by perceptions of control and trust and negatively shaped by perceived invasion. The study examined citizens’ acceptance of CT app at 2 time intervals. Before launch, the intention to adopt the app was positively influenced by trust and perceived privacy, and willingness to disclose personal information to the app was influenced by adoption intentions, trust, and perceived privacy. However, postlaunch use intentions were influenced only by prelaunch adoption intentions, whereas willingness to disclose personal information was influenced by trust and postlaunch use intentions but not by perceived privacy. Although the insignificant results may suggest that perceived privacy is only important before launch, and the influence of trust on use intentions diminishes over time, post hoc bootstrapping analysis revealed that both perceived privacy and trust had significant indirect relationships with use intentions and willingness to disclose information at T2. This suggests that both perceptions play a role in influencing behavioral intentions before and after the launch.

Contributions

Studies have shown that trust and privacy are important factors in the success of health surveillance technologies such as CT apps [11]. Our study leverages the procedural fairness theory to understand how citizens’ perceptions of trust and privacy emerge in the context of a CT app. This context is interesting, as the technology in question was introduced by the national government and backed by several organizations with the app’s potential benefits extending to the public at large. Therefore, it is important to look beyond the role of a single organization in driving perceptions of fairness to consider a broader set of antecedents that drive perceptions of trust and privacy in this context. Indeed, as research has shown the importance of trust and privacy in the success of mHealth and health technologies introduced by health care organizations and indeed national health systems [51], our study contributes to the broader health technology literature.

The first contribution of our study is the deeper understanding of how trust is formed in this context. Lack of trust in the government has been identified as a barrier to CT app adoption [18]. Thus, it is important to provide governments and public health organizations with insights into how trust in CT can be fostered [8]. Our findings bolster assertions regarding the important role of fairness perceptions and suggest that citizens’ trust perceptions regarding the app are formed based on their beliefs about the legitimacy of data collection, perceived autonomy, and perceived costs. Legitimacy is represented by citizens’ perceptions of the need for government surveillance and perceptions of the government’s motive for introducing the app, autonomy is captured by perceptions of control over one’s information in the app, and perceived costs to the individual relate to perceptions of personal intrusion.

The second contribution of our study is the investigation of how perceptions of privacy are formed. Many studies have highlighted the negative impact of privacy concerns on CT adoption [11,52-54]. However, we argue that privacy can be seen as a factor driving adoption if citizens believe that apps can afford them with some level of privacy. Our findings demonstrate that citizens’ privacy perceptions are shaped by trust in the app, which encompasses legitimacy perceptions and their perception of control offered by the app, and are negatively influenced by perceptions of intrusion. By highlighting the importance of fairness perceptions and elucidating the role of several perceptual factors at the governmental level (need for surveillance and government motives) and the app level (perceived intrusion and perceived control), which have been sparingly studied to date, our study advances our understanding of how privacy perceptions are developed in this context.

Understanding the factors driving CT app acceptance is paramount for future outbreaks [8]. The third contribution relates to understanding how citizens form intentions toward CT apps at different stages of the implementation process. Our study provides support for the influence of perceived privacy on individuals’ intentions to adopt an app and willingness to disclose personal information before launch and support for an indirect influence over time on future use and willingness to disclose data. This finding supports many studies that found that privacy concerns negatively impacted adoption intentions toward mHealth [51]. In the context of a national CT app, if individuals perceive that the app offers a sufficient state of privacy, they will express positive intentions toward adoption and information disclosure before and after the app launch.
simply, perceived privacy can have a sustained positive influence on behavioral intentions.

Trust has been widely studied within the privacy and other domains to understand individuals’ intentions to disclose information [42]. Trust in the app was found to influence individuals’ adoption and disclosure intentions before launch, indirectly influencing use intentions after launch and directly influencing disclosure intentions after launch. These findings suggest that the influence of initial trust perceptions prevails over time and may operate as a heuristic for interacting with the app on an ongoing basis. The stability of trust perceptions and their ongoing influence are a relatively nascent topic, although some theorists have suggested the possibility of trust as a heuristic [55]. Our research provides empirical evidence for this phenomenon and offers further support for claims that initial trust perceptions might be relatively robust and long acting [56].

Implications for Practice

The findings of this study have several practical implications. First, the trust tension between public good and the intrusiveness of surveillance technologies has led researchers to emphasize the importance of effective trust-building strategies when introducing surveillance programs [31]. Indeed, citizens in the United States and Germany have expressed concerns regarding possible surveillance stemming from CT apps [9]. Our study shows that citizens’ perceptions of trust and privacy can be influenced by fairness perceptions based on their beliefs regarding the need for surveillance and the government’s motives for introducing surveillance technologies, the perceived control they are offered over their personal information and negatively influenced by their perceptions of the intrusiveness of these technologies. Thus, governments should focus on transparency in their public health surveillance efforts, including the involvement of data protection authorities and civil liberties advocates throughout the project life cycle, potentially through a privacy advisory committee [57]. This transparency should be extended to communication with citizens on the need and purpose of a technology while stressing the control they have over their personal information. Our research suggests that early communications that shape first impressions are particularly important. Such practices not only comply with data protection laws, such as the EU (European Union) General Data Protection Regulation, but also foster a sense of trust and ultimately influence the use of technology.

Second, our findings highlight the positive influence of privacy perceptions on adoption and disclosure. Thus, we argue that privacy should not be viewed as a barrier to new technologies, such as mHealth or CT apps, but rather as an important consideration throughout the design, implementation, and postlaunch stages. Designers should ensure compliance with the regulatory requirements for consent and control. Governments and other organizations charged with introducing new technologies should ensure that they clearly communicate their compliance with regulations and the considerations of individuals’ personal data. Given that CT apps provide data on the location, copresence, and potentially ePHI of not only the focal person but also others that they have been in contact with, the principles of both necessity and proportionality would appear to be key. As per Lenca and Vayena [4], data collection must (1) be proportional to the seriousness of the public health threat, (2) be limited to what is necessary to achieve a specific public health objective, and (3) be scientifically justified. Policy makers and public health decision makers need to consider what communication and control mechanisms can be introduced to (1) build trust with the public and (2) repair trust, if necessary. This includes declaring what data will be collected and used while the app is live and by whom, confirming that data have been deleted, when no longer relevant (as is the case with COVID-19 data) or once the app is no longer required.

Third, in the context of technologies that require mass acceptance and willingness to share personal data, the focus cannot be placed solely on the number of downloads but must account for actual use and disclosure behaviors. Individuals’ intentions to download CT apps influence their willingness to disclose personal information both before and after launch. Once they have downloaded the app, it is critical that decision makers encourage use and that the widespread use of the app is linked, through public communication, to successful intervention strategies so that the benefits to the individual and society are reinforced.

Limitations and Directions for Future Research

This study has several limitations. First, other factors may influence privacy and trust perceptions or moderate the relationships between trust and privacy and adoption. Although it is not possible to consider all potential antecedents and intervening variables, it would be interesting to explore the role of other prominent perceptions, such as perceived sensitivity, as these apps require users to disclose identifying information, health information, and location information, all of which are arguably sensitive. Second, although our study considers 2 important technology use outcomes, before and after the launch, this approach has limitations. First, the collection of data from the same respondents at multiple time points inevitably led to a drop in responses. Although we sent repeated invitations during the second phase of data collection, the final sample that completed both surveys was smaller. Although this is commonplace within this approach and our sample characteristics are similar to the broader population of Ireland, we acknowledge that a large sample would be ideal and stress the importance of considering the sample size when drawing inferences from our study. Second, our 2 time points did not allow us to take full advantage of the potential to model longitudinal change trajectories over time. Further work is needed to incorporate time more fully into our understanding of how privacy and trust influence adoption and use behaviors.

Third, our study relies on individuals’ self-reported adoption and disclosure intentions. This approach is commonplace in the privacy and technology adoption literature streams, and it would not have been possible to study actual behaviors. However, we must acknowledge that intentions are not always matched with behaviors and that information disclosure is not always accurate or true. In other settings, it may be more feasible to collect more objective behavioral data, and we would encourage researchers to do so, particularly in settings where widespread adoption is
required for success. In addition, studies may go beyond our focus to understand disclosure behaviors at a deeper level and examine privacy-protective behaviors, such as withholding information or falsifying information. These protective behaviors are potentially dangerous in contexts such as CT apps because of the reliance on accurate data to track virus transmission.

Finally, our study explores a public health surveillance context where the focal person volunteers to participate and therefore has notice of the surveillance, control of their data and gives explicit consent. There are several conditions under which public health surveillance, including name reporting, may be undertaken without notice or explicit patient consent with well-established justifications in public health ethics, science, and law [58]. Even in the context of COVID-19, digital CT has not always been voluntary. In China, there is evidence of digital CT without notice or consent [59]. Furthermore, even when the focal person has notice and gives consent, contacts of the focal person have not given explicit consent. Although the primary focal person is subject to direct active surveillance, the secondary focal person is subject to passive indirect surveillance. In addition to the ethical issues that such practices raise, particularly where there is coordination and data exchange between private firms and the government [60], testing the theoretical framework developed in this study in this new context may provide a fruitful avenue of research. Similarly, aggregated anonymous spatiotemporal data sourced from commercial providers have been used as proxies for human movement and social interaction and as indicators of the effectiveness of social distancing interventions [61]. Although these data are currently anonymous, governments have already mandated access to identifiable data on the basis that the public interest overrides privacy rights [62]. This context may provide interesting insights and further extend our understanding of the limits of consumer acceptance of governmental health surveillance.

Conclusions

The COVID-19 pandemic was the first time governments implemented large-scale digital CT. Its success as a public health intervention depended on rapid technology adoption by a significant proportion of the population. Here, surveillance is active, and the target of government surveillance through COVID-19 digital CT apps is an active participant in sharing data with the government on their personal health status, location, and often their social network. The opportunity to study such an empirical context is not only rare but also the time frame for research is limited. Understanding the formation of individuals’ perceptions of trust and privacy in this context and how these perceptions influence their acceptance of digital CT apps is critical not only for informing the design of future digital CT initiatives but also for other situations that require rapid digital technology adoption by a significant proportion of society. If governments wish to leverage the power of digital technologies to control future public health threats, we recommend 3 principles to guide the design of both their surveillance initiatives and communications with the public—necessity, transparency, and proportionality—before and after the launch.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Survey items, validity, and reliability testing.

References


Abbreviations

- AVE: average variance extracted
- CT: contact tracing
- ePHI: electronic personal health information
- EU: European Union
- mHealth: mobile health
- PCT: Privacy Calculus Theory
- PTTT: propensity to trust technology
- TRA: theory of reasoned action

©Grace Fox, Lisa van der Werff, Pierangelo Rosati, Theo Lynn. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 18.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Investigating the Integration and the Long-Term Use of Smart Speakers in Older Adults’ Daily Practices: Qualitative Study

Fangyuan Chang¹, PhD; Lin Sheng¹, MEng; Zhenyu Gu¹, PhD
Interaction Design Lab, School of Design, Shanghai Jiao Tong University, Shanghai, China

Corresponding Author:
Zhenyu Gu, PhD
Interaction Design Lab
School of Design
Shanghai Jiao Tong University
800 Dongchuan Road
Shanghai, 200240
China
Phone: 86 13167232872
Email: zygu@sjtu.edu.cn

Abstract

Background: As smart speakers become more popular, there have been an increasing number of studies on how they may benefit older adults or how older adults perceive them. Despite the increasing ownership rates of smart speakers among older adults, studies that examine their integration and the long-term use in older adults’ daily practices are scarce.

Objective: This study aims to uncover the integration of smart speakers into the daily practices of older adults over the long term, contributing to an in-depth understanding of maintained technology use among this demographic.

Methods: To achieve these objectives, the study interviewed 20 older adults who had been using smart speakers for over 6 months. These semistructured interviews enabled participants to share their insights and experiences regarding the maintained use of smart speakers in the long term.

Results: We identified 4 dimensions of the long-term use of smart speakers among older adults, including functional integration, spatial integration, cognitive integration, and semantic integration. For the functional integration of smart speakers, the study reported different types of use, including entertainment, information collection, medication reminders, companionship, environment modification, and emergency calls. For the spatial integration of smart speakers, the study showed older adults’ agency in defining, changing, and reshaping daily practices through the spatial organization of smart speakers. For the cognitive integration of smart speakers, the findings showed the cognitive processes involved in adapting to and incorporating smart speakers into daily habits and routines. For the semantic integration of smart speakers, the findings revealed that older adults’ enjoyable user experience and strong bonds with the device contributed to their acceptance of occasional functional errors. Finally, the study proposed several suggestions for designers and developers to better design smart speakers that promote maintainable use behaviors among older adults.

Conclusions: On the basis of the findings, this study highlighted the importance of understanding how older adults use smart speakers and the practices through which they integrate them into their daily routines. The findings suggest that smart speakers can provide significant benefits for older adults, including increased convenience and improved quality of life. However, to promote maintainable use behaviors, designers and developers should consider more about the technology use contexts and the specific needs and preferences of older adults when designing these devices.

(Keywords: smart speaker; private home; older adults; long-term use; daily practices; smart speakers)
**Introduction**

Along with the advancement of speech technology and artificial intelligence, smart speakers such as Google Home and Amazon Echo are becoming integral to households [1]. Equipped with smart voice assistants such as Google Assistant and Amazon Alexa, these devices respond to voice commands, facilitating activities like playing music, answering questions, setting reminders, and controlling smart home appliances. Their speech input and output features enhance accessibility, especially for individuals with limited mobility and vision [2,3].

In the realm of health care, the adoption of smart speakers has opened avenues for significant advancements, akin to the transformative impact of mobile phones. These devices present unique advantages for health research in out-of-hospital environments, offering opportunities for chronic disease management, passive identification of medical emergencies, detection of behavioral and cognitive changes, and remote monitoring of respiratory diseases impacting public health [4,5]. A range of successful pilot studies has demonstrated the positive impact of smart speakers on users requiring social care, and adults with learning differences, showcasing potential cost savings and improved well-being [6,7].

Despite the potential benefits, scholars have highlighted a notable gap in knowledge concerning older adults’ experiences with smart speakers in long-term use, which poses challenges to assessing the long-term impact and implications of these devices on the well-being and quality of life of older adults. Specifically, while research has explored smart speaker use among various demographics, such as low-income populations [8,9], people with disabilities [10,11], parents [12,13], and children [14], studies specifically focusing on the older adult demographic remain relatively scarce [3,15]. Among the limited studies, the majority of the existing literature concentrates on the design of smart speakers for later life, encompassing discussions on effective conversational cues [16], privacy concerns [17], and the anthropomorphism of speakers [18]. Other studies have assessed the feasibility and usability of smart speakers for promoting active aging [19,20], investigated older adults’ first impressions of smart speakers [19,21], and identified influential factors regarding older adults’ attitudes toward smart speakers [22]. The most recent studies emphasized there is a missing knowledge about older adults’ experiences with smart speakers in long-term use. This knowledge gap hinders the development of tailored interventions and policies to maximize the benefits of smart speakers for this demographic, ensuring their inclusion in the digital revolution and promoting healthy aging in the digital era [23-25]. Consequently, there is a pressing need for comprehensive exploration into older adults’ long-term experiences with smart speakers, emphasizing their perceptions, attitudes, and the evolving nature of their interactions with this technology.

This study aims to understand how older adults routinely use smart speakers and integrate them into their daily lives. The outcomes can add to the emerging body of literature for a more comprehensive understanding of older adults’ use of smart speakers in the long term. We conducted semistructured interviews with 20 older adults who used the smart speaker for over half a year, to answer the following research questions: “How does the use of smart speakers integrate into older adults’ daily practices?” and “What design considerations can be generated from the long-term use of smart speakers by older adults?”

In summary, our study embarks on an exploration into the long-term experiences of older adults with smart speakers, filling a notable gap in the existing knowledge. While previous research has delved into aspects such as conversational cues, the feasibility of smart speakers for active aging, and the before-and-after adoption changes or the broader impact, the focus on the nuanced, day-to-day interactions of older adults with these devices over an extended period is a novel dimension that sets our study apart [26]. By unraveling the complexities and diverse experiences during older adults’ long-term engagement with smart speaker technology, we aim to contribute not only to the effective design of age-friendly devices but also to the broader discourse on the role of technology in promoting well-being among older populations.

**Methods**

**Study Design**

A qualitative research study was conducted to investigate the experiences of older adults’ long-term use of smart speakers in daily practices. Semistructured interviews were used to facilitate an in-depth exploration of older adults’ experiences, even when the study involved a relatively small number of participants [27,28]. The study adhered to the guidelines and reporting standards outlined in the Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist for qualitative studies [29].

**Ethical Considerations**

Ethical approval for the research was obtained from the research ethics committee of the Shanghai Jiao Tong University (H20223351). Informed consent was obtained from all participants.

**Participants and Recruitment**

The research was conducted in Shanghai, China. To ensure diverse perspectives, participants were recruited from 2 distinct communities of older people—one situated in a bustling urban environment and the other in a rural area on the outskirts of the city. In the urban community, the researchers partnered with a community coordinator who had established relationships with older adults. In the rural community, the research team collaborated with the head of the older people’s community, who was responsible for arranging community activities. The coordinator and the community head served as a liaison between the research team and the two communities to help in identifying suitable candidates.

The specific criteria for participant inclusion in the study remained consistent across both communities, requiring individuals to be aged 60 years or older and to have a minimum of 6 months of experience using a smart speaker in their homes. The time limit served to ensure that current practices of using
smart speakers were represented and to improve the reliability of experience recollection. Therefore, to be included in the study, participants needed to be aged 60 years or older and have had at least 6 months of experience using a smart speaker in their homes. Community members were thus excluded if they did not meet the age requirement, or had less than 6 months of experience with a smart speaker.

During the recruitment process, the research team first contacted the community coordinator or the community head, who shared the information about this research with all the members of the community. The older adults who met the inclusion criteria and showed their interest in participating in the study were recruited. The snowball approach was then used to identify more older adults using smart speakers for more than 6 months. Those who met the inclusion criteria were recruited as participants. The participants were compensated with gift cards (valued at around US $9) for interview participation. Ultimately, 20 older adults were recruited (Table 1). The sample size is in line with the recommendations by Guest et al [30] and Hennink et al [31], who pointed out that empirical data reached saturation within 20 interviews. Other qualitative literature about older adults’ use of technology used a similar number of participants [24,32].

### Table 1. Demographics of participants.

<table>
<thead>
<tr>
<th>Label</th>
<th>Sex</th>
<th>Age (years)</th>
<th>Years of experience&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Technical proficiency&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Used smart speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>OA1&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Female</td>
<td>72</td>
<td>2</td>
<td>Intermediate</td>
<td>Brand A</td>
</tr>
<tr>
<td>OA2</td>
<td>Female</td>
<td>66</td>
<td>0.5</td>
<td>Intermediate</td>
<td>Brand A</td>
</tr>
<tr>
<td>OA3</td>
<td>Female</td>
<td>69</td>
<td>1</td>
<td>Novice</td>
<td>Brand A</td>
</tr>
<tr>
<td>OA4</td>
<td>Female</td>
<td>71</td>
<td>1</td>
<td>Novice</td>
<td>Brand A</td>
</tr>
<tr>
<td>OA5</td>
<td>Male</td>
<td>73</td>
<td>3</td>
<td>Novice</td>
<td>Brand B</td>
</tr>
<tr>
<td>OA6</td>
<td>Male</td>
<td>67</td>
<td>2.5</td>
<td>Novice</td>
<td>Brand B</td>
</tr>
<tr>
<td>OA7</td>
<td>Male</td>
<td>68</td>
<td>0.5</td>
<td>Intermediate</td>
<td>Brand A</td>
</tr>
<tr>
<td>OA8</td>
<td>Female</td>
<td>61</td>
<td>0.5</td>
<td>Novice</td>
<td>Brand B</td>
</tr>
<tr>
<td>OA9</td>
<td>Female</td>
<td>73</td>
<td>2</td>
<td>Intermediate</td>
<td>Brand A</td>
</tr>
<tr>
<td>OA10</td>
<td>Female</td>
<td>66</td>
<td>2</td>
<td>Novice</td>
<td>Brand A</td>
</tr>
<tr>
<td>OA11</td>
<td>Male</td>
<td>77</td>
<td>1.5</td>
<td>Intermediate</td>
<td>Brand B</td>
</tr>
<tr>
<td>OA12</td>
<td>Female</td>
<td>65</td>
<td>2</td>
<td>Novice</td>
<td>Brand B</td>
</tr>
<tr>
<td>OA13</td>
<td>Male</td>
<td>73</td>
<td>1.5</td>
<td>Intermediate</td>
<td>Brand B</td>
</tr>
<tr>
<td>OA14</td>
<td>Female</td>
<td>64</td>
<td>1</td>
<td>Novice</td>
<td>Brand A</td>
</tr>
<tr>
<td>OA15</td>
<td>Female</td>
<td>73</td>
<td>1.5</td>
<td>Novice</td>
<td>Brand A</td>
</tr>
<tr>
<td>OA16</td>
<td>Male</td>
<td>73</td>
<td>1</td>
<td>Intermediate</td>
<td>Brand B</td>
</tr>
<tr>
<td>OA17</td>
<td>Male</td>
<td>62</td>
<td>0.5</td>
<td>Intermediate</td>
<td>Brand B</td>
</tr>
<tr>
<td>OA18</td>
<td>Male</td>
<td>64</td>
<td>1</td>
<td>Novice</td>
<td>Brand B</td>
</tr>
<tr>
<td>OA19</td>
<td>Female</td>
<td>69</td>
<td>2</td>
<td>Novice</td>
<td>Brand A</td>
</tr>
<tr>
<td>OA20</td>
<td>Female</td>
<td>70</td>
<td>1.5</td>
<td>Intermediate</td>
<td>Brand B</td>
</tr>
</tbody>
</table>

<sup>a</sup>The actual time that the participant starts to use the smart speaker regularly (self-reported).

<sup>b</sup>Individuals’ ability to discern the usefulness of a smart speaker and their comfort in using it for various purposes (evaluated by the Artificial Intelligence Literacy Questionnaire proposed in [33], Multimedia Appendix 1).

<sup>c</sup>OA: older adult.

### Data Collection

A literature review of technology use among older adults informed the development of a flexible and semistructured interview guide. The interview guide was developed in consultation with a professor with expertise in qualitative data collection and analysis. One preliminary interview was conducted with the included community coordinator (aged 58 years) to ensure the fluency of formal interviews (eg, whether questions are easy to understand). As the goal of our study was to better understand older adults’ long-term use of smart speakers in home environments, the research questions were developed encompassing their practices, experiences, values, and expectations about using the smart speaker (Multimedia Appendix 2).

All interviews received written consent and were conducted in a private confidential environment such as the community activity room and private homes between November 2022 and September 2023. The participants were encouraged to give examples to contextualize their daily practices of using the smart speaker. The interviews ranged in length between 30 and 60 minutes. Participants who had limited knowledge or experience in using smart speakers had shorter interviews. All interviews
were audio-recorded. During the interview, we conducted regular summaries to ensure the validity of the collected data.

**Data Analysis**

All interviews were transcribed. The interview transcripts were reported in a way to ensure the anonymity of the participants. The collected data were interpreted by inductive content analysis [34], with a focus on how older adults use the smart speaker in their home environments, and why the speaker is used in the observed way. The transcripts were analyzed in NVivo (version 12; Lumivero) following the process proposed by Braun and Clarke [27]. After a data familiarization stage, 2 authors (FC and LS) separately extracted text segments related to the research goal and categorized them into codes. Applying an iterative process, the relationships between codes were analyzed and the subthemes were formed by grouping related segments together. The 2 authors compared the codes and subthemes and resolved any differences in coding through discussion. The assignment of text segments to the subcategories was repeatedly checked to see if they reflected the same meaning. Data saturation was determined when no new themes and relationships among the interview data were found [35]. As a result, a list of 4 themes and 12 subthemes, upon which the authors achieved an agreement, was generated [36].

**Results**

**Overview**

We identified 4 dimensions of the long-term use of smart speakers among older adults, including functional integration, spatial integration, cognitive integration, and semantic integration (Table 2).
Table 2. Identified themes, subthemes, and code examples related to the long-term use of smart speakers among older adults.

<table>
<thead>
<tr>
<th>Themes and subthemes</th>
<th>Code examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Functional integration</strong></td>
<td></td>
</tr>
</tbody>
</table>
| Entertainment | • Using it for listening to music  
• Chatting with the small thing for fun  
• Telling some stories |
| Information collection | • Hear daily news  
• Weather information  
• A useful tool for knowing what is going on |
| Medication reminders | • The speaker can send out medication reminders  
• When I see it, I know, “Oh I need to take pills” |
| Companionship | • Enjoy hearing the speaker’s answers  
• The speaker is accompanying me, he is a friend  
• Love sitting in the couch, with her beside me |
| Environment modification | • Just need to say “Turn off the lights”  
• Connect it with the switch  
• Connect to the camera to show who is knocking on the door |
| Emergency calls | • Call for help if falling on the ground |
| **Spatial integration** | |
| Smart speaker location | • Decide the device’s location according to daily habits  
• Put it beside the medicine box for task convenience |
| Appliance connection | • Locate the smart speaker in the bedroom for device connection  
• Put the smart speaker beside the door for better internet  
• Put it on the bedside table for easily hearing the doorbells |
| **Cognitive integration** | |
| Learning and problem-solving | • Memorize the required voice commands  
• Handle the emerging technical hurdles |
| Mental adaptation | • Overcome skepticism and building trust  
• Realize the benefits of smart speakers for personal convenience |
| **Semantic integration** | |
| Enjoyable experience | • Save energy in memorizing things  
• Assist older adults in performing multitask operations |
| Affective bonds | • Feel a shared and caring presence through the technology use  
• Experience tranquility while reading with the smart speaker |

**Functional Integration of Smart Speakers**

Functional integration relates to the ways in which older adults use smart speakers to meet various functional needs in their daily lives. This dimension focuses on the practical and functional aspects of smart speaker use and how they are integrated into daily practices. Our findings revealed the diversity of use genres that had been developed in older adults’ use of smart speakers in their living environments. Though most participants were using a smart speaker of the same brand, the ways of using the device were multiple, including entertainment, information collection, medication reminders, companionship, environment modification, as well as emergency calls.

The most commonly mentioned use genre was entertainment. Almost all participants said they use the smart speaker to listen to music.

...I often listen to songs [via the smart speaker]. Some songs could be used for our square dance. [Female, age 69 years]

In addition to listening to music, 1 participant mentioned that he once used the smart speaker to recognize the name of a song.

...I was recalling the name of a song. My granddaughter told me if I sang the song’s melody, the speaker could tell me the song’s name. [Male, age 73 years]
Similar to the example above, using the smart speaker for information collection was widely found in the interviews. Our participants used the smart speaker to gain information such as daily weather, local news, and safety tips related to the COVID-19 pandemic.

The smart speaker was also commonly used for medication reminders. The participants indicated that they had to always remember when to take medicines before using the smart speaker. With the help of their families or relatives, they set up the smart speaker to remind them to take medicines at a specific time.

...I have to take medicines every day... Now the speaker can remind me of taking medicines. [Female, age 66 years]

Besides, the participants emphasized that they benefited from the companionship of the device. The following example showed how older adults saw the smart speaker as a person who accompanied them, and enjoyed the conversation with the smart speaker.

...My smart speaker is a well-behaved little girl, just like my granddaughter. She wakes me up every morning. I say good morning to her sometimes. Sometimes I ask her some questions that I know the answers... I am just curious about how she would reply to me, and enjoy hearing her answers. [Male, age 73 years]

Another use genre was about environment modification. A few participants used the smart speaker to control other smart appliances in their living environments. For instance, 1 participant explained that the smart speaker helped overcome the inconvenience caused by switches that were in distant places.

The switch in my bedroom is beside the door... I had to get out of bed, switch off the light, and then get into the bed in the darkness... Now I just need to say “turn off the lights.” [Female, age 72 years]

Finally, 1 participant mentioned that she had not used the smart speaker for emergency calls, but she knew she could call her daughters through the smart speaker or a “white button” when necessary.

I asked the smart speaker to call my daughter, her phone rang. I clicked the white button, her phone rang...[a white button, about 5 cm in diameter, is shown in her hand]. [Female, age 61 years]

Spatial Integration of Smart Speakers

Spatial integration relates to the ways in which older adults use the physical space and placement of their smart speakers to shape and organize their daily routines. This dimension focuses on the physical aspects of smart speaker use. Our findings suggested that participants located their smart speakers in different living spaces, including the living room, the bedroom, the study, the bathroom, and the balcony. By spatially organizing the speaker, and connecting it with other smart appliances or daily tasks in these contexts, the participants built up the use contexts under which they are comfortable interacting with the smart speakers, and actively changed and reshaped their daily practices.

The following example showed how participants decided the locations of the smart speaker and formed new daily practices according to the purposes for using the smart speaker and the existing daily practices.

...I wanted to take the pills, I had to get water from the kitchen, so I put the speaker, together with the medicine box, in the corner of my kitchen now. [Female, age 73 years]

Some participants incorporated the smart speaker by connecting it with the existing appliances in their living environments. In doing so, their practices of completing specific tasks were simplified.

...The light in my bedroom can be connected and controlled by the smart speaker, I thus put it in my bedroom. Now I just need to say “turn off the lights” when I want to sleep. [Female, age 72 years]

Cognitive Integration of Smart Speakers

Cognitive integration relates to the way in which older adults adapt to and incorporate smart speakers into daily habits and routines from a cognitive perspective. Participants shared their experiences of familiarizing themselves with the smart speaker’s user interface and functionalities. The impact of successful problem-solving was evident in the participants’ narratives. Many acknowledged that navigating through and resolving issues not only enhanced their understanding of the device but also strengthened their overall intention to use the smart speaker in the long term. Overcoming learning or technical challenges contributed to a sense of accomplishment and increased confidence in using the smart speaker.

Almost all participants mentioned the challenges in learning voice commands and the overall functionality of the device, as well as resolving technical hurdles. They also mentioned the importance of adept problem-solving skills when facing these challenges. Some participants recounted instances of encountering glitches such as software malfunctions, intermittent connectivity issues, and occasional misinterpretations of voice commands by the smart speaker. One participant shared a specific experience, recalling:

There were times when it simply didn’t respond, leaving me puzzled. I had to delve into troubleshooting methods to identify the issue. [Female, age 64 years]

Another participant highlighted the initial difficulty in grasping the full range of functions, expressing:

Mastering the voice commands posed a bit of a challenge. Instead of saying “XX, I need...” you have to initiate with “Hi, XX...” took me some time to unravel the full range of functions it offered. [Male, age 62 years]

In response to these challenges, older adults demonstrated resilience and resourcefulness in addressing technical issues. Problem-solving strategies ranged from seeking help from...
family members or friends to experimenting with different emotional and cognitive aspects of smart speaker use. Some engaged in hands-on exploration, experimenting with different commands to enhance their functional and mental understanding. Additionally, participants discussed seeking guidance from families, internet-based resources, and user manuals to expedite the adaptation process. As 1 participant shared:

*I watched some tutorial videos online and read the manual, and asked my family. It helped me get a better grasp of what the smart speaker could do.* [Female, age 65 years]

As smart speakers became integrated into daily routines, participants described making gradual adjustments. The learning curve transformed into a journey of mental adaptation, with older adults incorporating the device into activities such as setting reminders, checking the weather, or even incorporating it into their medication routines. One participant shared the experience of overcoming skepticism and building trust with the smart speaker over time.

*I was skeptical at first...But as I gradually used it, I found myself relying on it more.* [Male, age 73 years]

The transformative power of smart speakers was evident in their ability to enhance social interactions during family gatherings. One user realized the potential of the device in this context, stating:

*I realized the potential of using the smart speaker during family gatherings. Now, we use it to play music, share interesting facts, and even settle debates.* [Male, age 64 years]

Similarly, another user’s experience highlighted the versatile nature of the smart speaker. Initially used for simple tasks like checking the weather and setting alarms, the device became a gateway to a multitude of possibilities.

*Initially, I used it mainly for weather updates and setting alarms. But then, I discovered I could ask it to read audiobooks or provide cooking tips. It’s like unlocking a treasure trove of possibilities.* [Female, age 70 years]

Semantic Integration of Smart Speakers

Semantic integration refers to the emotional and meaningful connections that older adults develop with their smart speakers over time. This dimension focuses on the emotional and cognitive aspects of smart speaker use. According to our data, although the smart speakers had occasional functional errors that annoyed the participants, all participants acknowledged, appreciated, and valued the benefits offered by using the smart speakers. The benefits included not only enjoyable experiences to complete daily tasks but also bonds with the device from an affective dimension. The enjoyable user experiences and the strong bonds with the smart speaker enhanced the participants’ positive attitudes toward their device, making them feel that occasional functional errors were acceptable.

The participants claimed that the smart speaker enabled them enjoyable experiences such as helping them access more information and enjoy more forms of entertainment. Besides, some said the speaker empowered them to multitask and saved time and energy. Yet, looking solely at the daily tasks completed by smart speakers, participants thought the smart speakers were not unreplaceable. As 1 participant indicated:

*Apparently the smart speaker makes the completion of some daily tasks easier...but if only looking at the tasks it supports, I feel it doesn’t have that much uniqueness.* [Female, age 69 years]

The same person, however, emphasized that her affective bonds with the smart speaker meant a lot to her.

*...what I feel uniqueness is the companionship offered by the smart speaker...by using it, I feel as if someone is experiencing every day together with me, and taking care of me.* [Female, age 69 years]

Other participants also emphasized the importance of companionship offered by the smart speaker. Some stated that they started to develop bonds with the smart speaker because of its capability for task completion, but the bonds were strengthened and maintained through the companionship over time, driving them to maintain their use behaviors.

*...My son bought the smart speaker for me, because my friend in the community said it was convenient to listen to music.... Gradually [the speaker] becomes a "person" accompanying me.... I feel peace of mind when I am reading a book while he is just beside me, in the room together with me.* [Male, age 67 years]

The functional errors that emerged during the technology use were spontaneously mentioned by our participants. However, people looked at the errors with charity. Some even treated these errors as fun. For example, 1 participant mentioned that sometimes the smart speaker may be activated by the television sound.

*Once, the leading actor said “I don’t love you,” the smart speaker replied “That is heartbreaking” [laughing]... I like it, I think these so-called “functional errors” are acceptable, but sometimes are also fun.* [Male, age 67 years]

In this case, the smart speaker was not viewed as a pure device whose fluency in function operations was prioritized. Instead, the participant treated the smart speaker as a social agent as its inappropriate responses activated by television was unexpected social actions (ie, conversations between different devices).

**Discussion**

**Principal Findings**

The study aimed to investigate how smart speakers are integrated into the daily practices of older adults in the long term. The findings of our study illuminate a nuanced understanding of how older adults seamlessly integrate smart speakers into their lives, encompassing functional, spatial, cognitive, and semantic dimensions. This integration aligns with existing literature, shedding light on the multifaceted benefits and challenges associated with the adoption of smart speaker technology among older populations.
Integrate Multiple: Functional, Spatial, Cognitive, and Semantic Technology Integration

The functional integration of smart speakers among older adults is characterized by a rich diversity of use genres, reflecting the adaptability of these devices to meet various needs. Entertainment emerged as a dominant use genre, with participants expressing a universal affinity for using smart speakers to listen to music, transforming the device into a musical companion for activities like square dancing. This aligns with previous studies emphasizing the role of smart speakers in enhancing leisure activities for older adults [37]. Beyond entertainment, the device’s pivotal role in information collection, medication reminders, and environment modification resonates with literature, highlighting smart speakers as valuable tools for health management and home automation among older adults [38]. The potential for emergency calls, while not frequently used, aligns with the findings of studies emphasizing the importance of safety features for older adults [39]. The acceptance of smart speakers as reliable companions for various tasks aligns with the idea that these devices can address specific needs and preferences, contributing to the overall well-being of older users [40].

Spatial integration revealed how older adults strategically positioned smart speakers to shape and organize their daily routines. The participants allocated their speakers across diverse living spaces, from the living room to the bedroom, study, bathroom, and even the balcony. This spatial organization facilitated the creation of specific use contexts, where the smart speaker seamlessly blended into existing daily practices. This personalized approach aligns with literature emphasizing the importance of tailoring technology to older adults’ physical environments [41]. The integration of smart speakers with daily routines, such as pill-taking in the kitchen or controlling bedroom lighting, showcases the adaptability of older adults in incorporating technology seamlessly into their daily lives, emphasizing the active role older adults play in shaping their living environments to accommodate and optimize the technology functionality [42,43].

Cognitive integration illuminated the adaptive processes through which older adults familiarized themselves with the smart speaker’s functionalities and overcame learning challenges. The initial hurdles in mastering voice commands and comprehending the device’s full range of functions were common experiences. Yet, the participants’ narratives highlight their resilience and problem-solving acumen, aligning with studies that stress the importance of user support and educational resources in facilitating older adults’ technology adoption [44]. The gradual adjustments and incorporation of the smart speaker into daily habits reinforce the idea that cognitive integration is an ongoing process, transforming initial challenges into a journey of mental adaptation [45].

Semantic integration delved into the emotional and meaningful connections older adults forged with their smart speakers. Despite occasional functional errors, participants universally acknowledged and valued the benefits offered by these devices. The smart speaker’s role in facilitating enjoyable experiences, multitasking, and saving time and energy contributed to positive attitudes among older adults. The findings speak against the common discourse that technical issues and occasional functional errors could significantly diminish the overall positive impact and perceived value of smart speakers among older adults [46]. Instead, participants not only accepted these imperfections with understanding but also found moments of humor and enjoyment in the device’s occasional quirks. This resilience toward technical glitches underscores the robust emotional and meaningful connections formed between older adults and their smart speakers [24,47], highlighting that the perceived benefits and companionship offered by the technology far outweigh occasional operational hiccups.

To sum up, the integration of smart speakers among older adults is a holistic process that involves functional use, spatial organization, cognitive adaptation, and semantic connection with the device. Understanding the interplay of these dimensions provides a comprehensive insight into how smart speakers become integral components of older adults’ daily lives, offering not only practical functionalities but also emotional fulfillment and companionship.

Design Implications

Being Aware of the Heterogeneity in Older Adults’ Technology Use

In our study, older adults used the smart speaker in different physical spaces, through different actions, and for different purposes. The underlying reasons could be the diversity of personal spheres [33], the differences in sociocultural and socioeconomic situations [48], and the disparities in individual digital literacy [49]. The heterogeneity in older adults’ technology use may pose challenges for designers and developers to transform ideas and insights into concrete designs. Merely focusing on a specific issue met by older adults would limit the benefits of smart speakers to older adults. The complex entanglement of space, practices, and user needs that may affect technology use at different sites should be carefully considered. We thus suggest designers and developers to zoom out and get the big picture of older adults’ daily practices before zooming in to define the design problem to solve.

Serving for the Existing Activities in Older Adults’ Home Environments

Despite the discourse that smart speakers can support older adults to explore unexperienced activities, in our study, participants used the device to enhance the quality of their existing activities. For instance, the smart speaker supports medication reminding, turning on or off the lights, and music listening. Consequently, we suggest designers and developers to pay more attention to the issues or opportunities related to older adults’ existing activities at home, rather than new activities that are not commonly experienced among this group of population.

Attaching Importance to Additional Devices Centered on the Smart Speaker

Our findings show that some additional devices connecting to the smart speakers were used by older adults. For instance, a “white button,” designed for emergency calls was carried by a
participant all the time. The solution of connecting smart speakers to light switches in case of harm in darkness was also acknowledged. The 2 cases suggest that additional devices, especially the ones ensuring personal safety, have a market in the aging groups. Because many older adults have difficulties using the apps on mobile phones [3], we argue more additional devices centered on the smart speaker are needed.

**Leaving Space for the Agency of Older Adults**

Older adults are usually treated as inexpert technology users, and judged to be risky to do technology appropriation [50]. Yet, our findings show how participants spatially organized the smart speaker, and actively defined, changed, and reshaped their practices after using the device. In our cases, older adults created solutions that suit their specific needs [51]. These solutions were not predefined or preformatted by someone else, but specifically for older adults themselves [52]. We believe this may contribute to developing strong affective bonds between older adults and the device, driving their maintainable use behaviors, and promoting their resilient attitudes toward functional errors. Hence, we call for more studies about the association between older adults’ agency in technology use and their affective bonds with technology. We also encourage designers and developers to carefully consider the space for the agency of older adults in terms of technology appropriation.

**Getting to Know the Preferences of Communities**

The most commonly mentioned genre of using the smart speaker was listening to music in our study. This is understandable as many participants are from a square dancing community. It shows that the community where older adults are involved can implicitly manifest their preferences toward the purposes of technology use. Therefore, we encourage designers and developers to shift partial attention from individuals to communities when investigating the preferences of older adults.

As smart speakers swiftly reach the mainstream, understanding the detailed, nuanced experiences of older adults using these devices becomes paramount. Our research delves into the intricacies of daily use. By unraveling the complexities and diverse experiences during older adults’ long-term engagement with smart speaker technology, we aim to bridge the gap between technology and aging, offering insights that go beyond design considerations to impact policies and interventions. Moreover, our study seeks to contextualize the significance of understanding older adults’ experiences within the broader sociotechnological landscape. The potential benefits of these devices in health care, chronic disease management, and emergency identification align with broader societal challenges in health care accessibility and aging populations. In an age where digital advancements redefine societal norms, the integration of smart speakers into the lives of older adults is not just a technological trend; it is a pivotal factor in ensuring their inclusion in the digital revolution. Consequently, this research does not merely contribute to the attractiveness of smart speakers for older users but serves as a critical voice in the ongoing dialogue about healthy aging in the digital era.

**Conclusions**

This paper focuses on the integration of smart speakers into older adults’ daily practices. We aim to understand how older adults routinely use smart speakers and integrate them into their daily lives, which in turn enables us to propose suggestions for designers and developers. A total of 20 older adults who have used a smart speaker for over 6 months were interviewed. The findings demonstrate that older adults integrate smart speakers into daily use through functional integration, spatial integration, cognitive integration, and semantic integration. Based on the findings, we proposed several suggestions for designers and developers to better design smart speakers that promote maintainable use behaviors by older adults.

**Limitations**

Our findings must be evaluated within the context of several limitations. First, participants were recruited from 2 communities, potentially impacting the generalizability of the results. The risk of selection bias or unaccounted factors, such as the homogeneity of participant characteristics within the same facility, may have influenced interview responses. Yet, the study indeed highlighted the intricate nature of long-term technology use among older adults in these specific communities, offering valuable insights for designers and developers in the context of smart speaker design. Second, our study only investigated the long-term use of smart speakers, without paying much attention to the discontinuation of using smart speakers in the long term. The investigation of the discontinued use of smart speakers may reveal more insights into older adults’ technology use. However, we believe that it is important to examine the integration of smart speakers in older adults’ daily practices, as the findings may contribute to the maintainability of older adults’ technology use.

**Acknowledgments**

The authors would like to thank the participants in the study for sharing their experiences and perceptions. We also thank the Ageing Dancing community and Bright community for making the study possible. This research was funded by a Chinese postdoctoral grant (2023M732231) and a Design School Innovation grant (WH102243005/201/004).

**Conflicts of Interest**

None declared.

Multimedia Appendix 1

Artificial Intelligence Literacy Questionnaire (adapted).
References


7. Sunshine J. Smart speakers: the next frontier in mHealth. JMIR Mhealth Uhealth 2022;10(2):e28686 [FREE Full text] [doi: 10.2196/28686] [Medline: 35188467]


Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research

©Fangyuan Chang, Lin Sheng, Zhenyu Gu. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 12.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Hearing Rehabilitation With a Chat-Based Mobile Auditory Training Program in Experienced Hearing Aid Users: Prospective Randomized Controlled Study

Jae Sang Han1, MD, PhD; Ji Hyung Lim1, MD; Yeonji Kim1, MD; Aynur Aliyeva1,2, MD; Jae-Hyun Seo1, MD, PhD; Jaehyuk Lee3, BSc; Shi Nae Park1, MD, PhD

Corresponding Author:
Shi Nae Park, MD, PhD

Abstract

Background: Hearing rehabilitation with auditory training (AT) is necessary to improve speech perception ability in patients with hearing loss. However, face-to-face AT has not been widely implemented due to its high cost and personnel requirements. Therefore, there is a need for the development of a patient-friendly, mobile-based AT program.

Objective: In this study, we evaluated the effectiveness of hearing rehabilitation with our chat-based mobile AT (CMAT) program for speech perception performance among experienced hearing aid (HA) users.

Methods: A total of 42 adult patients with hearing loss who had worn bilateral HAs for more than 3 months were enrolled and randomly allocated to the AT or control group. In the AT group, CMAT was performed for 30 minutes a day for 2 months, while no intervention was provided in the control group. During the study, 2 patients from the AT group and 1 patient from the control group dropped out. At 0-, 1- and 2-month visits, results of hearing tests and speech perception tests, compliance, and questionnaires were prospectively collected and compared in the 2 groups.

Results: The AT group (n=19) showed better improvement in word and sentence perception tests compared to the control group (n=20; P=.04 and P=.03, respectively), while no significant difference was observed in phoneme and consonant perception tests (both P>.05). All participants were able to use CMAT without any difficulties, and 85% (17/20) of the AT group completed required training sessions. There were no changes in time or completion rate between the first and the second month of AT. No significant difference was observed between the 2 groups in questionnaire surveys.

Conclusions: After using the CMAT program, word and sentence perception performance was significantly improved in experienced HA users. In addition, CMAT showed high compliance and adherence over the 2-month study period. Further investigations are needed to validate long-term efficacy in a larger population.

Trial Registration: Clinical Research Information Service (CRiS) KCT0006509; https://cris.nih.go.kr/cris/search/detailSearch.do?seq=22110&search_page=L

(JMIR Mhealth Uhealth 2024;12:e50292) doi:10.2196/50292

KEYWORDS
hearing loss; hearing aids; hearing rehabilitation; auditory training; mobile program

Introduction

According to a World Health Organization report in 2021, it is predicted that a quarter of people will have some degree of hearing loss by 2050 [1]. Hearing loss is known to be related to the risk of vocational problems, depressed mood, and even cognitive impairment in adults [2-4]. Adult aural or hearing rehabilitation, a multifactorial strategy for patients with acquired hearing impairment, includes 4 main components: sensory management, instruction, perceptual training, and counseling [5]. Sensory management involves enhancing auditory function through the use of hearing aids (HAs), cochlear implants, and alternative devices. Additional perceptual training, including auditory training (AT), seems to be required because sensory management alone has limitations in improving speech perceptual ability [5,6].

Although AT is known to improve hearing performance and lower the HA return rate, it has been not sufficiently provided to patients with hearing loss due to its poor time- and cost-effectiveness [7-9]. In our previous randomized controlled trial, we reported significant improvements in consonant perception and subjective satisfaction after 8 weeks of AT.
consisting of in-hospital face-to-face education and at-home self-training among HA users [10]. However, we experienced the limitations of in-hospital training programs during the COVID-19 pandemic, and this led us to consider a mobile-based AT program.

Smartphone penetration is rapidly increasing worldwide, and South Korea is a leading county for smartphone use, with a penetration rate of around 93% in 2021 [11,12]. Among the older population, around 94% of those older than 60 years and 60% of those older than 70 years own smartphones, and most of them (80%) use the internet with a mobile device [12,13]. The increasing smartphone penetration rate is a global trend, and mobile-based health care programs are expected to be used effectively even in the older population [14]. We developed a new mobile AT program with a chat-based interface to increase compliance with hearing rehabilitation, with attention to the fact that most of the adult population is familiar with message apps on smartphones [15].

Several mobile and web-based AT programs have been introduced in the last decade. However, many mobile-based AT applications are simple gamified training programs for pediatric patients, and most of them have no clinical validation to support their efficacy [16]. Even though a state-of-the-art review showed AT to be effective for improving auditory perceptions [17], it is unclear whether mobile-based training will still be effective. Digital therapeutics are expected to change the future of the health care system; however, clinical validation should first be performed to prove their effectiveness [18]. In addition, the effect of AT on experienced HA users has been inconsistently reported, while the effect of AT on novel HA users is relatively clear. Of the 16 studies included in a state-of-the-art review on AT, only 3 investigated experienced HA users, and only 1 of these studies showed improvements in speech perception [19-21].

Therefore, this prospective randomized controlled study was conducted to investigate the efficacy of our novel chat-based mobile AT (CMAT) program for 3 aspects of hearing rehabilitation in experienced HA users: speech perception ability, subjective satisfaction, and training compliance.

Methods

Ethical Considerations

This prospective study was approved by the ethical committee of Seoul St. Mary’s Hospital (KC21EISI0525) and followed the tenets of the Declaration of Helsinki. The patient records and information were anonymized and deidentified before analysis. All participants provided written informed consent prior to commencement of the study and voluntarily participated in this clinical trial.

Participants

A prospective randomized controlled trial with an unblinded study setting was carried out. Bilateral HA users were recruited from the department of otorhinolaryngology–head and neck surgery at a tertiary referral center between September and December 2021. Study eligibility criteria were as follows: (1) age ≥20 years; (2) bilateral moderate to severe sensorineural hearing loss (mean threshold of pure tone audiometry measured at 500, 1000, 2000, and 4000 Hz was 41 to 80 dB hearing level, and the air-bone gap was less than 15 dB hearing loss); and (3) the patient had been using bilateral HAs for more than 3 months, demonstrated sufficient functional gain, and consistently used the HAs for more than 8 hours per day. Patients were excluded if they had fluctuation in hearing loss, brain tumor, or difficulty using the program or coordinating hearing tests.

The randomization was performed using premade random cards with a 1:1 allocation ratio prepared by a contract research organization (Medical Excellence). The cards were then opened by the health care provider after obtaining informed consent. Participants allocated to the AT group (ATG) were provided instructions about the CMAT program by a coauthor (JL) and were encouraged to use it daily for 8 weeks. One CMAT session consisted of 20 questions, took about 5 to 10 minutes, and was repeated 3 times a day. Those assigned to the control group (CG) did not receive the CMAT. All participants were asked to complete audiologic tests, speech perception tests, and questionnaires before enrollment and at 1 and 2 months after the start of the study, respectively.

Assuming that the effect of offline AT and CMAT would be similar, we determined the number of target participants in the same way as in our previous study [10]. For statistical significance at a .05 confidence level with 80% power, the sample size required for the 2 groups was estimated as 18 patients per group. Allowing for a 10% dropout rate, 42 patients were estimated to be required in total.

Chat-Based AT Program

Our CMAT program was developed in the form of a web-based program that can be accessed from computers, tablets, and smartphones. However, all participants were requested to connect with their smartphones to avoid device bias in this clinical trial. When a participant accessed a provided link, the ID and password given during participant registration were required to be entered, and a page was provided to additionally confirm the consent form for clinical trials and the collection of personal information. Afterward, a problem was presented in a chat-based interface. In the chat-based interface, we designed a conversational partner as a character representing a medical professional, intending to give users the feeling of exchanging chat bubbles with a health care provider. Participants could choose 1 of 4 answers, and a message was shown giving feedback after a problem was answered. All participants were instructed to train in a quiet place while wearing their HAs.

The rehabilitation program consisted of 2 parts: word and sentence training. For word training, a total of 1540 two-syllable Korean words that are frequently used in daily life were extracted based on data provided by the National Institute of the Korean Language. Four words with the same or similar vowels but different consonants were provided, and then one of these words was provided twice. Then, the participant was asked to choose the word that he or she heard. For sentence training, daily phrases consisting of 2- to 12-word phrases were provided, and the frequency of the words used was checked to correct or delete infrequently used words (eg, “We decided to meet at Seoul Station”). A total of 1225 sentences were
composed, and a question was made for each sentence. The sentences and related questions were provided once, and 3 words similar to the correct answer were shown with it (eg, “Question: Where would you like to meet?”; “Answers: (1) Sadang Station, (2) Sung-soo Station, (3) Sung-book Station, (4) Seoul Station”).

This CMAT program was presented as if chatting with a virtual character. If a participant answered the question correctly, the next question among the word and sentence training question pools was shown randomly. If a participant chose a wrong answer, the question was repeated up to 3 times. Ten-word and 10-sentence questions were randomly assigned to participants for solving within a single session. AT dosage was planned to be around 30 minutes of training per day, similar to our previous study, as solving 20 questions was expected to take approximately 10 minutes. However, if a participant wanted to solve more questions, they could engage in an additional session. Participants included in the ATG were instructed to repeat this session 3 times a day, spaced out, following a similar approach to our previous study [10]. A schema of the CMAT is shown in Figure 1.

Figure 1. Schematic diagram of chatbot-delivered auditory training program.

The Korean language was used in this program, and words and sentences were synthesized using Google Cloud text-to-speech software. To facilitate scalability for increases in the number of sentences or upgrades in the future, we determined that software-generated voices would be a reasonable choice. The speech was randomly generated using male or female voices.
provided by the software, and the default settings of the program were used for other suprasegmental characteristics of sentences, including speed, pitch, and volume. The log-in time and the number of questions solved were automatically collected by the program; however, prompts to use the CMAT were not provided.

**Outcome Measures**

**Study Design**

All participants completed audiologic tests, speech perception tests, and questionnaires at the 0-, 4-, and 8-week visits, and the results were collected using Excel (Microsoft). Among these measured outcomes, speech perception test results were used as the primary end point, and questionnaire results were used as the secondary end point.

**Audiologic Tests**

Pure tone audiograms (PTAs) without HAs and sound field threshold audiometry tests with HAs were conducted to exclude changes in hearing level and inappropriate function of HAs. Speech audiometry tests, including the speech reception threshold test and speech discrimination scores, were evaluated both with and without the participants wearing HAs.

**Speech Perception Tests**

A total of 5 speech perception tests were conducted by a female audiologist in quiet conditions while the patient wore an HA. The Ling Six Sound test was conducted for phoneme detection [22], and the Vowel and Consonant Imitation Test (VCIT) was used to measure phoneme perception [23]. To evaluate word perception ability, monosyllable and bisyllable open-set tests were used [24]. Sentence perception ability was evaluated using the Korean version of Central Institute for the Deaf (K-CID) test [25].

**Questionnaires**

Subjective benefits were measured by 3 validated questionnaires. The Korean version of the Hearing Handicap Inventory for the Elderly (K-HHIE) was used to measure situational and emotional handicaps in everyday life due to hearing problems [26]. In addition, the Korean versions of the International Outcome Inventory for Hearing Aids (K-IOI-HA) and Abbreviated Profile of Hearing Aid Benefit (APHAB) tests were conducted to evaluate the subjective satisfaction with HA use [27,28].

**Statistical Analysis**

Statistical analysis was conducted using SAS (version 9.4; SAS Institute). The Shapiro-Wilk test was used to examine the normality of the measured variables. Data were expressed as mean, SD, and percentage. P values were calculated using the chi-square test or Fisher exact test for categorical variables, the Mann-Whitney test or independent 2-tailed t test for pairs of independent variables, and the Wilcoxon rank sum test or paired t test for continuous variables. All speech perception tests were collected as percentage data, and as the data did not meet the normality assumption based on the Shapiro-Wilk test, we conducted additional analyses using arcsine transformation. Speech perception tests and questionnaire surveys were analyzed with a linear mixed-effect model and post hoc test. Random effects included the intercept, while fixed effects included group (ATG and CG), time (initial, 1 month, and 2 month), group x time interaction, and the covariate “aided 4 kHz threshold result of the right ear.” The inclusion of the aided 4 kHz threshold result of the right ear as a covariate was due to the significant differences between the 2 groups in the initial hearing test despite randomization. Further details regarding this will be provided in the Results section. Two covariance structures in the linear mixed model, variance components and compound symmetry, were compared. There were no substantial differences in the results of the fixed effect type 3 test, so we described the variance components estimation. The Akaike information criterion values representing model fit varied depending on the specified model, from 280 to 680. Bonferroni correction was used for multiple comparisons and adjusted P values are presented. Correlations between pairs of variables were analyzed with the Pearson correlation test. Differences were considered significant when the P value was less than .05.

**Results**

**Clinical Characteristics**

A total of 42 participants were initially enrolled after informed consent was obtained. The ATG and CG were each randomly allocated 21 patients, with 1 participant subsequently withdrawn from the ATG due to withdrawal of consent. During the 2-month study follow-up period, 1 participant in the ATG was excluded from the analysis due to low compliance (CMAT completion rate <50%), and 1 participant in the CG was excluded for refusing to perform speech perception tests. Therefore, 19 participants in the ATG and 20 participants in the CG were analyzed. The CONSORT (Consolidated Standards of Reporting Trials) flow diagram is shown in Figure 2. None of the participants in the ATG had any problems using CMAT during the study.
There were no significant differences between the 2 groups in age, sex, duration of hearing loss, type of HA being used, or duration of HA use. The detailed patient demographics are shown in Table 1.
**Table**. Clinical characteristics of the participants.

<table>
<thead>
<tr>
<th></th>
<th>Total (n=41)</th>
<th>Auditory training group (n=20)</th>
<th>Control group (n=21)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>70.0 (10.9)</td>
<td>70.5 (6.9)</td>
<td>69.5 (13.8)</td>
<td>.41a</td>
</tr>
<tr>
<td><strong>Sex, n</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>16</td>
<td>8</td>
<td>8</td>
<td>.90b</td>
</tr>
<tr>
<td>Female</td>
<td>25</td>
<td>12</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td><strong>Hearing aid type, n</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete in the canal</td>
<td>67</td>
<td>33</td>
<td>34</td>
<td>.75c</td>
</tr>
<tr>
<td>Invisible hearing aid</td>
<td>10</td>
<td>4</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Receiver in the canal</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Hearing loss duration (months), mean (SD)</td>
<td>103.0 (86.5)</td>
<td>107.9 (83.3)</td>
<td>98.3 (91.3)</td>
<td>.48a</td>
</tr>
<tr>
<td>Hearing aid use duration (months), mean (SD)</td>
<td>54.5 (62.1)</td>
<td>56.1 (72.2)</td>
<td>53.0 (50.7)</td>
<td>.80a</td>
</tr>
</tbody>
</table>

aMann-Whitney test.
bFisher exact test.
cChi-square test.

**Audiologic Evaluations**

Although we randomly allocated participants to the 2 groups, there were unintended differences in the baseline hearing tests. In the ATG, the high-frequency thresholds (3000, 4000, and 6000 Hz) of the right ear were significantly higher than in the CG in both unaided and aided conditions (independent t test or Mann-Whitney test, all P<.05). However, there was no significant difference in the left ear. In speech audiometry, mean values for speech discrimination score were higher in the CG without a statistically significant difference (independent t test or Mann-Whitney test, all P>.05). No significant changes over time were observed in the repeated audiologic tests under either unaided or aided conditions (linear mixed model, all P>.05). These audiologic test results are presented in Figure 3.
Speech Perception Tests

With the exception of the vowel imitation test, the initial mean scores for the remaining 5 tests were better in the CG. Although the Ling Six Sound test, monosyllable and bisyllable open-set tests, and the K-CID test showed no significant differences (adjusted $P > .05$ for all tests; Mann-Whitney test with Bonferroni correction), the initial consonant imitation test exhibited a statistically significant difference (adjusted $P = .003$; Mann-Whitney test with Bonferroni correction). Therefore, the aided 4 kHz threshold result of the right ear, which exhibited a significant difference in the initial PTA result ($P = .002$; Mann-Whitney test), was included as a variable in the analysis of the linear mixed model, considering that high-frequency hearing loss is a well-known factor affecting hearing perception [29].

When conducting a linear mixed model that includes the 4 kHz threshold of the right ear as a variable, we observed no significant difference in the initial scores of any speech perception tests (all $P > .05$).

In the Ling Six Sound test, the VCIT (vowel and consonant) and monosyllable test results showed no significant group differences (all $P > .05$, linear mixed model). However, in the
bisyllable condition, the ATG demonstrated significantly better results compared to the CG ($P=.04$, linear mixed model). Similarly, in the K-CID test, the ATG showed significantly greater improvement than the CG ($P=.03$, linear mixed model). These results are visualized in Figure 4, and detailed linear mixed model analyses, including post-hoc $P$ values, are presented in Table 2. Additionally, an arcsine transformation was applied to the percentage values, yielding almost the same results. Detailed values and statistical analyses of both the percentage and arcsine-transformed data are provided in Multimedia Appendix 1.

**Figure 4.** Results of speech perception tests. The auditory training group (ATG) showed significant improvement in speech perception compared to the control group (CG) in the intergroup analysis of the 2-syllable word recognition test and the K-CID (Korean version of Central Institute for the Deaf) test. ns: not significant; VCIT: Vowel and Consonant Imitation Test. *$P<.05$; linear mixed model.
<table>
<thead>
<tr>
<th></th>
<th>Group, post hoc P value</th>
<th>Time, post hoc P value</th>
<th>Auditory training group</th>
<th>Control group</th>
<th>Group × time, post hoc P value</th>
<th>F (df)</th>
<th>P value&lt;sup&gt;ab&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial</td>
<td>1 mo</td>
<td>2 mo</td>
<td>Initial minus 1 mo</td>
<td>1 mo minus 2 mo</td>
<td>Initial minus 1 mo</td>
<td>1 mo minus 2 mo</td>
</tr>
<tr>
<td>Ling Six Sound test</td>
<td>&gt;.99</td>
<td>.79</td>
<td>&gt;.99</td>
<td>&lt;.001</td>
<td>.004</td>
<td>&gt;.99</td>
<td>.11</td>
</tr>
</tbody>
</table>

<sup>a</sup>Group × time interaction effects compared the initial minus 1 month or initial minus 2 months.
<sup>b</sup>VCIT: Vowel and Consonant Imitation Test.
<sup>c</sup>K-CID: Korean version of Central Institute for the Deaf test (linear mixed model).

**Subjective Satisfaction**

There was no significant difference between the 2 groups in all 3 types of questionnaire surveys (all P>.05; linear mixed model). Detailed results are described in Multimedia Appendix 2.

**Compliance With CMAT Program**

Except for one participant in the ATG who withdrew consent, all participants (n=20) in the intervention group accessed the CMAT program in a mobile environment. One of 20 participants (5%) was excluded from the analysis in this study because the duration of CMAT use was less than 50% of the required AT. The remaining 19 participants (95%) completed at least 50% of the training sessions, and 17 (85%) completed more than the required AT.

Log-in times, completion rates, and scores of all participants were collected by the CMAT program and analyzed. Mean daily application log-in time was 20.3 (SD 11.4) minutes in the first month and 20.4 (SD 10.7) minutes in the second. The completion rate was higher than our recommendation; 123.7% (SD 71.4%) in the first month and 128.5% (SD 61.3%) in the second month. There were no statistically significant differences among these variables (P=.80 for completion rate and P=.50 for log-in time; Wilcoxon signed rank test) (Figure 5A). The analysis of compliance and speech perception tests revealed no significant correlation (Figure 5B).
Discussion

Principal Results

In this study, we developed a new CMAT program in the Korean language, and we confirmed that it was effective in improving speech perception abilities and had high compliance among experienced bilateral HA users.

Our study evaluated speech perception performance at the levels of phoneme, word, and sentence. When the tests were repeated, they all showed improvements in both the ATG and CG. This is a similar finding as our previous report, and it might have been caused by the participants adapting to the test materials [10,30]. Interaction effects implied that compared to the initial time point, at 1 month and 2 months there was a greater performance improvement in the ATG in the bisyllable test and K-CID test. However, no significant findings were observed for the Ling Six Sound test, VCIT, or monosyllable test. The prominent improvement in bisyllable word and sentence perception tests could be explained by the effect of AT having been achieved only to a limited degree for the trained task. The CMAT program used bisyllable words and sentences, which resulted in marked improvement on the tests assessing the same ability. Several previous studies suggested that the effect of AT
appeared only in trained tasks and did not lead to improvements in untrained tasks [21,31,32].

All 3 questionnaires measuring subjective hearing disability and HA satisfaction showed no improvement after 2 months of CMAT, a result that is contrary to our previous study. Our previous study investigated the effectiveness of hospital-based AT and showed significant improvements on the K-IOI-HA and K-HHIE questionnaires in the intervention group [10]. A systematic review suggested that telemedicine has similar feasibility as conventional hearing rehabilitation; however, this review paper was not focused on AT; it focused more on HA fitting, device testing, and counseling [33]. To the best of our knowledge, only 3 comparative studies have been conducted of telemedicine and conventional AT [34-36]. A pilot study comparing the feasibility of tablet-based AT and conventional face-to-face AT in adult cochlear implant recipients reported similar results to our study. Even though tablet-based AT showed better speech improvements in tablet-based training, subjective auditory ability measured by the Oldenburger Inventory-R questionnaire showed partial improvement only in the conventional AT group and not in the teletherapeutic group [36]. The reason why we could not find subjective speech perception improvement in the questionnaire seems to be the lack of interaction between the participants and health care providers, who encourage and reassure patients during each session of conventional AT programs. A relatively short study period might be another reason. A long-term follow-up study or a study with participants who are more exposed to health care providers during CMAT might be needed to support our suggestions in the future.

Of the 20 participants who agreed to CMAT, 17 (85%) completed all or more requested training times, and there was no difference in compliance during the 2-month study. Our completion rate is similar to previous studies [10,37], and it seems that mobile-based AT is easily accessible and that the chat-based interface gives users a more interactive impression [8].

In our study, there was no statistically significant correlation between CMAT compliance and listening performance, which might be due to the high and homogeneous completion rate of an average of 120% of the program in the ATG. However, previous studies suggested that training duration is related to the improvement in listening abilities. A study using the Listening and Communication Enhancement training program in veterans using HAs reported that 84% (42/50) of participants completed the required sessions, and improvement in off-task performance was significantly better in the group of participants who completed all the sessions [38]. Another study examined the use of ReadmyQuips among new HA users and reported a correlation between the words-in-noise test and training time but not between the hearing-in-noise test and training time [39].

Strengths and Limitations

The mobile-based AT program we introduced has several advantages that set it apart from other internet-based AT programs. The first advantage lies in the interface, which features chatting, thus catering to users familiar with mobile messengers and providing a more convenient experience. Simplistic problem-solving might bore users easily, and gamification could come across as juvenile for adult users. Additionally, our interface uses a format in which users chat with a character resembling a medical professional, aiming to give users the impression of receiving remote treatment from experts. The second advantage pertains to the quality and quantity of the provided problems. Other AT programs sometimes present meaningless word lists or scenarios where the sound completely differs from the presented words. However, our CMAT program was designed with the intention of using 1540 frequently used words provided by the National Institute of the Korean Language to offer assistance in users’ everyday lives. We aimed to enhance consonant discrimination abilities that are challenging for individuals with hearing impairments by creating problems that categorize cases where the vowels are similar but the consonants differ. Additionally, due to the vast number of questions, each exceeding 1000, participants might not encounter the same question more than once during the 2-month test period.

Nevertheless, CMAT also has its limitations. Recent artificial intelligence (AI)–powered language models like ChatGPT, which enable real-time interactions, were not integrated. The program did not incorporate various user-responsive technologies such as those found in AI-driven tools. Furthermore, additional settings such as background noise or speech speed adjustments were not possible in this version of CMAT. We refrained from including these adjustable settings to maintain the consistency of the training program and avoid introducing bias in the study. However, in the next version, we plan to incorporate features that allow users to adjust the program to their skill level.

From a clinical study perspective, there are several other advantages. Foremost among them is the fact that this is the first well-designed study conducted using a mobile-based AT program using the Korean language. This study also suggests the possibility that AT could be widely applied to the hearing rehabilitation process by showing the effect of AT on users with HA experience, which has been controversial in previous reports [19-21].

However, several limitations that emerged during the course of this study should also be taken into consideration. First, despite the randomization, the high-frequency hearing threshold in the right ear was significantly higher in the ATG than the CG, indicating that the CG score was better in most initial speech perception tests. We considered a subgroup analysis such as participant matching to avoid differences in initial hearing threshold, but we could not conduct such an analysis due to the limitations of a randomized prospective study with a relatively small number of participants. Instead, we tried to minimize unintentional bias by adjusting the 4 kHz threshold of the PTA when analyzing with the linear mixed model; however, there is a possibility that the effect of CMAT was underestimated due to differences in initial hearing threshold. Further research using a larger number of participants is needed in the future. Second, a 2-month training time and follow-up duration might not be long enough to show the effect of CMAT on all types of hearing performance and subjective satisfaction. Therefore, a future study to evaluate the long-term effects of CMAT with a longer
training time is needed. Third, this study was not blinded, and the potential risk of type I error should be considered. Lastly, fitting verification results, such as real-ear measurements, were not included, and data on HA use time during the trial were not included in this analysis. Instead, fitting validation was conducted using sound field tests, and only those who had been using HAs for more than 3 months and for at least 8 hours a day were included in this study. Incorporating more objective fitting verification and HA use time monitoring with data logging would be beneficial in future studies.

Several points for improving this CMAT program can be considered in the future. First and foremost, AI technology can be adapted to enable 2-way interactions between the user and the program. The AI system can generate an almost infinite number of new questions and adjust the question difficulty level based on the user’s responses. Additionally, while this study used a chat-based interface for adults, more casual training methods, such as gamified training, could be adopted for pediatric patients with hearing loss.

Conclusions
In this study, we developed a novel mobile-accessible AT program with a chat-based interface and showed that it enhanced speech perception ability in experienced HA users. CMAT seems to increase the accessibility of and compliance to AT, and we expect that it will be used in all stages of the hearing rehabilitation process.

Acknowledgments
This work was supported by the Technology Development Program (S2964607), funded by the Ministry of SMEs and Startups of Korea. The funding organization had no role in the design or conduct of the study; in the collection, analysis, or interpretation of the data; in the decision to submit the article for publication; or in the preparation, review, or approval of the article.

Data Availability
The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available because they contain information that could compromise the privacy of the research participants.

Conflicts of Interest
JL is the developer of the software used in this study (Nara Information Co, Ltd).

Multimedia Appendix 1
Linear mixed model analysis of speech perception tests after arcsine transformation.
[DOCX File, 33 KB - mhealth_v121e50292_app1.docx]

Multimedia Appendix 2
Detailed results of 3 questionnaire surveys.
[DOCX File, 25 KB - mhealth_v121e50292_app2.docx]

Checklist 1
CONSORT eHEALTH checklist (V 1.6.1).
[PDF File, 1032 KB - mhealth_v121e50292_app3.pdf]

References


Abbreviations

- A1: artificial intelligence
- APHAB: Abbreviated Profile of Hearing Aid Benefit
- AT: auditory training
- ATG: auditory training group
- CG: control group
- CMAT: chat-based mobile auditory training
- CONSORT: Consolidated Standards of Reporting Trials
- HA: hearing aid
- K-CID: Korean version of Central Institute for the Deaf
- K-HHIE: Korean version of the Hearing Handicap Inventory for the Elderly
- K-IOI-HA: Korean version of International Outcome Inventory for Hearing Aids
- PTA: pure tone audiogram
- VCIT: Vowel and Consonant Imitation Test

© Jae Sang Han, Ji Hyung Lim, Yeonji Kim, Aynur Aliyeva, Jae-Hyun Seo, Jaehyuk Lee, Shi Nae Park. Originally published in JMRI mHealth and uHealth (https://mhealth.jmir.org), 9.2.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMRI mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Original Paper

Documentation Completeness and Nurses’ Perceptions of a Novel Electronic App for Medical Resuscitation in the Emergency Room: Mixed Methods Approach

Kin Cheung1*, PhD; Chak Sum Yip2,3*, MSN

1School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)
2Accident and Emergency Department, Tuen Mun Hospital, Hong Kong, China (Hong Kong)
3Quality & Safety Office, The Hong Kong Children's Hospital, Hong Kong, China (Hong Kong)
*all authors contributed equally

Corresponding Author:
Kin Cheung, PhD
School of Nursing
The Hong Kong Polytechnic University
11 Yuk Choi Road, Hung Hom
Hong Kong
China (Hong Kong)
Phone: 852 2766 6773
Email: kin.cheung@polyu.edu.hk

Abstract

Background: Complete documentation of critical care events in the accident and emergency department (AED) is essential. Due to the fast-paced and complex nature of resuscitation cases, missing data is a common issue during emergency situations.

Objective: This study aimed to evaluate the impact of a tablet-based resuscitation record on documentation completeness during medical resuscitations and nurses’ perceptions of the use of the tablet app.

Methods: A mixed methods approach was adopted. To collect quantitative data, randomized retrospective reviews of paper-based resuscitation records before implementation of the tablet (Pre-App Paper; n=176), paper-based resuscitation records after implementation of the tablet (Post-App Paper; n=176), and electronic tablet-based resuscitation records (Post-App Electronic; n=176) using a documentation completeness checklist were conducted. The checklist was validated by 4 experts in the emergency medicine field. The content validity index (CVI) was calculated using the scale CVI (S-CVI). The universal agreement S-CVI was 0.822, and the average S-CVI was 0.939. The checklist consisted of the following 5 domains: basic information, vital signs, procedures, investigations, and medications. To collect qualitative data, nurses’ perceptions of the app for electronic resuscitation documentation were obtained using individual interviews. Reporting of the qualitative data was guided by Consolidated Criteria for Reporting Qualitative Studies (COREQ) to enhance rigor.

Results: A significantly higher documentation rate in all 5 domains (ie, basic information, vital signs, procedures, investigations, and medications) was present with Post-App Electronic than with Post-App Paper, but there were no significant differences in the 5 domains between Pre-App Paper and Post-App Paper. The qualitative analysis resulted in main categories of “advantages of tablet-based documentation of resuscitation records,” “challenges with tablet-based documentation of resuscitation records,” and “areas for improvement of tablet-based resuscitation records.”

Conclusions: This study demonstrated that higher documentation completion rates are achieved with electronic tablet-based resuscitation records than with traditional paper records. During the transition period, the nurse documenters faced general problems with resuscitation documentation such as multitasking and unique challenges such as software updates and a need to familiarize themselves with the app’s layout. Automation should be considered during future app development to improve documentation and redistribute more time for patient care. Nurses should continue to provide feedback on the app’s usability and functionality during app refinement to ensure a successful transition and future development of electronic documentation records.

(JMIR Mhealth Uhealth 2024;12:e46744) doi:10.2196/46744
KEYWORDS

tablet computer; nursing documentation; paper resuscitation record; electronic resuscitation record; medical resuscitation; electronic medical record; documentation; resuscitation; electronic health record; nurses’ perception; traditional paper record; nurse

Introduction

Background

The completeness of documentation of critical care events in the accident and emergency department (AED) is essential for (1) the continuity of patient care, (2) medicolegal issues [1], (3) improving accessibility to critical information needed for research [2], and (4) serving as evidence for quality outcome measures [3]. Traditionally, documentation is performed on paper. Due to the fast-paced and complex nature of resuscitation cases, missing data is a common issue during emergency situations. As much as 60% of essential data fields in prehospital paper records can be incomplete [4]. A study in a trauma center also found incompleteness in 18% of the mandatory elements for trauma resuscitation [5].

Over the last 2 decades, there has been a global trend of switching to electronic medical records (EMRs). It was estimated that about 46% of AEDs in the United States used EMRs in 2010 [6]. This percentage is expected to increase in the future. Fully functional EMRs have been shown to improve efficiency in AEDs [7]. Despite the growth in usage, very few studies have explored the impact of EMRs in AED settings. Furthermore, the perceptions toward EMRs are mixed. A study found that nurses and physicians generally had a negative perception toward EMRs in the AED. EMRs are considered to be ineffective, redundant, and prone to error [8]. In contrast, another study found that nurses perceived that their productivity increased and care was better coordinated after implementing EMRs [9].

Among the limited studies in the area, a retrospective review of trauma resuscitations in AED settings showed that EMRs can improve documentation completeness [5,10]. However, there is a lack of studies on medical resuscitations, which are more common than trauma resuscitations [11]. Medical resuscitations are performed on triage category I and II patients with life-threatening conditions such as myocardial infarction, sepsis, and stroke [12]. Medical resuscitations differ from trauma resuscitations in that they do not follow a single protocol. The differences in management protocols can make the process of documenting medical resuscitations different from that of trauma resuscitations.

The EMR system used by the Hospital Authority of Hong Kong is called the Clinical Management System (CMS). It is an integrated platform that allows clinical users to manage the following daily clinical activities [13]: (1) obtain clinical data including consultation notes, laboratory, and imaging results; (2) document clinical activities; and (3) provide clinical decision support.

In AEDs in Hong Kong, medical records are still part paper and part electronic, with the patient’s clinical notes being documented on paper. This type of mixed documentation has been shown to hinder effective communication and utilization of information in either record [14].

Recently, there has been a trend of switching to electronic documentation in AEDs. In 2020, 3 of 18 AEDs in Hong Kong had switched to an EMR system called the eAED. It was expected that, by 2023, about two-thirds of AEDs would have switched to the eAED. The eAED is meant to replace paper charts previously used to document a patient’s progress [15]. Despite the gradual adoption of EMRs in AEDs in Hong Kong, the use of electronic documentation during medical resuscitations has not occurred owing to the time-critical, fast-paced nature and lack of a suitable application.

However, with advances in computer processing power, a tablet-based system could fill the gap. Documentation efficiency and data precision have improved when a tablet-based app was used, in comparison with a desktop EMR, during a simulation [16]. In Hong Kong, a tablet-based system called “eResus” is being developed by the Hong Kong Hospital Authority for medical and trauma resuscitation documentation. With the implementation of the eAED and eResus, documentation in AEDs would become fully electronic.

Aim and Objectives

This study aimed to evaluate the impact of a tablet-based app on documentation completeness during medical resuscitations. The research questions were the following:

- What are the differences between paper and electronic tablet-based records on the levels of documentation completeness?
- What are the perceptions of emergency room nurses regarding documentation completeness when using eResus?

Hypothesis

This study hypothesized that the completeness of resuscitation documentation using electronic tablet-based records would be higher than that using paper records.

Methods

Design

To answer research question 1, a randomized retrospective review of paper and electronic resuscitation medical records (N=528) was conducted using a documentation completeness checklist. The study was implemented during the transition from paper to electronic documentation, when only triage category II cases would be documented using the tablet-based app called eResus. Therefore, triage category II records were collected before (from November 2020 to December 2020) and after (from February 2021 to March 2021) implementation of the tablet-based eResus app. Paper records were collected before (Pre-App Paper) and 1 month after (Post-App Paper) implementation of eResus, while electronic records were collected 1 month after (Post-App Electronic) implementation of eResus. We randomly selected 176 records each for the

To answer research question 2, emergency nurses’ perceptions of the advantages, challenges, and areas for improvement of the electronic app for resuscitation documentation were obtained in individual interviews conducted in mid-April 2021, 3 months after the implementation of eResus. Reporting of the qualitative findings was guided by the Consolidated Criteria for Reporting Qualitative Studies (COREQ) [17], as delineated in the following sections, to enhance rigor.

**Ethical Considerations**

Ethical clearance (NTWC/REC/20098) from the study hospital and the Human Subjects Ethics Sub-committee of the Hong Kong Polytechnic University (HSEARS200200826001) was obtained before the commencement of the study.

**Data**

**Quantitative Data**

A documentation completeness checklist was established based on the literature and a review of the department’s current medical resuscitation event documentation. The checklist consisted of 5 essential domains (ie, basic information, vital signs, procedures, investigations, and medications) of medical resuscitation as illustrated in Multimedia Appendix 1. Face and content validity of the checklist were determined by 4 experts in the emergency medicine field [18]. Experts were invited based on the following criteria: (1) worked in an AED and (2) published at least one article related to the accident and emergency field. The expert panel consisted of 1 associate consultant, 1 medical officer, and 2 advanced practice nurses (1 of which was a Fellow in Emergency Nursing). The content validity index (CVI) was calculated using the scale CVI (S-CVI). The S-CVI is calculated based on the number of items in the scale rated by the expert as “quite relevant” or “highly relevant” [19]. The S-CVI was further analyzed by universal agreement (UA) among experts (S-CVI/UA) and the average (S-CVI/Ave). The checklist’s S-CVI/UA was 0.822, and the S-CVI/Ave was 0.939.

The medical resuscitation documents were reviewed against the validated checklist by a researcher (CSY), and intrarater reliability was determined to ensure consistency. Intrarater agreement was calculated using the Cohen kappa [19]. We evaluated 5 cases at week 0 and week 2. The agreement between the 2 records was considered acceptable at a $\kappa$ of 0.884 (95% CI 0.671-1.105; $P<.001$).

For each resuscitation documentation review, the researcher provided a dichotomous response of “Yes or No” for each item and identified the level of completeness based on the checklist. The patient diagnosis, length of medical resuscitation, initial triage category, and demographics including age and gender were also collected as part of the basic information. However, patient names and identification numbers were not collected. Data were kept anonymous by assigning codes only identifiable to the researcher.

**Qualitative Data**

Emergency room nurses were guided to discuss their thoughts on the eResus app’s features for documentation completeness through individual interviews with an onsite nurse who was one of the researchers (CSY). They understood the aim of the study, and their experience with the app was explored. Participants’ demographic data including age, gender, years of experience after graduation, and years of experience in their current specialty were collected for subsequent data analysis.

**Sample Size**

For the quantitative data, the Chi-square test was used to compare the differences in documentation completeness between the 3 groups. Based on the findings from a level 1 pediatric trauma center in 2015 [10], with an $\alpha$ of 0.05 and power of 0.80, a minimum sample size of 153 medical records per record set was required. To ensure an adequate sample size, 176 patient records were included in each of the paper and electronic record sets, resulting in a total sample size of 528 records (ie, 176 records each for the Pre-App Paper, Post-App Paper, and Post-App Electronic record sets).

For qualitative data, data saturation is the criterion to determine the sample size. Data are considered saturated when no new theoretical insights are gained from new data [20]. For this study, data saturation was achieved after 10 individual interviews, and 2 more interviews were conducted to confirm the data saturation.

**Recruitment**

The study was conducted in the AED of 1 hospital in Hong Kong. It is one the major local trauma centers providing 24-hour accident and emergency services and serves more than 190,000 patients per year, with over 300 resuscitation cases per month [21]. The tablet app was scheduled to be implemented in June 2020 but was postponed due to COVID-19. The app was eventually implemented in January 2021.

**Quantitative Data Collection Method**

In this study, we reviewed 2 types of resuscitation documents, namely paper and tablet-based resuscitation records. Completed resuscitation documents in paper format were attached to the patient’s CMS record by optical scanning as per usual practice. These records were stored in the CMS.

Training prior to the implementation of the electronic resuscitation record could lead to bias toward improved completeness of electronic documents [10]. Therefore, paper documentation records were collected before and 1 month after implementation of eResus to address this issue. First, baseline paper resuscitation records were collected prior to implementation of the eResus app (Pre-App Paper). After the implementation of the eResus app with training, there was a washout period of 1 month. After 1 month, the paper (Post-App Paper) and tablet-based (Post-App Electronic) resuscitation records were collected for analysis. Both paper and tablet resuscitation records involving trauma team activation or triage category I cases were excluded because the app did not cover these 2 types of cases at this stage.
For the purpose of this study, 3 lists of 3 months of case lists including eligible medical records from the Pre-App Paper, Post-App Paper, and Post-App Electronic record sets were retrospectively generated from the CMS and assigned a serial number. From each group, 176 records were randomly selected using a random number generator.

**Qualitative Data Collection Method**

In terms of qualitative data collection, purposive sampling was applied. The researcher conducted individual, voice-recorded interviews with each emergency nurse 3 months after the eResus implementation. The nurses were provided an explanation of the study, and written consent was obtained. The inclusion criteria included nurses (1) working in the AED of the hospital, (2) with experience using the eResus app, (3) who spoke Cantonese and were able to read English, and (4) working in their current position for more than 3 months.

Invitation emails were sent to colleagues. Eligible colleagues who replied to the email or expressed interest were invited to be interviewed according to their years of experience. Individual interviews were conducted in a quiet room or via Zoom. Each interview lasted about 1 hour or stopped when the interviewee felt that their viewpoint had been fully expressed. The interview guide is shown in Textbox 1.

### Textbox 1. Interview guide.

| Opening question: |
|-------------------|-------------------|
| Can you tell me your experience with using the eResus app until now? |

| Guiding questions: |
|-------------------|-------------------|
| 1. What are the main advantages and challenges with achieving high documentation completeness when using eResus in the resuscitation room? |
| 2. How do you think eResus can be improved to help you achieve better documentation completeness? |

**Data Analysis Methods**

For the quantitative data analysis, SPSS version 25 (IBM Corp) was used. Descriptive statistics such as means, standard deviations, frequencies, and percentages were used to present the study variables. Normality was tested using the Kolmogorov-Smirnov test, and the data were found to be not normally distributed. The Mann Whitney U test was used to compare mean ranks for age and clinical characteristics between the Pre-App Paper and Post-App Paper record sets as well as between the Post-App Paper and Post-App Electronic record sets to ensure the clinical characteristics of the 3 groups were comparable. Subsequently, Chi-square tests were used to compare the differences in proportions, such as the percentage of completeness between the Pre-App Paper and Post-App Paper record sets to determine any historical bias or effect from training and then between the Post-App Paper and Post-App Electronic record sets. Results with a P value < .05 were considered significant.

Each resuscitation record was manually reviewed against the study checklist for data element completeness. Each record was reviewed individually. Any incomplete data element was entered as an incomplete domain for the respective domain of the 5 domains, namely basic information, vital signs, procedures, investigations, and medications. For example, for records of the administration of 2 medications that use the same route, the record was treated as 2 separate data entries. If 1 of the data items (such as 1 missing medication name) was incomplete, the medication domain for that case was entered as incomplete. The number of entries for each domain of the resuscitation documentation was analyzed, delineating sections that were recorded at higher or lower frequencies.

For the qualitative data analysis, content analysis was performed [22]. First, the interview was audio-recorded and transcribed verbatim into Chinese. NVivo Pro 12 was used for data analysis. The researcher read through the transcript multiple times to become immersed in the data. Participants’ experiences with the eResus app, challenges, and possible solutions were extracted and summarized into meaning units. Third, each meaning unit was condensed and labelled with codes. Fourth, subcategories were identified by comparing the similarities and differences between different codes. Finally, the latent meanings of the subcategories were sorted into themes.

**Results**

### Quantitative Results

Tables 1 and 2 present the characteristics of the 5 domains of resuscitation documentation. Comparisons were made between the Pre-App Paper and Post-App Paper record sets. There were no significant differences in characteristics or documentation completion between the Pre-App Paper and Post-App Paper record sets (Table 3).
Table 1. Comparisons using the Mann-Whitney U test among the 3 resuscitation record sets in patient age; length of resuscitation; and total numbers of vital sign entries, procedures, investigations, and medications in the resuscitation documentation for patients requiring medical resuscitation (N=528).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Range (Mean (SD))</td>
<td>Range (Mean (SD))</td>
<td>U</td>
<td>P value</td>
<td>Range (Mean (SD))</td>
</tr>
<tr>
<td>Basic information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td>4-98 (61.6 (20.6))</td>
<td>3-101 (61.6 (22.0))</td>
<td>15,121</td>
<td>.84</td>
<td>4-100 (62.5 (21.1))</td>
</tr>
<tr>
<td>Length of resuscita-</td>
<td>3-216 (40.6 (30.2))</td>
<td>5-175 (41.7 (28.6))</td>
<td>14,673</td>
<td>.39</td>
<td>8-367 (40.5 (37.8))</td>
</tr>
<tr>
<td>Total number of vital</td>
<td>1-58 (10.9 (8.0))</td>
<td>2-47 (11.1 (7.1))</td>
<td>15,089</td>
<td>.68</td>
<td>2-72 (9.9 (7.6))</td>
</tr>
<tr>
<td>sign entries</td>
<td></td>
<td></td>
<td>15,050</td>
<td>.61</td>
<td>0-10 (1.5 (1.2))</td>
</tr>
<tr>
<td>Total number of proce-</td>
<td>0-7 (1.83 (1.34))</td>
<td>0-8 (1.76 (1.26))</td>
<td>15,091</td>
<td>.67</td>
<td>0-11 (5.70 (1.91))</td>
</tr>
<tr>
<td>dures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of investi-</td>
<td>0-11 (5.72 (1.89))</td>
<td>0-11 (5.49 (1.92))</td>
<td>175 (99.4)</td>
<td>16.06 (2)</td>
<td>75 (99.4)</td>
</tr>
<tr>
<td>gations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of medica-</td>
<td>0-12 (1.65 (2.07))</td>
<td>0-15 (1.70 (2.31))</td>
<td>176 (100)</td>
<td>63.50 (2)</td>
<td>176 (100)</td>
</tr>
<tr>
<td>tions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Gender differences among the 3 groups of resuscitation records for patients requiring medical resuscitation (N=528), as assessed using the Chi-square test.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Pre-App Paper record set (n=176), n (%)</th>
<th>Post-App Paper record set (n=176), n (%)</th>
<th>Post-App Electronic record set (n=176), n (%)</th>
<th>Difference among the groups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>χ2 (df)</td>
<td>χ2 (df)</td>
<td>χ2 (df)</td>
<td>P value</td>
</tr>
<tr>
<td>Female</td>
<td>84 (47.7)</td>
<td>84 (47.7)</td>
<td>86 (48.9)</td>
<td>0.06 (2)</td>
</tr>
</tbody>
</table>

Table 3. Differences in completion of the 5 domains of documentation between paper and electronic resuscitation records (N=528).

<table>
<thead>
<tr>
<th>5 domains</th>
<th>Pre-App Paper record set (n=176), n (%)</th>
<th>Post-App Paper record set (n=176), n (%)</th>
<th>Post-App Electronic record set (n=176), n (%)</th>
<th>Differences among groups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>χ2 (df)</td>
<td>χ2 (df)</td>
<td>χ2 (df)</td>
<td>P value</td>
</tr>
<tr>
<td>Basic information</td>
<td>113 (64.2)</td>
<td>105 (59.7)</td>
<td>128 (72.7)</td>
<td>6.86 (2)</td>
</tr>
<tr>
<td>Vital sign</td>
<td>116 (65.9)</td>
<td>108 (61.4)</td>
<td>158 (89.8)</td>
<td>40.97 (2)</td>
</tr>
<tr>
<td>Procedures</td>
<td>123 (69.9)</td>
<td>127 (72.2)</td>
<td>176 (100)</td>
<td>63.50 (2)</td>
</tr>
<tr>
<td>Investigations</td>
<td>101 (57.4)</td>
<td>93 (52.8)</td>
<td>128 (72.7)</td>
<td>16.06 (2)</td>
</tr>
<tr>
<td>Medications</td>
<td>158 (89.8)</td>
<td>163 (92.6)</td>
<td>175 (99.4)</td>
<td>15.24 (2)</td>
</tr>
</tbody>
</table>

For the post-app comparison, there were no significant differences in the characteristics, except the number of vital sign entries, between the Post-App Paper and Post-App Electronic record sets (Tables 1 and 2). To answer research question 1, there was a significantly higher completion rate for all 5 domains in the Post-App Electronic record set than in Post-App Paper record set (Table 3).

Qualitative Results

Categories

The objective of the qualitative study was to explore nurses’ perceptions of the use of eResus for documentation completeness. The main categories identified were “advantages of tablet-based documentation of resuscitation records,” “challenges with tablet-based documentation of resuscitation records,” and “areas for improvement of tablet-based resuscitation records” (Textbox 2).
Advantages of tablet-based documentation of resuscitation records

- Structural guidance for documentation
- Easy to review and edit documentation
- Comparable mobility to paper and superior to desktop

Challenges with tablet-based documentation of resuscitation records

- System loading speed and stability
- Familiarization with the app layout

Areas for improvement of tablet-based resuscitation records

- Need for speedy documentation and automated documentation

Data saturation was achieved after conducting individual interviews with 12 nurses. The mean age of the participants was 26.9 (SD 2.68) years, and 9 participants were female. Participants’ mean length of work experience was 4.46 (SD 2.20) years, with a range of 2.5 years to 8.5 years. Their mean length of work experience in AED was 3.3 (SD 1.87) years, with a range of 1.5 years to 8.5 years.

Advantages of Tablet-Based Documentation of Resuscitation Records

Structural Guidance for Documentation

The electronic app included an extensive database that encompassed the essential aspects of resuscitation documentation. The participants appreciated the app’s preset data fields that prompted users to input essential data during documentation.

(During documentation of blood glucose,) the interface displayed all the data field such as time, result, performer. You definitely cannot forget to input. [D168-169]

(After urinary catheter insertion) I may forget to write urinary output..., but eResus would prompt you if you did not enter. [H 170-173]

The application made sure that you have 2 colleagues to countercheck the medication and documented their name before administration. [C164-165]

The built-in logic set by emergency physicians and nurses provided clinical management support and guidance to users during documentation. Certain data fields were auto filled, saving more time for nursing care.

I found it convenient because the application would lead you how to input data in a step-by-step fashion. [A49]

After inputting the systolic blood pressure, it would automatically divert you to the diastolic blood pressure. [C70-71]

When asystole rhythm was chosen, the data field on blood pressure, pulse etc. would be prohibited from inputting...We no longer have to write “undetectable” over and over again. [E47-52]

Furthermore, the electronic app reduced the need for verbal order prescriptions and allowed structured electronic prescriptions, which are less prone to error during documentation and administration.

Verbal order was prone to miscommunication, distraction, and error in administration. [C146-147]

In the past, I would have to remember or write down physician’s verbal order... but now the drug name, dosage, infusion speed etc. would all be on the screen. [K 41-47]

Easy to Review and Edit Documentation

Medical resuscitation documentation has to be done contemporaneously during resuscitation. Electronic documentation can ensure legibility compared with handwriting, and users were able to review specific aspects of the documentation for completeness using in-app features.

Colleagues’ handwriting could be illegible; maybe everyone was in a hurry. And colleagues could misspell words, which could affect handover to ward colleagues. [E22-24]

Someone may accidently splash alcohol onto the paper chart, and the word would become illegible. [H209-210]

The application has a filter function which allows you to choose vital signs, allowing you to review vital sign inputs and trends or procedures, allowing you to review whether you have forgotten to document something. [K123-125]

Fragmented information was conveyed to the documenter from various sources, in a random sequence. Not all users can correctly recall the exact sequence of medical resuscitation as they document. Electronic documentation allowed the users time to edit the sequence rather than having to rewrite the whole resuscitation event on a new paper resuscitation record.

If the handwriting was too ugly and the time sequence is too out of place, such as the medication administration time did not align to the corresponding row, then I would cross out the whole paper chart and rewrite it. But now, eResus can easily amend it. [I 104-106]
Comparable Mobility to Paper and Superior to Desktop

Patients requiring medical resuscitation would often need to be transferred to another department for investigation or intervention. A tablet app can provide the mobility needed to document in various locations.

Let’s say the patient has to be escorted to computed tomography (CT). I would take the table to the CT suite (to continue the document). When we returned to the resuscitation bay, I could use the Bluetooth keyboard to continue the document. It’s better than desktop. [L 169-171]

You can bring it (to CT) like paper...don’t even need to bring pen, just use your fingers. [J 179-180]

Challenges With Tablet-Based Documentation of Resuscitation Records

System Loading Speed and Stability

Participants embraced the transition to electronic documentation. However, participants reported technical challenges due to the internet connection or app coding issues when using the electronic app that could compromise documentation completeness. The fast pace of medical resuscitations and contemporaneous nature of the documentation exacerbated the problem.

It has some technical problems...there was a time when it kept crashing and could not input data. [A58-60]

Sometimes, switching between different tab pages is rather slow. [F60]

The patient was...in asystole, we were conducting chest compression, and administering medication, but the application was still loading. [I 116-117]

Familiarization With the App Layout

Navigating through the various tab bars, interface, and data fields of the app was different from the paper resuscitation record that presented all the data fields on the same page. Users were required to tab multiple times to access the desired data fields on the tablet, which was more time-consuming. All participants received training prior to using the electronic app that could compromise documentation completeness. The fast pace of medical resuscitations and contemporaneous nature of the documentation exacerbated the problem.

I have to tab this and that before I could input data...if it was handwritten, it would be much quicker. [H 53-57]

This application has different tabs and options, which require a bit of thinking...it is like using a phone. [C33-43]

When you first encounter the application, you would need to spend time to learn the layout. But after you have become familiar with it, you would find the documentation process very smooth. [D147-149]

Areas for Improvement of Tablet-Based Resuscitation Records

All participants reported that speed of documentation was an important aspect in resuscitation documentation. During medical resuscitations, the case nurse was required to perform patient care and document contemporaneously. These resuscitation events were highly demanding and required speedy documentation:

Sometime, the documentation with eResus could take up lots of time. There were cases when we needed to document lots of medication right at the beginning. The application may not be able to document events in real time. [A63-64]

When you have many items pending documentation, you would be naturally prone to incomplete documentation. [B76-77]

The multitasking nature of the nursing practice posed competing demands between managing patient care and documentation, which required the nurses to compromise. Nurses decreased the frequency of taking vital signs. One participant said:

When handling less critical cases...I would take vital signs every 5 minutes (instead of 3) so that I can be more at ease when managing both the patient and documentation. [D113-117]

Users appreciated the auto retrieval of data from the Hospital Authority’s network and the auto fill of relevant fields. Relevant data previously inputted into the app were prepopulated either automatically or after the user’s approval:

It would auto-capture allergy status from CMS, an electronic health record system used in Hong Kong). [G 102-103]

It can retrieve the previous (vital sign) data. Then, I can tab it and paste it onto the data field...such as Glasgow Coma Scale score, etc. [H 44-46]

Faster documentation speed can improve documentation completeness and overall resuscitation quality. Participants believed that speed and improved care could be gained by automation:

If a multimonitor could automatically record vital signs and transfer data into the application, the user would spend less time inputting data and more time looking after the patient...or checking the (resuscitation) record for incompleteness. [D254-262]

(Automated vital sign recording) would be useful...But some factors could affect the reading’s accuracy; it should allow health care workers to verify the readings prior to documenting. [K 217-220]

Documentation speed can also be gained by flattening the user interface (UI). Participants found that, although organizing data fields into different categories and layers was logical, it made navigating through the layers inevitably slow. A more direct, intuitive UI is needed to improve the speed of data entry:

(The commonly used) items should be accessible with one tap. [G194]
The commonly performed investigations...that has many data fields should be more easily accessible.

[L 216-220]

Discussion

Principal Findings

This is the first study, particularly in Asia, to compare the completion rates of documentation between paper and tablet-based resuscitation records in the emergency room. Our results indicate that electronic documentation is promising, with a higher completion rate than with the paper format.

Our study results support the hypothesis that tablet-based documentation of resuscitation records results in a higher documentation completion rate than paper formats in all 5 domains. Previous studies in AED settings have been conducted to compare the completion rates of key data elements for trauma resuscitation records between paper and electronic formats for adult [23] and pediatric [10] trauma cases. Both studies found areas for improvement and degradation in the key data elements. However, it is difficult to directly compare the studies since the outcome measures were different and none of the studies in AED settings used tablet-based devices. Nevertheless, our study results were consistent with those of previous studies that supported that, with electronic resuscitation records, documentation completion rates were higher, particularly basic information such as case start time and disposal but not for serial vital signs [10,23]. Interestingly, although no difference was found in the completeness of documenting vital signs and interventions in these previous studies, our study showed improvement in the vital signs, procedures, investigations, and medication domains. This may be due to the differences in the inherent design of the EMRs and the use of a tablet-based device instead of desktops, as explained by interview participants. In addition, contrary to the concern of bias for improved documentation completeness caused by training [10], our study showed no statistically significant differences between the Pre-App Paper and Post-App Paper record sets. This indicates that missing data with the paper format could be consistent since the documentation format has not changed (such as using the same paper format).

Our qualitative results further explain the reasons why the tablet-based device could improve documentation completeness. The structural design of the tablet-based resuscitation record provided guidance that contributed to the completeness. This guidance provides support to the documenter via various clinical support features such as structured prescriptions, preset data fields, and preset documentation logic. This structural guidance was developed by consulting local emergency physicians and nurses working in the AED. The guidance mimicked the normal workflow and thus supported the documentation process. Similar results were found in a previous study in which nurses had higher confidence using the EMR when they perceived that their suggestions were used to customize the system [24]. This also implies that the tablet-based device will be considered useful if it is country, institution, and department-specific.

However, similar to other studies, our qualitative findings supported that nurses have to multitask during work, which has been shown to compromise documentation completeness [8]. Furthermore, documentation in an EMR was perceived to be more time-consuming and complex [25]. Our study participants also experienced similar concerns with slow app loading speeds and needing to navigate through various tabbed pages, which increases the complexity of documentation.

Our participants further suggested that future development of the app should include automation features that would spare the documenter from manually inputting individual data into the app. Automating data input can reduce the documenters need to tab multiple times before finding the desired data field. This would be particularly useful for vital signs, which were the most frequently documented in this study. The UI should be flattened to facilitate input of other common data fields, which supports the concept that data fields that are more frequently recorded should be located in readily accessible spots [26]. With the automation of vital signs and an improved UI, the documenter would be able to spend more time on patient care and document in real time.

Implications for Emergency Room Nurses

Most emergency room nurses believed that the transition from paper to electronic charting can improve the quality of resuscitation documentation and patient safety. This study clearly demonstrates the potential of electronic charting to achieve that. During the transition period, the nurse documenters faced general problems with resuscitation documentation such as multitasking and unique challenges such as software updates and a subsequent need to become familiar with the app’s layout. Therefore, systematic, periodic needs assessments of nurse documenters using tablet-based devices, followed by corresponding training, should be conducted. Emergency room nurses should also be actively involved in the development and implementation phases to ensure success in the transition and future development of electronic documentation. Automation functions should be considered during the development of future apps to improve documentation and redistribute more time for patient care.

Limitations

This study has some limitations. It was not able to demonstrate the effect on documentation accuracy, rates of medication errors, the quality of patient care, or the process of clinical decision-making. Furthermore, since the sample was obtained from 1 AED only, the study findings may not be generalizable to other AEDs or other acute ward settings where the staffing and workflow may be different. In addition, this study excluded trauma cases and cardiopulmonary resuscitation cases; thus, its findings cannot be generalized to all resuscitation room situations in the AED.

Conclusions

This study demonstrates that a statistically higher completion rate in 5 domains essential to resuscitation documentation was achieved with a tablet-based device than with the traditional paper resuscitation documentation. Refinement of the device should be ongoing and include consultation with the users. Further studies can expand the scope to involve all medical resuscitation cases across AEDs.
Acknowledgments

We acknowledge all the emergency room nurses who participated in the interviews in this study. This study did not receive any funding.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Documentation completeness checklist.

References


Abbreviations

- AED: accident and emergency department
- CMS: Clinical Management System
- COREQ: Consolidated Criteria for Reporting Qualitative Studies
- CT: computed tomography
- CVI: content validity index
- EMR: electronic medical record
- S-CVI: scale content validity index
- UA: universal agreement
- UI: user interface

©Kin Cheung, Chak Sum Yip. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 05.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Health Care Professionals’ Experiences With a Mobile Self-Care Solution for Low Complex Orthopedic Injuries: Mixed Methods Study

Jelle Spierings¹, MD; Gijs Willinge², MD; Marike Kokke¹, MD; Bas Twigt², MD, PhD; Wendela de Lange³, BSc; Thijs Geerdink², MD, PhD; Detlef van der Velde¹, MD, PhD; Sjoerd Repping⁴, MD, PhD; Carel Goslings², MD, PhD

¹Department of Traumasurgery, St Antonius Hospital, Utrecht, Netherlands
²Onze Lieve Vrouwe Gasthuis, The Department of Trauma Surgery, Amsterdam, Netherlands
³The Healthcare Innovation Center, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, Netherlands
⁴Department of Public Health, Amsterdam UMC, Amsterdam, Netherlands

*these authors contributed equally

Corresponding Author:
Jelle Spierings, MD
Department of Traumasurgery
St Antonius Hospital
Soestwetering 1
Utrecht, 3543AZ
Netherlands
Phone: 31 883203000
Email: j.spierings@antoniusziekenhuis.nl

Abstract

Background: To cope with the rising number of patients with trauma in an already constrained Dutch health care system, Direct Discharge (DD) has been introduced in over 25 hospitals in the Netherlands since 2019. With DD, no routine follow-up appointments are scheduled after the emergency department (ED) visit, and patients are supported through information leaflets, a smartphone app, and a telephone helpline. DD reduces secondary health care use, with comparable patient satisfaction and primary health care use. Currently, little is known about the experiences of in-hospital health care professionals with DD.

Objective: The aim of this study was to explore the experiences of health care professionals with the DD protocol to enhance durable adoption and improve the protocol.

Methods: We conducted a mixed methods study parallel to the implementation of DD in 3 hospitals. Data were collected through a preimplementation survey, a postimplementation survey, and semistructured interviews. Quantitative data were reported descriptively, and qualitative data were reported using thematic analysis. Outcomes included the Bowen feasibility parameters: implementation, acceptability, preliminary efficacy, demand, and applicability. Preimplementation expectations were compared with postimplementation experiences. Health care professionals involved in the daily clinical care of patients with low-complex, stable injuries were eligible for this study.

Results: Of the 217 eligible health care professionals, 128 started the primary survey, 37 completed both surveys (response rate of 17%), and 15 participated in semistructured interviews. Health care professionals expressed satisfaction with the DD protocol (median 7.8, IQR 6.8-8.9) on a 10-point scale, with 82% (30/37) of participants noting improved information quality and uniformity and 73% (27/37) of patients perceiving reduced outpatient follow-up and imaging. DD was perceived as safe by 79% (28/37) of participants in its current form, but a feedback system to reassure health care professionals that patients had recovered adequately was suggested to improve DD. The introduction of DD had varying effects on workload and job satisfaction among different occupations. Health care professionals expressed intentions to continue using DD due to increased efficiency, patient empowerment, and self-management.

Conclusions: Health care professionals perceive DD as an acceptable, applicable, safe, and efficacious alternative to traditional treatment. A numerical in-app feedback system (eg, in-app communication tools or recovery scores) could alleviate health care professionals’ concerns about adequate recovery and further improve DD protocols. DD can reduce health care use, which is...
important in times of constrained resources. Nonetheless, both advantages and disadvantages should be considered while evaluating this type of treatment. In the future, clinicians and policy makers can use these insights to further optimize and implement DD in clinical practice and guidelines.

*(JMIR Mhealth Uhealth 2024;12:e51510) doi:10.2196/51510*

**KEYWORDS**

application; direct discharge; eHealth; experience; healthcare professional; mixed method study; orthopaedic surgery; orthopaedic; policy; policymaker; self-care application; self-care; trauma surgery; utilization; virtual fracture clinic

**Introduction**

The global increase in the number of patients with trauma presents a major challenge to the already strained health care systems [1,2]. To achieve more sustainable health care, digital alternatives to face-to-face outpatient follow-up have been introduced as a supported strategy [3-5]. These alternatives, also known as “eHealth,” are defined by Eysenbach [6] as the intersection of medical informatics, public health, and business, referring to health services and information delivered or enhanced through the internet and related technologies. In a broader sense, the term characterizes not only a technical development but also a state of mind; a way of thinking; an attitude; and a commitment to networked, global thinking to improve health care locally, regionally, and worldwide by using information and communication technology.

Based on a British example, a Dutch teaching hospital implemented the Direct Discharge (DD) protocol to maintain the quality of care for patients with trauma in 2019 [7,8]. The DD protocol involves discharging patients from the emergency department (ED) without scheduled follow-up while providing patients with information through a self-care mobile eHealth app. DD significantly reduces secondary health care use (SHU) with similar levels of patient satisfaction and primary health care use (PUH) [9-11]. Based on these results and catalyzed by the COVID-19 pandemic, over 25 Dutch hospitals have implemented this protocol since 2019 [12].

The successful and sustainable adoption of digital health technology is complex and influenced by various factors at organizational, technological, and social levels [13,14]. This complexity is widely recognized in eHealth and eHealth evaluation frameworks [15-17]. The organizational and logistic benefits and patient satisfaction scores following DD have been well described in the literature [18-20]. However, the social aspects of the DD protocol for health care professionals remain underreported [21]. An in-depth exploration is warranted to better understand the adoption of the DD protocol within its social context, including insight into the experiences of stakeholders. The aim of this study was to explore the experiences of health care professionals with the DD protocol parallel to the implementation of this protocol in 3 Dutch hospitals to enhance durable adoption and improve the protocol.

**Methods**

**Design**

An observational mixed methods study was conducted among health care professionals from August 2021 to June 2022, parallel to the implementation of DD in 3 hospitals. Both quantitative and qualitative data were collected and analyzed separately by a quantitative team (GW and JS) and a qualitative team (WDL and Elke Mathijssen). The Bowen feasibility framework was used to organize both data sources with the following parameters: implementation, acceptability, preliminary efficacy, demand, and applicability [22]. After separate analyses, quantitative and qualitative data were triangulated with the Pillar Integration Process [23]. This study was reported according to the Good Reporting on a Mixed Methods Study (GRAMMS) criteria (Multimedia Appendix 1) [24].

**Context**

The 3 participating centers were urban, level-2 trauma centers with up to 3 locations per hospital, treating between 1200 and 1800 patients with low-complex traumatic musculoskeletal injuries annually. Each center had a similar size and structure. All centers consisted of 3 locations, with 1 large location focusing on low-to-high complex traumatic injuries and having an ED with more rooms compared with the other locations. The 2 other locations were smaller and had no particular focus on patients with trauma, but they treated low-complex patients with trauma if they sought care at these locations. Per center, all 3 locations have 1 team taking care of all patients. These teams consisted of (orthopedic) trauma surgeons, residents, plaster technicians, ED physicians, and ED nurses. In total, 217 eligible health care professionals were exposed to DD based on data provided by participating hospitals. Changes in tasks per health care professional are described in Figure 1. These changes apply per center, including all 3 hospitals per center. The variance in the number of employees was correlated with the size of the hospital. The 2 centers implemented DD in September 2021 and 1 in March 2022.
Figure 1. Treatment protocols before and after implementation of DD and changes in location, involved stakeholders and tasks. DD: Direct Discharge; ED: emergency department.

### Traditional Treatment
Before DD was implemented, patients were treated according to the local trauma protocols. These protocols consisted of immobilization or support with either a cast, sling, bandage, or splint and brief information about the injury at the ED. At least 1 outpatient follow-up appointment was scheduled at the plaster room or in the outpatient clinic within 2 weeks after the injury for review, extensive information, and definitive management planning.

### Direct Discharge Protocol
This protocol was derived from the British model of a Virtual Fracture Clinic (VFC) and adapted to the Dutch health care setting in 2019 [25]. In its Dutch adaptation, DD includes 11 treatment protocols for low-complex, stable traumatic orthopedic injuries with additional injury-related criteria (Multimedia Appendix 2) [25]. If patients met the injury-related inclusion criteria in Multimedia Appendix 2 and spoke Dutch or English fluently, they were included; no further predefined restrictions (eg, age or comorbidity) were used. Patients were excluded from the protocol at the ED if they had initial treatment in another hospital; follow-up in another hospital (eg, closer to home); multiple injuries; a reason for follow-up other than the injury (eg, social-care reasons); an eye-, motor-, or verbal-score <15 at presentation; or intoxication. With DD, patients were discharged directly from the ED without routine outpatient follow-up. They received a removable orthosis or a sling (eg, brace instead of a cast) and extensive information at the ED, summarized in a mobile self-care app (the VFC app). Patient eligibility for the protocol was re-evaluated on the next workday (within 24 hours) by a team consisting of an (orthopedic) trauma surgeon and a radiologist. If patients were incorrectly discharged directly based on the injury-related or social inclusion criteria during the second review the next day, then they were contacted by phone and scheduled for a face-to-face appointment. This re-evaluation was a standard procedure in both protocols and was used to check previous decisions of young doctors by a senior group of medical professionals based on the radiographs and electronic patient record.

### The VFC App
The VFC app provides self-care assistance through information, videos, and a helpline and can be downloaded for free at the Google Play Store and iOS App Store (Figure 2). Injury-specific leaflets with recovery information, treatment rules, and red flags were included. Furthermore, frequently asked questions, audiovisual exercise, immobilization, and analgesic instructions were included to assist patients. If patients required human contact in addition to the information, a helpline by phone operated by a health care professional was available during working hours. The VFC app aimed to increase self-management and self-care during recovery and to substitute face-to-face follow-up.
Figure 2. English and Dutch in-app screenshots of the Virtual Fracture Clinic app used in the Direct Discharge protocol.

Implementation Tools and Materials
Interested centers contacted 1 of the trauma surgeons or the main email address of the initiating Onze Lieve Vrouwe Gasthuis (OLVG) Hospital, Amsterdam, the Netherlands. The OLVG Hospital created several tools to assist the implementation in other hospitals. The tools were created to streamline the implementation process as hospitals had similar questions during implementation and medical doctors struggled with the implementation of eHealth in their daily practice, partly due to inexperience with implementation and accompanying barriers and facilitators. These tools consisted of an email address that health care professionals could contact, an implementation guide, a digital PowerPoint presentation with an overview of the concept, and an information set with standardized information. This information set included posters, pocket cards, training guides, smart phrases for electronic patient records, standardized discharge letters, and a tool to personalize the layout of the VFC app. Participating centers started implementation preparations 3 months before the actual start of DD as the standard of care. Based on the experiences of the implementation in previous centers, 3 months was considered a sufficient amount of time to implement DD. There were minor differences between hospitals to optimize local fit (eg, the available hours of the staffed helpline).

Study Population
Health care professionals involved in the in-hospital treatment of patients with eligible low-complex, stable traumatic orthopedic injuries on a daily basis were included in this study (Figure 1). Health care professionals were excluded from the final analysis if they did not provide the correct contact details or did not complete both surveys.

Sampling and Recruitment
A total of 4 weeks before the implementation of DD, the 217 potentially eligible health care professionals were asked to participate in the study through education moments, e-learnings, and by email. The health care professionals were given a survey distribution link by email to a preimplementation survey in Research Electronic Data Capture (REDCap; Vanderbilt University), a digital survey system [26]. Potentially eligible participants were remembered twice, 1 week after the initial email. Participants were excluded if they did not complete both surveys or if they did not provide any contact details to send the second survey to. Within the survey, consent for an additional semistructured interview was asked. Age, sex, occupation, medical specialty, and hospital were used to select a purposive sample among health care professionals who consented. During sample selection for the interviews, ED nurses and plaster technicians were underrepresented. Through an open call, nonresponders were recruited through email, after which the definitive sample was selected. Eligible health care professionals were contacted to schedule a web-based, semistructured interview. Health care professionals were reminded through email to complete the survey. The second survey was sent 3 months after the implementation of the DD protocol (Figure 3). We aimed to collect completed quantitative data from 100 health care professionals and qualitative data from 15 health care professionals.
Data Collection

Data were collected from surveys and semistructured interviews.

Surveys

A total of 2 surveys, a preimplementation survey and a postimplementation survey (Multimedia Appendix 3), with 46 questions, including close-ended questions, multiple-choice questions, 5-point Likert scales, visual analogue scales, and free-text questions, measured 5 Bowen feasibility parameters. As no golden standard for the evaluation of innovations exists, the surveys and topic list were developed by 4 researchers (JS, GW, BT, and TG) and checked by 2 experts on relevance: a professor in trauma surgery (CG) and an associate professor in process evaluations of health care innovations (Jaap Trappenburg). We pretested the survey with 5 health care professionals to improve clarity. After data collection, the preimplementation expectations were compared with the postimplementation experiences.

Semistructured Interviews

A total of 2 independent researchers from “The Healthcare Innovation Centre,” not involved in daily clinical practice or the VFC research team, conducted digital semistructured interviews to minimize social-desirability bias. The web-based interviews were held at least 8-12 weeks after the implementation of DD to warrant an optimal recall. The interviews were guided by a topic list based on literature, including the previously mentioned Bowen feasibility parameters (Multimedia Appendix 4). The research team piloted the topic list for clarity and completeness and modified it during data collection. A pragmatic choice was made to interview 5 different types of health care professionals from each participating center, which was deemed sufficient to get a good understanding of the experiences of health care professionals and to reach saturation. No new themes were identified in the last step of the analysis, indicating that saturation was reached.

Data Analysis

Quantitative Data

Quantitative data were analyzed using SPSS (version 27; IBM Corporation) [27]. Baseline characteristics and outcomes were reported descriptively using numbers and proportions for categorical variables and mean (SD) or median (IQR) for continuous variables as appropriate. The normal distribution of continuous data was assessed with a visual analysis of boxplots. The paired 2-tailed t test or Mann-Whitney U test was used to determine the statistical significance of parametric variables for normally and nonnormally distributed data, respectively.

Qualitative Data

Qualitative data were analyzed using an inductive approach. Data analysis started after the first 5 interviews. Interviews were audiotaped, summarized, and analyzed using NVivo (version 12; QSR International) [28]. A total of 2 researchers (WDL and Elke Mathijsen) used inductive analysis with methods to ensure reliability and validity [29,30]. The data was independently analyzed by 1 researcher (WDL), and another researcher (Elke Mathijsen) reviewed the analysis. Discrepancies and remarks were discussed until they reached a consensus about the interpretations of the data. Memos were made to track research decisions during analysis. Code saturation was reached when no new categories or themes emerged from the new raw data [31]. We considered 15 interviews sufficient to reach saturation and get a good understanding of the experiences of professionals. Therefore, the number of interviews was limited by a pragmatic choice of available time. The final themes were used to describe the value and feasibility of DD from the perspective of involved health care professionals.

Triangulation

After the separate quantitative and qualitative analyses, the findings were triangulated with a simplified approach of the Pillar Integration Process technique [23]. This approach uses a transparent and rigorous 4-stage technique for integrating and presenting qualitative and quantitative findings in a joint display (Microsoft Excel, 2018; Microsoft Corporation) [32]. A researcher (JS) presented the quantitative findings per study parameter, and another presented the qualitative findings (WDL). Dissimilarities and self-contained themes were objectified. These themes were merged by 1 of the researchers (Elke Mathijsen) into a meaningful narrative (the pillar), which was reviewed by 2 researchers (JS and WDL).

Patient and Public Involvement

Patients or health care professionals were not involved in the design, intervention, research question, or outcome measures of this study.
**Ethical Considerations**

This study, including the process analysis, was reviewed and approved by the Medical Ethical Committee of Utrecht, Netherlands (W21.261).

Patients provided consent for participation in the research, and could opt out at any time after request by e-mail. The original consent or IRB approval covers secondary analysis without additional consent. Data is de-identified. A data key is stored at the local hospitals in a secured map and coded file. This is only accessible to JS and GW. The accessible data has been de-identified as far as possible (e.g., age in years instead of date of birth). Patients received no compensation to participate in this research.

**Results**

**Demographics**

Of the 217 estimated eligible health care professionals, 128 started the primary survey, 42 did not complete the primary survey, and 49 did not complete both surveys (Figure 4). Of the 37 included health care professionals (response rate of 17%), 23 (62%) were female, and the median age was 38 (IQR 32-45) years. Current occupations were medical specialists (14/37, 38%), residents (14/37, 38%), plaster technicians (7/37, 19%), and ED nurses (2/37, 5%; Table 1). The baseline characteristics of health care professionals who solely filled out the primary surveys did not vary statistically significantly in age ($P=.98$) or sex ($P=.28$) as compared to those who filled out both surveys. A total of 15 health care professionals, 5 per hospital, participated in the web-based, semistructured interviews, of which 60% (9/15) were female.

![Figure 4. Flow diagram of the included health care professionals in the evaluation of the Direct Discharge (DD) protocol.](image-url)
Table 1. Baseline characteristics of health care professionals included in a mixed methods evaluation of Direct Discharge.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Survey (n=37)</th>
<th>Interview (n=15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female, n (%)</td>
<td>23 (62)</td>
<td>9 (60)</td>
</tr>
<tr>
<td>Age (years), median (IQR)</td>
<td>38 (32-45)</td>
<td>40 (32-44)</td>
</tr>
<tr>
<td>Hospital, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospital A</td>
<td>19 (51)</td>
<td>5 (33)</td>
</tr>
<tr>
<td>Hospital B</td>
<td>11 (30)</td>
<td>5 (33)</td>
</tr>
<tr>
<td>Hospital C</td>
<td>7 (19)</td>
<td>5 (33)</td>
</tr>
<tr>
<td>Current function, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trauma surgeon</td>
<td>5 (13)</td>
<td>3 (20)</td>
</tr>
<tr>
<td>Orthopedic surgeon</td>
<td>1 (3)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Emergency physician</td>
<td>8 (22)</td>
<td>2 (13)</td>
</tr>
<tr>
<td>Surgery resident in training</td>
<td>8 (22)</td>
<td>3 (20)</td>
</tr>
<tr>
<td>Surgery resident not in training</td>
<td>6 (16)</td>
<td>1 (7)</td>
</tr>
<tr>
<td>Plaster technician</td>
<td>7 (19)</td>
<td>3 (20)</td>
</tr>
<tr>
<td>ED(^a) nurse</td>
<td>2 (5)</td>
<td>3 (20)</td>
</tr>
<tr>
<td>Total</td>
<td>37 (100)</td>
<td>16 (100)</td>
</tr>
</tbody>
</table>

\(^a\)ED: emergency department.

**Implementation**

Qualitative data showed that the implementation strategy varied between hospitals and was adjusted to improve the local fit (e.g., available hours of the helpline or brace brands). Changes in tasks, immobilization material, and the number of follow-up appointments influenced the implementation experience the most. Some nurses reported that coworkers had difficulties applying the braces, sometimes because their schooling during implementation was suboptimal or delayed, for example, because someone was ill (quote 1 in Table 2). Health care professionals reported that adequate schooling and involvement of the previous responsible health care professional in the implementation process were essential to executing the DD protocol adequately. Mainly because the ED nurse, now responsible for applying the braces, had less experience immobilizing fractures than the plaster technician.
Table 2. Health care professionals’ quotes and associated feasibility parameters.

<table>
<thead>
<tr>
<th>Theme and quote number</th>
<th>Quote</th>
<th>Health care professional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implementation of Direct Discharge among health care professionals</td>
<td>Some of my colleagues have difficulty with the materials. How does it work and what goes where? It takes a bit longer for some of them to get the hang of it. Having to learn so many new things sometimes causes resistance.</td>
<td>ED nurse</td>
</tr>
<tr>
<td>Acceptability of the Direct Discharge among health care professionals</td>
<td>This is a significant improvement for patients and appeals to their autonomy and control, as well as their own influence on the healing process. I believe it is motivating and in line with the current times.</td>
<td>ED nurse</td>
</tr>
<tr>
<td></td>
<td>I no longer have to do these routine outpatient clinical check-ups. I could only provide limited contributions besides providing information, allowing me to have more peace and tranquility in the clinic. I can use that time for other patients to add more value.</td>
<td>Trauma surgeon</td>
</tr>
<tr>
<td></td>
<td>Both among doctors and nurses, DD has been widely embraced and well implemented, but for both professions, it requires valuable extra minutes due to additional explanation. Currently, the workload is very high.</td>
<td>Resident</td>
</tr>
<tr>
<td></td>
<td>But if that care is taken away from us, I do believe that we have a responsibility towards the patients to ensure the proper transition of that care.</td>
<td>Plaster technician</td>
</tr>
<tr>
<td>Preliminary efficacy of the Direct Discharge among health care professionals</td>
<td>We now have a tool in our hands to change healthcare without it deteriorating, which convinces people who tended towards over-treatment.</td>
<td>ED physician</td>
</tr>
<tr>
<td></td>
<td>The quality is not affected, assuming the doctor was already good. It is mainly more efficient. Information provision has improved. It has become more modern. I think DD is not worse, but we’re not certain yet.</td>
<td>Trauma surgeon</td>
</tr>
<tr>
<td></td>
<td>We are sometimes called about 2-3 times per day on the fracture line. I don’t think that’s a bad score.</td>
<td>Plaster technician</td>
</tr>
<tr>
<td>Demand for Direct Discharge among health care professionals</td>
<td>I was already familiar with DD because I was looking for good and reliable information on injuries during my training as an ED nurse. I came across the article published by OLVG and started using the app in my work.</td>
<td>ED nurse</td>
</tr>
<tr>
<td></td>
<td>Every day, a few patients are treated through the app. I am starting to notice the reduction in daily practice!</td>
<td>Trauma surgeon</td>
</tr>
<tr>
<td></td>
<td>In my work, it has changed that we see less patients, but new things have also been added. However, we no longer see minor injuries. The easier type of care has decreased a bit.</td>
<td>Plaster technician</td>
</tr>
<tr>
<td>Applicability of the Direct Discharge among health care professionals</td>
<td>I see the advantage of a brace instead of a cast, a great improvement. I would also prefer DD myself.</td>
<td>Resident</td>
</tr>
<tr>
<td></td>
<td>The walking boot is difficult to fit, which poses a risk of misuse. This results in most complications being caused by misuse. What does that do to the recovery process?</td>
<td>ED physician</td>
</tr>
</tbody>
</table>

**Acceptability**

**Overview**

Quantitative data showed a median satisfaction with treatment of 7.8 (IQR 6.8-8.9) on a 10-point scale. This finding complemented qualitative data, as most health care professionals were satisfied with DD (quote 2 in Table 2). Health care professionals suggested that in-app patient feedback, such as pain scores or patient-reported outcome measures, could further increase health care professional satisfaction levels and address health care professionals’ possible concerns about adequate recovery. Some health care professionals were hesitant about DD due to its novelty, limited education during implementation, changes in work activities, and concerns about the short- and long-term outcomes of certain injuries.

**VFC App Acceptability**

The introduction of the VFC app at the ED yielded both advantages and disadvantages. Both qualitative and quantitative data showed that introducing the VFC app and treatment information in the ED required more time than the traditional treatment (19/37, 52%; Multimedia Appendix 5). Qualitative data showed that this time decreased once DD was fully implemented but remained longer than traditional treatment. Despite the increased time to inform patients, the perceived reduction in logistics at the outpatient clinic was perceived as more valuable (quotes 3 and 4 in Table 2). Both data sources underline the benefits of the app in terms of uniform, on-demand, and adequate information for patients after discharge. The downsides of the VFC app include less personal health care and the current limited language availability (Dutch and English).
Workload and Job Satisfaction

Qualitative data showed that workload and job satisfaction decreased slightly after the introduction of DD. Quantitative data did not support this, as no statistically significant differences between expectations and experiences were found regarding workload ($P=.37$) and increased job satisfaction ($P=.42$). Plaster technicians reported that the introduction of DD has led to losing a “fun” part of their job. Some reported they could provide less service for patients and felt responsible for educating ED nurses who had less experience with the immobilization of fractures compared with them (quote 5 in Table 2).

Preliminary Efficacy

Quality of Care

Quantitative data showed that the quality of care with the DD protocol is perceived as comparable to traditional treatment (25/37, 67%). Health care professionals reported an improvement in the quality of information and uniformity (30/37, 82%; Multimedia Appendix 6). No statistically significant differences were found between expectations and experiences of quality of care ($P=.86$), quality of information ($P=.42$), and quantity of information ($P=.18$). Qualitative data supported these findings, highlighting the benefits of uniform, injury-specific information (quotes 6 and 7 in Table 2). In both the survey and interviews, most (27/37, 73%) health care professionals reported a reduction in outpatient follow-up and injury-related imaging. Qualitative data showed that (orthopedic) trauma surgeons perceived the reduction as beneficial, whereas some plaster technicians experienced the reduction as a disadvantage. Most residents and ED physicians reported that the logistical benefits at the outpatient clinic outweighed the slight increase in time at the ED.

Perceived Safety

Both data sources showed that most (29/37, 79%) health care professionals perceived DD as safe and that sufficient scientific evidence exists to treat patients safely. Health care professionals assumed that patients had fully recovered if they did not contact the hospital again. Nevertheless, they proposed a numerical feedback system in the VFC app to ensure adequate recovery and alleviate concerns about the poor long-term functional outcomes of their patients. The frequency of helpline use was low ($\leq$5 times per week; quote 8 in Table 2). Some residents reported that the introduction of DD decreased their exposure to low-complex traumatic injuries, which might influence their learning curve in the future. Furthermore, the lack of follow-up introduced the tendency for some residents to be more explanatory at the ED. However, some residents stated that during crowding at the ED, they limited their explanation to the VFC app with minimal instructions. The frequency of calling varied per injury, with patients with a greenstick fracture rarely requiring contact. Reasons to call the helpline were similar among the 3 centers. Patient questions were related to a poor recall of the ED visit, suboptimal information provision at the ED, doubts about their recovery, or insufficient reading of the app’s content. The daily multidisciplinary radiologic evaluation and helpline were considered effective safety nets.

Demand

Qualitative analysis revealed that some health care professionals had previous knowledge or experience with the DD protocol (quote 9 in Table 2). The COVID-19 pandemic positively influenced their perceptions of digitally assisted care. Some health care professionals reported that it fits the general demand to develop more efficient outpatient follow-up models and that the DD protocol is an example of “tomorrow’s health care.” Reduced outpatient follow-up, hospital use (eg, treatment rooms or parking lots), and staff were mentioned as benefits. Health care professionals expected the DD protocol to stay and intended to continue using it. Health care professionals reported that the DD protocol could improve patient empowerment and enhance self-management and independence, especially among younger patients. Health care professionals had different experiences integrating the DD protocol into their daily activities. Orthopedic trauma surgeons perceived a decrease in patients at the outpatient clinic following DD implementation and stated that their workload was starting to decrease due to DD (quote 10 in Table 2). Plaster technicians also reported a decrease in patients. However, they perceived this sometimes as a disadvantage as they no longer treated these low-complex injuries, which is considered a loss of a fun part of their job. No reduction in workload was perceived among plaster technicians due to additional tasks and an increase in a patient population with more complex injuries (quote 11 in Table 2).

Residents reported that extensive information has changed from the ED and outpatient clinic to solely the ED. The ED nurses reported applying fewer casts and more orthoses, such as braces or walking boots.

Applicability

Before implementation, some health care professionals expressed concerns about the incorrect use of immobilization materials. However, after implementation, the types of immobilization materials were perceived as adequate, with 74% (27/37) of health care professionals finding the braces easy to use for patients (Multimedia Appendix 7). The braces were less immobilizing than a cast, which was perceived as a benefit (quote 12 in Table 2). The less immobilizing treatment regimens could result in an earlier return to full function. However, some were concerned that the braces could lead to inferior long-term functional outcomes, even though the scientific evidence for these concerns was lacking (quote 13 in Table 2).

Discussion

Principal Findings

Dutch health care professionals considered the DD protocol a safe and satisfactory alternative to traditional treatment, leading to a perceived decrease in SHU. Although providing information with DD required more (explanation) time for residents and physicians at the ED, the logistical benefits (ie, reduced number of follow-up appointments) outweighed the slight increase in time. Before and after implementation, no statistically significant differences were found regarding workload or job satisfaction.
However, qualitative data reported benefits in workload and job satisfaction for trauma surgeons and residents, and disadvantages for some plaster technicians in terms of job satisfaction and workload. Furthermore, health care professionals reported increased quality of information and comparable quality of care. Almost all health care professionals would like to continue using DD after implementation. To improve DD, several new app functionalities were suggested, and the earlier involvement of stakeholders who performed new tasks was suggested.

**Comparison With Existing Literature**

The study results show that DD is an acceptable alternative to traditional follow-up for health care professionals. Health care professionals reported similar, high satisfaction scores with treatment as previously reported patient satisfaction scores for similar protocols [18,33]. To further improve health care professionals’ satisfaction scores, health care professionals in this study suggested a short numerical feedback system to monitor injury recovery remotely (e.g., recovery questions after 3 months). In the literature, health care professionals have also mentioned this as an important feature of eHealth developed for patients with musculoskeletal injuries [34,35]. A challenge to incorporating this is that these functions would require a more enhanced app that complies with current laws and regulations for data storage and requires substantial financial investment. In addition, this request might also be a sign of early-stage adoption, where health care professionals feel a bit uncertain about the patients they would normally see for follow-up but who are now out of sight. As time progresses, it seems likely they will feel more comfortable not seeing these patients anymore, as this is then considered standard of care.

Preliminary efficacy was partly in line with the literature [18,20,33]. Most health care professionals considered DD a safe alternative to face-to-face follow-up, leading to a perceived reduction of SHU. This finding is consistent with previous articles, which concluded that DD is safe based on low, comparable complication numbers, with significantly reduced SHU compared to traditional treatment [9,18,33]. The perceived reduction in SHU varied among stakeholders. The reduction was perceived as most beneficial in qualitative data regarding workload for (orthopedic) trauma surgeons and least beneficial for plaster technicians, as they enjoyed this particular part of their workload. This study has been unable to demonstrate statistically significant differences regarding workload and job satisfaction before and after implementation. Even though the introduction of DD at the ED increased valuable treatment time for some stakeholders, which decreased but remained longer after implementation, almost all reported that the benefits of less SHU outweighed this downside. These findings are interesting, as the increase in time has been reported as one of the most important personal barriers to implementing and adopting eHealth but was not reported as an important disadvantage in this study [14,35].

Similar quality of care and increased quality of information after the introduction of DD are in accordance with earlier findings [11,20,36]. Particular advantages of DD were uniform, injury-specific, and on-demand information, consistent with a study pointing out the current heterogeneity in treatments for these injuries and the demand for uniform treatment [37]. The disadvantages of DD were the care being less personal, the loss of care activities for several stakeholders, and the inability to monitor recovery. A previous study described that VFCs might influence learning curves as residents become less experienced with the follow-up and full recovery of these injuries [38]. This was not supported by our results. However, a lack of follow-up has led to a tendency for some (often inexperienced) residents to be more explanatory at the ED to ensure the patient had received all information during the only contact moment. To reassure themselves and identify the small group of patients with persistent complaints, some health care professionals suggested a feedback system to ensure they had recovered adequately. Such a system should not increase workload but only filter patients with remaining questions, pain, or complications. This should be developed with patients and health care professionals and could contain anchor-based questions or patient-reported outcome measures with predefined cut-off values.

Health care professionals report that DD fits the public demand to develop more efficient (outpatient) care, and digital assistance could help deal with constrained resources. This partly aligns with the literature, as attitudes toward the usefulness of eHealth vary [39-41]. Almost all health care professionals would like to continue using DD after its implementation. This finding complements current literature showing the widespread implementation of DD protocols and VFCs in the United Kingdom, Australia, New Zealand, and India [8,33,42,43]. This further strengthens the idea of at least adequate acceptance and the general applicability of this concept in different countries. The applicability of DD varied among stakeholders and was influenced by the increase or loss of tasks after implementation of the protocol and their experience with that particular increase or loss of tasks. Even though evaluation of applicability for health care professionals and health care professional satisfaction with the protocol seems vital to determining feasibility, it has not been reported previously for DD protocols, despite the many centers that have implemented similar protocols. Health care professionals reported that early stakeholder involvement during implementation could be beneficial to cope with the changes in tasks and transfer of knowledge, which aligned with the results of Logishetty [21] reporting the importance of early stakeholder involvement during VFC implementation in a quality improvement approach. The lack of early involvement of stakeholders and limited feasibility among health care professionals are known risks for nonadoption or abandonment [3,44,45].

**Strengths and Limitations**

This study has several strengths. To date, this is the first study to explore the in-depth views of several stakeholders involved in DD protocols and VFCs. The COVID-19 pandemic has accelerated the implementation of DD in the Netherlands, emphasizing the demand for a shift in patient care, where eHealth alternatives have become the new standard [46]. The views and perceptions of health care professionals about different eHealth services are valuable to further tailor these services to their needs and preferences. A second strength of
this study is the multidisciplinary involvement of the research team during evaluation and analysis. This approach ensured that the data were analyzed from all possible angles. A third strength is a mixed methods approach to evaluating DD because the separate collection of both data sources combined with the triangulation increased the likelihood of our results being a realistic representation of the daily clinical activities of the involved health care professionals. The fourth strength is the use of a validated framework to evaluate DD, which allowed for a structured insight into each feasibility parameter. The fifth strength was the heterogeneity in the sizes and locations of the three study sites. This allowed us to investigate the applicability within different types of hospitals.

This study also has limitations that need to be addressed. The first limitation is the small quantitative sample size and low response rate. Even though the response rate among health care professionals is 10% to 15% lower compared to patient studies, the quantitative sample size remains limited [47]. Nevertheless, for almost all parameters, quantitative and qualitative data were similar, indicating a realistic response from most health care professionals. A third limitation is that ED nurses were not involved in the initial design of this study and did not provide quantitative data. Nevertheless, they have been included in the qualitative data to strengthen the study results.

**Implications for Clinicians and Policy Makers and Future Perspectives**

DD protocols reduce SHU without negatively influencing PHU, satisfaction, complications, or functional outcomes [17]. By assisting patients through the VFC app, they can receive care at home, potentially reducing health care costs [9]. Health care costs in the Netherlands have increased drastically, particularly due to specialized in-hospital care. In addition, there is a growing shortage of personnel. Digitally assisted solutions are suggested to cope with the rise of these costs and limited personnel, but implementation remains difficult [14]. The technical (eg, compliance and complication numbers) and logistic outcomes of this concept are well explored, but the social and cultural elements of DDs have not been explored. These findings help optimize future implementation strategies for eHealth in orthopedic and trauma surgery by providing preconditions and learning lessons such as early stakeholder involvement. These findings might be generalizable for other short treatment processes in other (surgical) departments with high volumes of relatively low-complex surgical patients (eg, low-complex dermatology or otorhinolaryngology).

As DD is introduced rapidly to cope with constrained financial and human resources and health care professionals expect DD to stay, a thorough evaluation of caregivers and patients is essential to ensuring sustainable adoption. Results of this study have improved our DD protocol and have led to 2 corresponding major points of improvement among all stakeholders and many points of improvement, such as ideas for illustrations, adjusted language (levels), and adjustable font size from 1 or more health care professionals. Future studies should focus on co-designing numerical feedback with patients and health care professionals. Furthermore, a thorough evaluation of patient perspectives should be performed to gain insight into the end users’ perspective on this innovation. Particularly to investigate the potential link between different levels of literacy (health literacy, digital literacy, and literacy) and health equity, as suggested in previous studies [48,49].

**Conclusion**

Health care professionals perceive DD as an acceptable, applicable, safe, and efficacious alternative to traditional treatment. A short numerical feedback system could alleviate concerns about a full recovery and further improve DD protocols. DD can reduce SHU, which is important in times of constrained resources. Nonetheless, both advantages and disadvantages should be considered while evaluating this type of protocol. In the future, clinicians and policy makers can use these insights to further optimize and implement DD and VFC in clinical practice and guidelines.

**Conflicts of Interest**

The institutions of the authors affiliated with St Antonius Hospital and Onze Lieve Vrouwe Gasthuis (OLVG) Hospital have received funding to support this work through an unrestricted grant from the Dutch organization for health care sciences and health care innovation, ZonMw, the Hague, Netherlands (516012524).

Multimedia Appendix 1
Assessment of Good Reporting on a Mixed Methods Study (GRAMMS) criteria for the study.
[DOCX File, 24 KB - mhealth_v12i1e51510_app1.docx ]

Multimedia Appendix 2
Additional criteria and immobilization for low-complex, traumatic orthopedic injuries included in the Direct Discharge Protocol.
[DOCX File, 15 KB - mhealth_v12i1e51510_app2.docx ]

Multimedia Appendix 3
Surveys used to evaluate Direct Discharge among health care professionals.
[DOCX File, 140 KB - mhealth_v12i1e51510_app3.docx ]

Multimedia Appendix 4
Topic list for health care professionals the Direct Discharge protocol.

Multimedia Appendix 5
5-Point Likert scale regarding applicability of the Direct Discharge among health care professionals.

Multimedia Appendix 6
5-Point Likert scale regarding preliminary efficacy of the Direct Discharge among health care professionals.

Multimedia Appendix 7
5-Point Likert scale regarding applicability of the Direct Discharge among health care professionals.

References


The Effectiveness of a Cell Phone eHealth App in Changing Knowledge, Stigmatizing Attitudes, and Intention to Seek Help Associated With Obsessive-Compulsive Disorder: Pilot Questionnaire Study

Antonio Chaves, PhD; Sandra Arnáez, PhD; Gemma García-Soriano, PhD

1Departamento de Orientación Educativa, IES Cid Campeador, Conselleria d'Educació, Cultura i Esport, Valencia, Spain
2Departamento de Personalidad, Evaluación y Tratamientos Psicológicos, Universitat de València, Valencia, Spain

Corresponding Author:
Gemma García-Soriano, PhD
Departamento de Personalidad, Evaluación y Tratamientos Psicológicos
Universitat de València
Avda Blasco Ibáñez, 21
Valencia, 46010
Spain
Phone: 34 963983389
Email: gemma.garcia@uv.es

Abstract

Background: Obsessive-compulsive disorder (OCD) is a disabling disorder associated with high interference in people’s lives. However, patients with OCD either do not seek help or delay seeking help. Research suggests that this could be explained by poor mental health literacy about the disorder and the associated stigma.

Objective: This study aims to evaluate the feasibility, acceptability, and preliminary effectiveness of a mental health mobile app, esTOCma, developed to improve knowledge about OCD and its treatment, increase help-seeking intention, and reduce stigmatizing attitudes and social distance associated with OCD.

Methods: We used preintervention, postintervention, and 3-month follow-up assessments in this single-arm pilot intervention. Overall, 90 participants were recruited from the community using the snowball sampling method. We used esTOCma to defeat the “stigma monster” over the course of 10 missions. The participants completed the sociodemographic information and Obsessive-Compulsive Inventory–Revised at preassessment and an acceptability questionnaire at postassessment. All other measures were completed at the preassessment, postassessment, and 3-month follow-up (i.e., the Spanish Mental Illness Stigma Attribution Questionnaire–27, the General Help-Seeking Questionnaire, the Social Distance Scale, and the Mental Health Literacy Questionnaire).

Results: Of the 90 participants from the community that were assessed for eligibility, 86% (n=78) were allocated to intervention. Of these 78 participants, 79% (n=62) completed the game and answered the postintervention assessment (completer group). Overall, 69% (43/62) of the participants also completed the 3-month follow-up assessment. The participants completing the study were older (P=.003) and had a higher baseline knowledge of OCD (P=.05). The participants took an average of 13.64 (SD 10.50) days to complete the intervention, including the pre- and postassessments. The participants spent an average of 4.56 (SD 3.33) days completing the 10 missions included in the app. Each mission took a mean of between 2 (SD 3.01) and 9.35 (SD 3.06) minutes. The app was rated as useful or very useful by the vast majority of participants 90% (56/62). Moreover, 90% (56/62) of the participants reported that they had learned or learned a lot, and 98% (61/62) of the participants reported that they would recommend the app to a friend. Repeated measures ANOVA (43/62, 69%) showed that after the intervention participants showed an increased knowledge of mental health and intention to seek help as well as fewer stigmatizing attitudes and less social distance.

Conclusions: Preliminary data show that esTOCma is a feasible and acceptable app, and after completing its 10 missions, there is an increase in the understanding of OCD and help-seeking intention along with a decrease in the social stigma and social distance associated with OCD that lasts for at least 3 months. The results support the potential of technology-based interventions to increase the intention to seek help and reduce the stigma associated with OCD. A larger, community-controlled study is also recommended.
**Introduction**

Background

Obsessive-compulsive disorder (OCD) is a clinically heterogeneous condition characterized by obsessions, compulsions or both that cause clinically significant levels of distress or functional impairment [1,2]. At present, effective treatments exist for OCD [3,4]; however, many people delay seeking treatment [5,6], and this may contribute to its chronic course [7,8]. Research suggests that this delay in seeking treatment by patients with OCD may be explained by social stigma and poor mental health literacy (MHL) about the disorder [5,9,10].

To date, many interventions have proved to be effective in reducing stigma associated with mental disorders, producing knowledge, and achieving attitudinal improvements [11-13]. Data show that contact and education strategies produce small-to-medium short-term reductions in stigmatizing attitudes, and there is limited evidence on long-term effectiveness [14,15]. A few of these interventions have benefited from the advantages of new technologies (eg, video games and electronic contact with patients), showing medium effects on reducing social stigma and suggesting that new technologies are a useful tool to decrease stigmatizing attitudes toward mental disorders, at least in the young population (the mean age of participants ranged between 15.7 and 24 years) [16]. However, none of these studies have examined the benefits of using app-based interventions. Furthermore, most of these interventions have focused on reducing the stigma of mental disorders such as schizophrenia [17-19], depression [20-22] or bipolar disorder [23,24].

In the case of OCD, programs are scarce, and most proposals have studied the impact of the educational mechanism [25-28], showing a reduction in stigmatizing attitudes and an improvement in participants’ MHL. Furthermore, a proposal has reported a significant reduction in social stigma and social distance using an indirect contact strategy through a video of a patient with OCD and a family member talking about their experience with the problem [29]. However, none of these interventions are based on innovative technologies, and only 1 of them includes >1 intervention strategy (ie, contact and education) [29]. In this context, as a response to the limited interventions focused on OCD and the need to bridge the gap between the onset of symptoms and seeking help, a gamified mental health app named esTOCma has been developed [30] (refer to the Methods section). In this way, the intervention will take advantage of mobile mental health interventions as low-cost tools that are available 24 hours a day for a large number of people [31,32], in addition to including gaming benefits such as providing immediate feedback, motivating users to achieve goals, and being easy to use [33].

**Objective and Hypothesis**

The aim of this study is to explore the feasibility, acceptability, and preliminary effectiveness of the beta version of the eHealth mobile app esTOCma. Regarding the effectiveness, and based on the reviewed literature on interventions to reduce stigma and increase knowledge of OCD [25-29] and other mental disorders [11,14,16], we hypothesize that the esTOCma intervention will (1) improve knowledge associated with OCD and its treatment, (2) decrease stigmatizing attitudes and social distance, and (3) increase intention to seek help. Moreover, we hypothesize that changes will be maintained at the 3-month follow-up assessment.

**Methods**

**Study Design**

This study was a single-arm pilot intervention with 3 measures at the pretest, posttest, and 3-month follow-up. Data were collected from people residing in Spain.

**Participants and Procedure**

Data were collected from a convenience sample. Participants were recruited from the general community and university setting by snowball sampling after providing relevant information via face-to-face classes and inviting the participants to share information about the study with their acquaintances. The inclusion criteria for this study were as follows: (1) being aged >18 years, (2) residing in Spain, (3) owning a smartphone, and (4) self-reporting not having an OCD diagnosis. Interested participants were invited to participate in a study consisting of downloading an app (esTOCma) in an Android Package Kit file format, playing with it, and completing a set of questionnaires before and after using the app. The participants performed all the tasks individually at home and at their own pace at the time they deemed most convenient. Furthermore, they chose the rate at which they completed the game, although the app recommends completing 1 mission a day.

The participants signing the informed consent form were given an identification number automatically generated by the app. The data from the questionnaires and the game were matched with the personal player ID. The participants were randomized by the app using a sampling without replacement method to 1 of 6 vignettes describing a person with obsessive-compulsive (OC) symptoms from 1 out of 6 types of content (ie, aggression or harm; sexual; religious, blasphemous, or immoral; contamination or washing; doubts or checking; or superstition, symmetry, or order). The vignettes consisted of descriptions of patients presenting with OC symptoms, with similar severity. Furthermore, interference and impairment in quality of life were assessed. The vignettes were selected from real clinical cases [34]. All of them meet the diagnosis criteria following the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, and the labels of obsessions or compulsions were avoided. All the described patients were referred to as A; sex...
was not specified, and they were middle aged. Furthermore, all descriptions included between 166 and 175 words.

After reading the assigned vignette, the participants were asked by the app to complete the preintervention measures. Most of them (i.e., Spanish Mental Illness Stigma Attribution Questionnaire–27 [AQ-27-E], General Help-Seeking Questionnaire [GHSQ], Social Distance Scale [SDS], and MHL questionnaire) were answered in reference to the assigned vignette. Only after completing the preassessment were the participants able to begin playing. The participants who completed the 10 missions and finished the game were asked through the app to complete the postintervention assessment and the 3-month follow-up assessment. To encourage the participants to complete the game, they were given entries to a prize draw for a voucher to spend on the web.

Figure 1. The missions, set along a mountain road.

The game is organized into three different intervention mechanisms: (1) psychoeducation (general information about OCD, OCD heterogeneity, OCD dimensionality, OCD cognitive model, evidence-based treatments, and options for seeking help in OCD), which includes 5 missions, and 1 of them—mission 3—including a video explaining the OCD cognitive model (1 min); (2) indirect contact (including 6 videos of approximately 2 to 3 minutes of 3 patients diagnosed with OCD who talk about their own experience with OCD: symptom description, interference, how long it took them to seek help, their experience

Intervention Program

The esTOCma beta version is a serious game whose content and videos were developed by Doctor of Philosophy–level clinical psychologists who are experts on OCD, together with a usability expert. A professional designer developed the graphic elements, and a computer engineer developed the app. After testing multiple prototypes, this version was developed (Multimedia Appendix 1).

During the game, participants are asked to fight against the OCD "stigma monster" with their knowledge by accomplishing 10 missions and freeing the 10 characters who are affected by the esTOCma monster, a creature that feeds on false beliefs and misinformation in society [35] (Figure 1). The participants are guided through the game by a woman who describes herself as an OCD expert.
in disclosing that they had a disorder, and experience with psychotherapy), which is organized into 2 missions; and (3) cognitive restructuring to replace dysfunctional beliefs related to rejection toward people with OCD, variables involved in OCD development, and treatment options and effectiveness, which includes 3 missions.

The missions are organized as follows: (1) the expert describes the objective of the mission and introduces the character to be freed, (2) the expert presents activities associated with the mission (between 6 and 8 activities, mostly consisting of reading a text and answering questions about it), and (3) the freed character appears. During the game, users receive basic internet-based rewards (ie, a key to free the character), together with a message of reinforcement and a video of the character actually being freed from eSTOCma. At the end of the game, there is new visual reinforcement through a video in which all the characters are freed from the monster. Moreover, users receive reinforcement through a diploma that certifies them as OCD experts. A further description of the game can be found in the study by Chaves et al [35].

**Measures**

The participants completed the sociodemographic information and Obsessive-Compulsive Inventory–Revised at preassessment and an acceptability questionnaire at postassessment. All other measures were completed at the preassessment, postassessment, and 3-month follow-up assessment.

**Sociodemographic Variables**

The sociodemographic variables include gender, age, educational level, level of information and communications technology (ICT) knowledge (from 1 little or none to 5 expert level), and the question of whether they have an OCD diagnosis.

**Acceptability Questionnaire**

It consisted of 3 questions developed ad hoc to assess the acceptability of the eSTOCma app. The survey included 2 questions related to usefulness (ie, “Did you find the app useful?” [from very useful to not useful at all] and “Did you learn from the app?” [from I learned a lot to I learned nothing]) and 1 question related to satisfaction with the app (ie, “Would you recommend this app to a friend?” [from a lot to not at all]). The questions were multiple-choice questions with 4 alternatives.

**AQ-27-E Measures**

This measures the social stigma associated with a vignette describing a person showing OC symptoms through 27 items rated on a Likert-type scale ranging from 1 to 9 [36,37]. It includes 9 subscales with 3 items: responsibility, pity, anger, dangerousness, fear, no help, coercion, segregation, and avoidance. Higher scores indicate higher social stigma. In this study, the AQ-27-E scales showed acceptable to excellent internal consistency across subscales (from 0.70 [anger, 3-month follow-up] to 0.91 [fear, 3-month follow-up]), except for responsibility (ranging from 0.40 [postassessment] to 0.63 [preassessment]) and pity (ranging from 0.36 [preassessment] to 0.52 [3-month follow-up]).

**GHSQ Measures**

This measures the intention to seek help from 10 different sources with regard to the specific content described in a vignette describing a person showing OC symptoms [38,39]. Participants rated 10 items regarding their help-seeking intentions if they were experiencing from symptoms similar to those described in the vignette on a 7-point Likert-type scale ranging from 1 (extremely unlikely) to 7 (extremely likely). Scores were calculated by summing up the items and dividing by 10. Higher scores indicate a higher intention to seek help. In this, the GHSQ showed acceptable internal consistency at the different assessment points (Cronbach α ranging from 0.68 [preassessment] to 0.75 [postassessment]).

**SDS Measures**

This assesses an individual’s willingness to interact with an individual with a mental disorder described in a vignette across 7 different situations on a 4-point Likert scale from 0 (definitely willing) to 3 (definitely unwilling) [40]. The total score has been calculated by adding the scores and dividing by the 7 items, with higher scores indicating a greater preference for social distance. In this, the SDS showed between good and excellent internal consistency at all assessment points; Cronbach α ranged from 0.86 at postassessment to 0.93 at the 3-month follow-up.

**MHL Questionnaire**

This is an instrument developed for this study based on previous studies [25,41] and assesses 1 of the components defined by Kutcher et al [42] as MHL but only referring to OCD: the understanding of OCD and its treatment. Part 1 has 4 multiple-choice questions including between 2 and 7 alternative answers, only 1 of which is correct. The questions refer to the assigned vignette (person A) and evaluate the following dimensions: (1) problem recognition (ie, “What happens to A is cause for concern?”; response alternatives: yes and no), (2) OCD identification (ie, “What do you think might be happening to A?”) This question includes 7 response alternatives: family problems, adjustment problems, anxiety disorder, generalized anxiety disorder, schizophrenia, OCD, and depression), (3) perception of causality, and (4) effective treatment option. Part 2 has four multiple-choice questions with 3 alternative answers, only 1 of which is correct, and refers to participants’ general knowledge of OCD: (1) identification of OCD as a mental disorder (ie, OCD is [a] a learning disorder, [b] a mental disorder, or [c] a set of manias); (2) definition of obsession; (3) definition of a compulsion; and (4) role played by compulsions and other control strategies in the maintenance of obsessions. A total score has been calculated as the sum of the correct answers, thus ranging from 0 to 8.

**Obsessive-Compulsive Inventory–Revised Measures**

This is an 18-item self-report questionnaire assessing distress caused by OC symptoms and rated on a 5-point Likert scale ranging from 0 (not at all) to 4 (extremely) [43,44]. A total score was calculated. The total score of the Obsessive-Compulsive Inventory-Revised showed excellent internal consistency at all assessment points in this study (Cronbach α ranging from 0.89 [postassessment] to 0.90 [preassessment]).
Statistical Analysis

Descriptive statistics (eg, means, SD, and frequencies or percentages) were used to analyze demographic data, study variables, and app use patterns. Chi-square and 1-tailed $t$ tests were used to test whether the groups (completers vs noncompleters) had preexisting differences. The change in study variables over time was determined by a repeated measures ANOVA. Partial eta–squared was used to report the effect size of the intervention on the dependent measures. A mixed model was used to determine whether the pattern of use of the app affected its effectiveness. The within-participants factor was time, and the between-participants factor was whether the user followed the 1 mission per day recommendation. The statistical significance level was set at $P=.05$. SPSS Statistics (version 26; IBM Corp) was used for statistical analysis.

Ethical Considerations

All procedures described in the study have been approved by the Human Research Ethics Committee of the University of Valencia, Spain (1276901). All study participants provided informed consent before study enrollment.

Results

Feasibility

Recruitment

A total of 90 participants were enrolled in the study; they downloaded the app and met the inclusion criteria. Of these 90 participants, 86% (78/90) were allocated to intervention and 13% (12/90) were excluded as they did not provide informed consent or did not complete the preassessment intervention (Figure 2). Of these 78 participants, 20% (16/78) did not complete the game or the postassessment (noncompleter group) and 79% (62/78) completed the game and answered the postintervention assessment (completer group). In the completer group, 69% (43/62) of the participants completed the 3-month follow-up assessment, and the remaining 30% (19/43) of the participants were lost at the 3-month follow-up.

**Figure 2.** Participants’ flowchart.
Participants’ Characteristics
The completer group had a mean age of 36.74 years, ranging from 18 to 71 years; were mostly women; and had university studies and moderate ICT knowledge. Differences in sociodemographic characteristics and preintervention assessment (baseline) between the completer group and the noncompleter group were calculated (Table 1).

Table 1. Differences between the completers and noncompleters on demographic and study variables (N=90).

<table>
<thead>
<tr>
<th>Variable or measure</th>
<th>Completers (n=62)</th>
<th>Noncompleters (n=16)</th>
<th>Chi-square (df)</th>
<th>t test (df)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (y), mean (SD)</td>
<td>36.74 (14.41)</td>
<td>27.19 (9.26)</td>
<td>10.4 (2)</td>
<td>N/A*</td>
<td>.003</td>
</tr>
<tr>
<td>Gender (women), n (%)</td>
<td>38 (61)</td>
<td>7 (44)</td>
<td>2.0 (2)</td>
<td>N/A</td>
<td>.35</td>
</tr>
<tr>
<td>Education level, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary</td>
<td>3 (5)</td>
<td>0 (0)</td>
<td>2.2 (2)</td>
<td>N/A</td>
<td>.31</td>
</tr>
<tr>
<td>Secondary</td>
<td>11 (18)</td>
<td>1 (6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University</td>
<td>48 (77)</td>
<td>15 (94)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge of ICT(^b), n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.58</td>
</tr>
<tr>
<td>Little or none</td>
<td>1 (2)</td>
<td>0 (0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>5 (8)</td>
<td>0 (0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>35 (57)</td>
<td>10 (63)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advanced</td>
<td>18 (29)</td>
<td>4 (25)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expert level</td>
<td>3 (5)</td>
<td>2 (13)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MHL(^c) (total(^d), mean (SD)</td>
<td>6.79 (1.50)</td>
<td>5.93 (1.61)</td>
<td>N/A</td>
<td>1.992 (76, 36.104)</td>
<td>.05</td>
</tr>
<tr>
<td>AQ-27-E(^e) (total(^f), mean (SD)</td>
<td>84.77 (24.81)</td>
<td>77.25 (25.52)</td>
<td>N/A</td>
<td>1.075 (76, 36.104)</td>
<td>.28</td>
</tr>
<tr>
<td>SDS(^g) (total(^h), mean (SD)</td>
<td>1.00 (0.64)</td>
<td>0.78 (0.70)</td>
<td>N/A</td>
<td>1.207 (76, 36.104)</td>
<td>.23</td>
</tr>
<tr>
<td>GHSQ(^i) (total(^j), mean (SD)</td>
<td>4.20 (0.92)</td>
<td>4.27 (0.83)</td>
<td>N/A</td>
<td>0.308 (76, 36.104)</td>
<td>.38</td>
</tr>
<tr>
<td>OCI-R(^k) (total(^l), mean (SD)</td>
<td>18.17 (13.23)</td>
<td>20.06 (10.70)</td>
<td>N/A</td>
<td>0.526 (76, 36.104)</td>
<td>.60</td>
</tr>
</tbody>
</table>

\(^a\text{N/A: not applicable.}\)
\(^b\text{ICT: information and communications technology.}\)
\(^c\text{MHL: Mental Health Literacy Questionnaire.}\)
\(^d\text{Total score ranging from 0 to 8.}\)
\(^e\text{AQ-27-E: Spanish Mental Illness Stigma Attribution Questionnaire–27.}\)
\(^f\text{Total scoring from 27 to 243.}\)
\(^g\text{SDS: Social Distance Scale.}\)
\(^h\text{Total score ranging from 0 to 3.}\)
\(^i\text{GHSQ: General Help-Seeking Questionnaire.}\)
\(^j\text{Total scoring from 1 to 7.}\)
\(^k\text{OCI-R: Obsessive-Compulsive Inventory–Revised.}\)
\(^l\text{Total scoring from 0 to 72.}\)

Statistically significant differences were only observed in age, with the completer group being older, and MHL scores, which were higher in the completer group. In addition, in the completer group, we explored the differences between the participants who completed the 3-month assessment (43/62, 69%) and those who did not (19/62, 31%). The only discernible difference observed was in age ($t_{60}=2.167; P=.01$), with participants who completed the follow-up assessment being older.

Of the 43 participants included in the effectiveness analyses, most were women (n=28, 65%) with a mean age of 39.30 (SD 14.58; range 21-71) years, with university-level education (n=35, 81%), and with ICT knowledge between moderate (n=23, 54%) and advanced (n=15, 35%; only 3 (7%) participants described having between little or none and low ICT knowledge level).

App Use Pattern
An analysis of the app use pattern was conducted with the completer group. First, we explored the number of days that elapsed from the preassessment to the postassessment. The participants took a mean of 13.64 (SD 10.50; range 1-44) days to complete the app (which includes having done the preassessment, the 10 missions and the postassessment), with a mode of 1, that is, the most frequent pattern was conducting the pre- and postassessments and missions of the app in 1 day. Second, we analyzed the number of days the person spent performing the missions. The participants spent between 1 and
10 days performing the missions, with a mean of 4.56 (SD 3.33) days and a mode of 1. In total, 30% (19/62) of the participants completed the 10 missions in 1 day, whereas another 21% (13/62) of the participants spent between 9 and 10 days completing the app, that is, approximately 1 session per day.

Finally, we explored the minutes spent performing each mission, first excluding the participants who stayed on 1 mission for >20 minutes, as we assumed that they left the app open without using it. The data on the participants completing the missions in <20 minutes is presented in Table 2.

Table 2. Time (in min) spent in completing each of the 10 missions for participants who complete the mission in <20 minutes.

<table>
<thead>
<tr>
<th>Mission</th>
<th>Participants, n (%)(^a)</th>
<th>Time (min), mean (SD; range)</th>
<th>Mode (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mission 1</td>
<td>58 (94)</td>
<td>3.93 (1.02; 1-6)</td>
<td>4</td>
</tr>
<tr>
<td>Mission 2</td>
<td>54 (87)</td>
<td>4.44 (1.90; 2-10)</td>
<td>4</td>
</tr>
<tr>
<td>Mission 3</td>
<td>35 (56)</td>
<td>3.80 (1.37; 1-8)</td>
<td>3</td>
</tr>
<tr>
<td>Mission 4</td>
<td>35 (56)</td>
<td>2.06 (1.64; 0-10)</td>
<td>2</td>
</tr>
<tr>
<td>Mission 5</td>
<td>35 (56)</td>
<td>5 (3.11; 2-20)</td>
<td>4</td>
</tr>
<tr>
<td>Mission 6</td>
<td>56 (96)</td>
<td>8.71 (1.78; 7-20)</td>
<td>8</td>
</tr>
<tr>
<td>Mission 7</td>
<td>29 (47)</td>
<td>9.35 (3.06; 7-20)</td>
<td>8</td>
</tr>
<tr>
<td>Mission 8</td>
<td>33 (53)</td>
<td>3.85 (1.48; 2-8)</td>
<td>3</td>
</tr>
<tr>
<td>Mission 9</td>
<td>58 (94)</td>
<td>2 (3.01; 0-19)</td>
<td>0</td>
</tr>
<tr>
<td>Mission 10</td>
<td>56 (90)</td>
<td>2.09 (2.26; 0-9)</td>
<td>0</td>
</tr>
</tbody>
</table>

\(^a\)n (%) of participants completing the mission in ≤20 minutes.

The data are displayed in minutes.

Between the participants who completed missions in ≤20 minutes, and considering the mode, missions included in module 1 (ie, missions 1 to 5) took between 2 and 4 minutes, in module 2 (ie, missions 6 to 7) took 8 minutes, and in module 3 (ie, missions 8 to 10) took between <1 minute and 3 minutes.

Acceptability

After completing the app until the end of the game, most participants (56/62, 90%) perceived the app as useful or very useful. Moreover, 90% (56/62) of the participants indicated that they had learned or learned a lot, and 98% (61/62) of the participants indicated that they would recommend the app to a friend.

Preliminary Effectiveness: Differences Between Pre- and Postintervention Assessments and 3-Month Follow-Up

Repeated measures ANOVA was conducted to examine differences between pre- and postintervention and the 3-month follow-up on knowledge of OCD and its treatments (MHL questionnaire), stigmatizing attitudes (AQ-27-E), social distance (SDS), and intention to seek help (GHSQ) associated with OCD (Table 3). The results showed statistically significant differences (\(P \leq .05\)) in all variables with medium-to-large effect sizes, except for the second part of the MHL questionnaire and the pity and coercion subscales (AQ-27-E). In general, the results show that using esTOCma until the end of the game results in an increase in MHL and intention to seek treatment (GHSQ) and a decrease in stigmatizing attitudes (AQ-27-E) and social distance desire (SDS). Post hoc pairwise comparisons showed statistically significant differences between preintervention and the other 2 assessment points (postintervention and 3-month follow-up). No significant differences were observed between postintervention and the 3-month follow-up in the variables assessed, except for the MHL total score, in which follow-up scores did not differ from pre- to postintervention.
Table 3. Means (SDs) and repeated measures ANOVA on pre-, post-, and 3-month follow-up intervention scores (n=43).

<table>
<thead>
<tr>
<th>Variable or measure</th>
<th>Pretreatment, mean (SD)</th>
<th>Posttreatment, mean (SD)</th>
<th>3 month follow-up, mean (SD)</th>
<th>F test (df)a</th>
<th>P value</th>
<th>( \eta^2_b )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHLc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part 1d</td>
<td>3.32 (0.80)e</td>
<td>3.72 (0.54)e</td>
<td>3.62 (0.61)e</td>
<td>6.523 (1.566, 65.769)</td>
<td>.005</td>
<td>0.134</td>
</tr>
<tr>
<td>Part 2f</td>
<td>3.58 (0.69)</td>
<td>3.72 (0.50)</td>
<td>3.77 (0.57)</td>
<td>1.896 (1.513, 63.526)</td>
<td>.16</td>
<td>0.043</td>
</tr>
<tr>
<td>Total score</td>
<td>6.90 (1.34)e</td>
<td>7.44 (0.93)e</td>
<td>7.39 (0.90)e</td>
<td>5.754 (1.325, 55.662)</td>
<td>.01</td>
<td>0.120</td>
</tr>
<tr>
<td>AQ-27-Eg</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Responsibility</td>
<td>9.04 (4.07)e</td>
<td>6.97 (3.70)e</td>
<td>7.30 (4.35)e</td>
<td>8.103 (2, 84)</td>
<td>.001</td>
<td>0.162</td>
</tr>
<tr>
<td>Pity</td>
<td>17.30 (4.15)</td>
<td>18.62 (4.36)</td>
<td>18.20 (4.68)</td>
<td>1.984 (1.84, 76.81)</td>
<td>.14</td>
<td>0.045</td>
</tr>
<tr>
<td>Anger</td>
<td>8.23 (4.44)e</td>
<td>5.81 (3.42)e</td>
<td>5.76 (3.19)e</td>
<td>10.554 (1.614, 76.813)</td>
<td>&lt;.001</td>
<td>0.201</td>
</tr>
<tr>
<td>Dangerousness</td>
<td>7.11 (4.31)e</td>
<td>4.72 (2.65)e</td>
<td>5.37 (4.01)e</td>
<td>7.386 (2, 84)</td>
<td>&lt;.001</td>
<td>0.150</td>
</tr>
<tr>
<td>Fear</td>
<td>6.39 (4.26)e</td>
<td>5.00 (3.72)e</td>
<td>4.23 (2.42)e</td>
<td>6.489 (2, 84)</td>
<td>.002</td>
<td>0.134</td>
</tr>
<tr>
<td>No help</td>
<td>8.09 (4.68)e</td>
<td>5.69 (3.32)e</td>
<td>5.83 (4.05)e</td>
<td>9.320 (2, 84)</td>
<td>&lt;.001</td>
<td>0.182</td>
</tr>
<tr>
<td>Coercion</td>
<td>12.00 (5.30)</td>
<td>11.18 (5.50)</td>
<td>10.93 (6.47)</td>
<td>0.983 (1.490, 62.601)</td>
<td>.35</td>
<td>0.023</td>
</tr>
<tr>
<td>Segregation</td>
<td>5.16 (3.92)e</td>
<td>4.00 (1.96)e</td>
<td>4.04 (2.22)e</td>
<td>4.762 (1.555, 65.297)</td>
<td>.01</td>
<td>0.102</td>
</tr>
<tr>
<td>Avoidance</td>
<td>11.88 (5.66)e</td>
<td>8.58 (6.35)e</td>
<td>7.76 (5.48)e</td>
<td>16.938 (2, 84)</td>
<td>&lt;.001</td>
<td>0.287</td>
</tr>
<tr>
<td>SDSb</td>
<td>0.99 (0.58)</td>
<td>0.64 (0.65)</td>
<td>0.59 (0.63)</td>
<td>10.597 (2, 84)</td>
<td>&lt;.001</td>
<td>0.201</td>
</tr>
<tr>
<td>GHSQi</td>
<td>4.33 (0.84)e</td>
<td>4.84 (0.86)e</td>
<td>4.58 (1.03)e</td>
<td>6.818 (2, 84)</td>
<td>.002</td>
<td>0.140</td>
</tr>
</tbody>
</table>

aData were Greenhouse-Geisser corrected where appropriate.
\( \eta^2_b \): partial eta squared for within-subject contrasts (ANOVA).
*MHL: Mental Health Literacy Questionnaire, total score ranging from 0 to 8.

Part 1 scoring from 0 to 4.

Significant differences among groups (P ≤ .05).
Part 2 scoring from 0 to 4.

AQ-27-E: Spanish Mental Illness Stigma Attribution Questionnaire–27; subscales ranging from 3 to 27.

SDS: Social Distance Scale, ranging from 0 to 3.

GHSQ: General Help-Seeking Questionnaire, scoring from 1 to 7.

Finally, we examined whether the pattern of use of the app influenced the effectiveness of the intervention with a 2 (group: individuals who follow the recommendation of 1 mission per day [ie, 9/10, % days] [12/N, %; percentile 75] vs individuals who did it in another way [31/N, %]) × 3 (time: pre-, post-, and follow-up assessments) repeated measures mixed ANOVA. The results show that there was no significant group×time interaction (\( F_{26,16} = 0.838; \ P = .67 \)). Univariate follow-up analyses also indicated no significant group×time effect for any of the measured variables (\( P > .05 \)).

Discussion

Principal Findings

This study is the first to investigate the feasibility, acceptability, and effectiveness of a mobile health app designed to enhance several variables related to OCD, which were identified by Kutcher et al [42] as components of MHL: understanding OCD and its treatments, decreasing stigma associated with OCD, and enhancing help-seeking effectiveness. Our findings show that esTOCma was feasible and acceptable and that after using it until game completion, there was a positive change in the variables of interest, which lasted for at least 3 months.

Of the participants allocated to the intervention, approximately 80% (78/90) completed the app, and of the who completed the intervention, approximately 70% (62/78) completed the follow-up assessment. This adherence rate is similar to or higher than that reported in other internet-based studies with self-help interventions, with dropouts being one of the main challenges of interventions with mental health apps [45-47].
moderate ICT knowledge. Regarding the pattern of use, although it was recommended that participants complete 1 mission per day and participants were reminded of this recommendation after finishing each mission, the most common pattern of use was to complete the app, including pre- and postevaluations, in 1 day. In fact, only 21% (13/62) of the participants completing the intervention followed the recommendation to perform 1 mission per day. It seems that it is more comfortable for participants to perform more missions per day; otherwise, they forget to complete the mission the following day. In fact, the participants completed their participation in the study within a range of 1 to 44 days.

Regarding the time invested in each mission, although missions were quick to complete and always took <10 minutes, the participants often left missions midway and continued later the same day or days later. Furthermore, certain missions were abandoned in the middle more frequently, suggesting that they could be reformulated to make them more “attractive.” This was the case for mission 7, which was interrupted by >50% (31/62) of the participants, as well as for missions 3, 4, 5, and 8. The pattern of use and interruptions do not seem to be associated with the intervention mechanism on which the mission is based but perhaps with the content or the duration of the mission. If we analyze missions 1, 2, 6, 9, and 10, those that were carried out without interruption, we see that missions 1 and 2 are the initial ones and deal with content describing obsessions and compulsions; mission 6 includes the first videos that also describe symptomatology and interference, without additional theoretical content to read; and missions 9 and 10 (cognitive restructuring) are characterized by being very brief and with less theoretical content than missions 1 to 5 (psychoeducation). It seems that the description of symptomatology, as well as more dynamic and shorter missions, result in more attractive missions or at least in missions that capture participants’ attention to a greater extent.

In general, the participants seemed satisfied with the app, as approximately 100% (62/62) would recommend it to a friend and >90% (56/62) consider it useful and that they have learned about OCD.

Regarding the effectiveness of the app, the results show an increase in OCD knowledge. Our data match those of the previous interventions that have found increases in general knowledge of OCD after offering written information about the disorder [25,26]. The data are also consistent with other technology interventions that have increased MHL levels on different mental health problems [48-50]. However, the differences were not statistically significant in those questions that asked about OCD in general (eg, the definition of an obsession), that is, not referring to the description of a person showing OC symptoms. This could be due to a ceiling effect, as the scores were already high in the preassessment evaluation. In fact, they were higher among those participants who played with the app until the game was over in comparison with those who did not finish the game. In this sense, our first hypothesis was only partially supported.

Regarding the hypothesized decrease in stigmatizing attitudes, the results support our hypothesis, as they suggest that after completing the 10 missions of the app, there was a decrease in some stereotypes or public attitudes, such as the perception of OC symptoms as dangerous; emotional reactions of anger or fear toward people showing OC symptoms; discriminating behaviors such as the intention of not helping, segregating, or avoiding people showing OC symptoms; or the desire to maintain social distance. Although a video-based intervention decreased social distance desire [29] and 2 interventions centered on reading the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition diagnostic criteria for OCD decreased negative attitudes about violent and sexual thoughts [28,51], other interventions consisting of reading information on OCD (psychoeducation) showed small changes in stigmatizing attitudes [25]. The results suggest that the esTOCma intervention, which includes psychoeducation but also incorporates components such as contact, seems to change stigmatizing attitudes to a greater extent, with medium-to-large effect sizes.

Our data are also consistent with previous research also using new technologies to reduce stigma associated with other mental disorders that have reported a decrease in dangerousness, anger, fear, segregation, and avoidance [15,52,53]; a decrease in stigma as a general measure [53-55]; an increase in the help factor [52]; or a decrease in social distance [52,56]. This is a remarkable result, as traditional antistigma interventions not using innovative technologies often report small-to-medium effect sizes [14,52]. Moreover, the results are based on a community sample with an average age higher than those used in studies reporting interventions using innovative technologies [16], suggesting that these types of interventions could also be useful for older people.

In addition to the changes in stigmatizing attitudes after using esTOCma, there were no significant changes in 2 of the social stigma dimensions measured by the AQ-27-E: pity and coercion. Previous studies have also reported a lack of changes in pity [57]. In fact, it has been suggested that pity should be considered, at least in some contexts, not to be a factor of stigma but rather a reflection of compassion and the capacity to empathize with people with mental health problems [58], and research has shown associations between pity and the tendency to help [37]. Regarding coercion, there is no significant change after the intervention in the belief that people with OCD should receive treatment, even if they refuse it (coercion). This is consistent with previous interventions in the OCD field [25].

Finally, the results also support our third hypothesis, as there was a large effect size improvement in the intention to seek help when experiencing symptoms similar to those represented in the different vignettes. Thus, increasing knowledge of OCD and its treatment could have acted as a help-seeking facilitator [59,60]. To the best of our knowledge, there are no interventions to improve help seeking associated with OCD, and thus, our results are of great relevance, as research shows that early help seeking is associated with a better treatment response, earlier remission of symptoms, and recovery from the disorder [9,61]. The data are consistent with other studies that, through mental health apps [62] and other technology-based interventions, have improved the intention to seek help for other mental health problems [63].
Moreover, our study showed a maintained effect in the study variables at the 3-month follow-up assessment. This is especially relevant as most of the interventions do not include a follow-up assessment [16,50,63,64], and only 1 intervention on OCD has included a follow-up assessment [29].

Limitations and Recommendations

This study has limitations. As a pilot study, the study sample was small and did not include a control group. There was also a considerable dropout rate that could be associated with the fact that participants forgot to complete the app (to do all 10 missions) as well as the large number of assessment questionnaires included. Furthermore, considering that OCD is a heterogeneous disorder and that there are differences in stigma and OCD recognition between different types of content [41,64,65], we decided to randomize participants to 6 vignettes that represent OCD heterogeneity. However, this decision could be considered as a limitation of the design of this study that could affect the effectiveness of the data.

Despite these limitations, our results provide valuable information for a broader implementation of the app. First, as future lines of work, we propose making an effort to motivate participants, given that higher adherence was associated with initial knowledge of OCD or sensitivity toward mental health problems. Second, a potential strategy to attract a younger audience, as adherence was also associated with older age, could involve transforming the app into a serious game, aligning esTOCma with the immersive experience of video games. Actively engaging participants through gamification elements could further contribute to increasing adherence. Looking ahead, evolving esTOCma into a serious game with an adventure game theme holds promise for future developments and for addressing the stigma associated with OCD. Third, pop-ups and emails to remind participants to “play” could be incorporated, as those participants who completed the participation took up to 44 days to finish an app recommended to be conducted in 10 days. Fourth, given that the number of missions per day seems not to influence the effectiveness of the app, the recommendation to play 1 mission per day could be omitted, as this is not the preferred pattern of use of participants. Finally, given that some missions were completed with interruptions by approximately half of the participants, it would be recommended to inform users of how long each mission will last. Furthermore, missions 3 to 5, 7, and 8 should be revised, as although they were quick to complete (between 3 and 8 min), the participants often left them midway and continued later. They could perhaps be shortened or made more dynamic.

Future Research

The results suggest the potential for the use of this intervention app and provide the basis for developing a larger randomized controlled study to validate the use of esTOCma (version 1.0), as has been proposed [35]. Importantly, we will also explore differences in OCD types of content through a controlled study, as research shows that there are differences in OCD recognition and stigma considering the different OCD types of content [41,64,65]. Furthermore, it will be of interest to explore the effectiveness of each intervention mechanism (ie, psychoeducation, indirect contact, and cognitive restructuring) to understand whether there are differences between them. This analysis would allow us to develop a new app that would eliminate those mechanisms that are less effective or improve them. Future studies should also examine the usefulness of this app in OCD cohorts, especially in the first stages of diagnosis, and their families. Through esTOCma, people with OCD could receive (or reinforce) cognitive-behavioral psychoeducation and be assisted in their search for empirically-based treatments. Furthermore, esTOCma could eliminate self-stigma, improving their quality of life. In terms of family involvement, esTOCma could assist them in comprehending and managing OCD symptoms as well as support their relatives with OCD in seeking effective treatment.

Conclusions

Our findings show that esTOCma is a feasible and acceptable app and that after completing its 10 missions, there is an increase in the understanding of OCD and help-seeking intention as well as a decrease in the social stigma and social distance associated with OCD that lasts for at least 3 months. These changes might result in less delay in seeking help and a better treatment response to the problem and prognosis. Moreover, providing mental health knowledge to the community population as to the nature and universality of intrusive thoughts may protect and prevent the general population from developing OCD and, furthermore, reduce the economic and personal costs associated with OCD.

Acknowledgments

Funding for this study (RTI2018-098349-B-I00 and PID2021-124409OB-I00) was provided by the Ministry of Science and Innovation—State Research Agency of Spain and by the European Union European Regional Development Fund: “A way of making Europe.” None of the funders had any role in the design, preparation, review, or approval of the manuscript or in the decision to submit the manuscript for publication.

Authors' Contributions

The work was conceptualized by AC and GG-S. The methodology and research, and the writing of the original draft were carried out by AC, GG-S and SA. Formal analyses were developed by SA. The final manuscript was reviewed and edited by GG-S.

Conflicts of Interest

None declared.
Multimedia Appendix 1
Screenshots of the esTOCma app with examples of the Psychoeducation mechanism.

[PDF File (Adobe PDF File), 960 KB - mhealth_v12i1e48027_app1.pdf ]

References


16. Rodríguez-Rivas ME, Cangas AJ, Varela JJ, Valdebenito S. Innovative technology-based interventions to reduce stigma toward people with mental illness: systematic review and meta-analysis. JMIR Serious Games 2022 May 30;10(2):e35099 [FREE Full text] [doi: 10.2196/35099] [Medline: 35635744]


39. Pacheco del Castillo LA. Dominican college students’ experiences of distress, help-seeking and stigma. Western Michigan University. 2017. URL: https://scholarworks.wmich.edu/cgi/viewcontent.cgi?article=4111&context=dissertations [accessed 2024-03-05]


Abbreviations

- AQ-27-E: Spanish Mental Illness Stigma Attribution Questionnaire–27
- GHSQ: General Help-Seeking Questionnaire
- ICT: information and communications technology
- MHL: mental health literacy
- OC: obsessive-compulsive
- OCD: obsessive-compulsive disorder
- SDS: Social Distance Scale

©Antonio Chaves, Sandra Arnáez, Gemma García-Soriano. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 29.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Evidence of How Physicians and Their Patients Adopt mHealth Apps in Germany: Exploratory Qualitative Study

Tanja Schroeder¹,², BA, MA; Maximilian Haug², BEng, MSc; Andrew Georgiou¹, BA, DipArts, MSc, PhD; Karla Seaman¹, BPharm, MSCE, PhD; Heiko Gewald², BSc, MBA, EMBSc, PhD

¹Centre for Health Systems and Safety Research, Australian Institute of Health Innovation, Macquarie University, Sydney, Australia
²Institute for Digital Innovation, Faculty of Information Management, Neu-Ulm University of Applied Sciences, Neu-Ulm, Germany

Corresponding Author:
Tanja Schroeder, BA, MA
Centre for Health Systems and Safety Research
Australian Institute of Health Innovation
Macquarie University
Level 6, 75 Talavera Road
Sydney, 2109
Australia
Phone: 61 2 9850 ext 6281
Email: tanja.schroeder@mq.edu.au

Abstract

Background: The enactment of the “Act to Improve Healthcare Provision through Digitalisation and Innovation” (Digital Healthcare Act; Digitale-Versorgung-Gesetz [DVG]) in Germany has introduced a paradigm shift in medical practice, allowing physicians to prescribe mobile health (mHealth) apps alongside traditional medications. This transformation imposes a dual responsibility on physicians to acquaint themselves with qualifying apps and align them with patient diagnoses, while requiring patients to adhere to the prescribed app use, similar to pharmaceutical adherence. This transition, particularly challenging for older generations who are less skilled with technology, underscores a significant evolution in Germany’s medical landscape.

Objective: This study aims to investigate physicians’ responses to this novel treatment option, their strategies for adapting to this form of prescription, and the willingness of patients to adhere to prescribed mHealth apps.

Methods: Using an exploratory qualitative study design, we conducted semistructured interviews with 28 physicians and 30 potential patients aged 50 years and older from August 2020 to June 2021.

Results: The findings reveal several factors influencing the adoption of mHealth apps, prompting a nuanced understanding of adoption research. Notably, both physicians and patients demonstrated a lack of information regarding mHealth apps and their positive health impacts, contributing to a deficiency in trust. Physicians’ self-perceived digital competence and their evaluation of patients’ digital proficiency emerge as pivotal factors influencing the prescription of mHealth apps.

Conclusions: Our study provides comprehensive insights into the prescription process and the fundamental factors shaping the adoption of mHealth apps in Germany. The identified information gaps on both the physicians’ and patients’ sides contribute to a trust deficit and hindered digital competence. This research advances the understanding of adoption dynamics regarding digital health technologies and highlights crucial considerations for the successful integration of digital health apps into medical practice.

JMIR Mhealth Uhealth 2024;12:e48345 doi:10.2196/48345

KEYWORDS
mobile health apps; DiGA; adoption; prescription; mHealth; aging and individual differences

Introduction

Mobile devices have enormous potential to enhance the way patients receive medical care and health education [1]. Mobile health (mHealth) is a dynamic and expanding area of health care with short innovation cycles [1,2]. mHealth is closely related to telemedicine and eHealth. The fundamental difference lies in how the related services are delivered to the patient. According to the World Health Organization (WHO), mHealth expands the spectrum of functionalities to include a mobile component. Thus, mHealth is delivered through any mobile
device, from simple mobile phones, smartphones, and tablets to wearable devices used in health care settings [2].

To accommodate the growing number of mHealth apps (ie, software products for smartphones designed to support good health), Germany passed the “Act to Improve Healthcare Provision through Digitalisation and Innovation” (Digital Healthcare Act; Digitale-Versorgung-Gesetz [DVG]) in December 2019 and the Digital Health Applications Ordinance (Digitale-Gesundheitsanwendungen-Verordnung [DiGAV]) in October 2020 [3,4]. This ordinance enables physicians to prescribe mHealth apps (in German referred to as Digitale Gesundheitsanwendungen [DiGAs; digital health applications]) to their patients in the same way as any other medicine. A DiGA supports the recognition, monitoring, treatment, and alleviation of diseases, injuries, or disabilities [5]. To qualify as a DiGA, an mHealth app has to successfully pass the evaluation of the Federal Institute for Drugs and Medical Devices (Bundesinstitut für Arzneimittel und Medizinprodukte; BfArM). Only then will an app be included in the list of reimbursable digital health apps, the so-called DiGA directory. In September 2023, this directory comprised 48 admitted DiGAs [6]. DiGAs cover a wide array of the medical spectra. Popular examples are “Endo-App” (to treat endometriosis), “Kaia COPD” (for chronic obstructive pulmonary diseases), or “sinCephalea” (for the treatment of migraine).

The introduction of the “Act to Improve Healthcare Provision through Digitalisation and Innovation” represents a significant and innovative change to the German health care system [7]. The new legal framework elevates DiGAs to the ranks of medical devices [5]. Therefore, if physicians want to prescribe a DiGA, they are required to thoroughly inform themselves about which apps can help their patients, how they are to be used, and how a positive influence is expected to manifest. Patients are required and need to be able to adhere to the prescription; for example, they have to use the DiGA as stated by the physician. Thus, the physician needs to assess (implicitly or explicitly) whether the patient is likely to use the DiGA as prescribed. This does not only include adherence to, for example, training intervals (for orthopedic DiGAs) but also the general technology savviness of the patient, for example, whether the patient is able to download the app, install it on the smartphone, maintain updates, and so on. This question is specifically challenging when it comes to older users, who are often regarded as being not technology savvy.

Consequently, this innovation raises several questions for technology adoption research. Typically, adoption research concerns the individual user’s decision whether to use technology, either mandatory [8,9] or voluntary [10]. Now, a concerned third party (the physician) decides on behalf of the user whether the patient is expected to be willing and able to use an app on their smartphone voluntarily. Therefore, the physician’s assessment now includes not only the medical relevance of the DiGA but also whether the user might be able and willing to use it as prescribed. Although the latter assessment seems easy for younger people, the case is much more difficult for older patients. Potential doctor misperception raises concerns about a possible digital divide and ageism by doctors. A recent report by the WHO and the United Nations (UN) raises awareness of this issue and urges action to combat ageism, as it negatively impacts well-being and can lead to premature death and higher health care costs [11].

In the health care context, the resulting research questions are specifically relevant, as the influence of age on the adoption of mHealth apps has not yet received sufficient attention in the scientific discourse [12]. Therefore, this research addresses the complementary research questions:

1. What factors enable or hinder physicians to prescribe DiGAs?
2. What factors enable or hinder older users’ adoption of DiGAs?

Methods

Study Design

We developed an exploratory qualitative study design to answer the research questions and gain insights through semistructured interviews with representatives of the 2 relevant stakeholder groups: physicians and patients. The study adhered to the standards for reporting on qualitative research [13].

Study Setting

The data for our research were gathered in Germany.

Participants

The first qualitative study was conducted from August 2020 to June 2021. We interviewed 28 physicians (demographics are provided in Table 1) to assess how DiGAs can improve the health of their patients. Physicians were chosen as research objects because of their unique role as prescribers of DiGAs. The interviewees had different backgrounds in terms of IT affinity and previous experience with mHealth apps in general.
Table 1. Demographics of the physicians.

<table>
<thead>
<tr>
<th>Code</th>
<th>Specialty</th>
<th>Sex</th>
<th>Age (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>General practitioner and cardiologist</td>
<td>Male</td>
<td>63</td>
</tr>
<tr>
<td>E2</td>
<td>Dermatologist</td>
<td>Male</td>
<td>67</td>
</tr>
<tr>
<td>E3</td>
<td>Dermatologist</td>
<td>Male</td>
<td>45</td>
</tr>
<tr>
<td>E4</td>
<td>Urologist</td>
<td>Male</td>
<td>48</td>
</tr>
<tr>
<td>E5</td>
<td>General practitioner</td>
<td>Female</td>
<td>63</td>
</tr>
<tr>
<td>E6</td>
<td>Dermatologist</td>
<td>Female</td>
<td>54</td>
</tr>
<tr>
<td>E7</td>
<td>Dermatologist</td>
<td>Male</td>
<td>41</td>
</tr>
<tr>
<td>E8</td>
<td>Internist and gerontologist</td>
<td>Male</td>
<td>38</td>
</tr>
<tr>
<td>E9</td>
<td>Orthopedist</td>
<td>Male</td>
<td>53</td>
</tr>
<tr>
<td>E10</td>
<td>General practitioner</td>
<td>Male</td>
<td>64</td>
</tr>
<tr>
<td>E11</td>
<td>Geriatric therapist</td>
<td>Female</td>
<td>58</td>
</tr>
<tr>
<td>E12</td>
<td>Urologist</td>
<td>Male</td>
<td>38</td>
</tr>
<tr>
<td>E13</td>
<td>General practitioner and pain therapist</td>
<td>Male</td>
<td>59</td>
</tr>
<tr>
<td>E14</td>
<td>Assistant doctor cardiology</td>
<td>Male</td>
<td>35</td>
</tr>
<tr>
<td>E15</td>
<td>Urologist</td>
<td>Male</td>
<td>42</td>
</tr>
<tr>
<td>E16</td>
<td>General practitioner</td>
<td>Male</td>
<td>68</td>
</tr>
<tr>
<td>E17</td>
<td>Dermatologist</td>
<td>Female</td>
<td>36</td>
</tr>
<tr>
<td>E18</td>
<td>General practitioner</td>
<td>Female</td>
<td>38</td>
</tr>
<tr>
<td>E19</td>
<td>General practitioner</td>
<td>Female</td>
<td>59</td>
</tr>
<tr>
<td>E20</td>
<td>General practitioner</td>
<td>Female</td>
<td>41</td>
</tr>
<tr>
<td>E21</td>
<td>General practitioner</td>
<td>Female</td>
<td>45</td>
</tr>
<tr>
<td>E22</td>
<td>General practitioner</td>
<td>Male</td>
<td>45</td>
</tr>
<tr>
<td>E23</td>
<td>Neurologist</td>
<td>Male</td>
<td>44</td>
</tr>
<tr>
<td>E24</td>
<td>General practitioner</td>
<td>Male</td>
<td>67</td>
</tr>
<tr>
<td>E25</td>
<td>Neurologist</td>
<td>Male</td>
<td>65</td>
</tr>
<tr>
<td>E26</td>
<td>Neurologist</td>
<td>Male</td>
<td>42</td>
</tr>
<tr>
<td>E27</td>
<td>Molecular neurologist</td>
<td>Female</td>
<td>45</td>
</tr>
<tr>
<td>E28</td>
<td>General practitioner</td>
<td>Female</td>
<td>37</td>
</tr>
</tbody>
</table>

In the second study, we interviewed patients (ie, the potential users of a DiGA). To reflect the specific issues of aging patients, we chose interview partners over 50 years of age. Evidence shows that from this age onward, chronic diseases increase significantly [14]. Therefore, this age group is likely to represent a large part of the target group for the prescription of DiGAs. Furthermore, studies suggest that there are still age disparities in attitudes toward technology and that the aging population is often less comfortable using technology [15].

We conducted 30 interviews to determine the factors that influence potential patients to adopt DiGAs or not. Demographics of the interviewees are given in Table 2.
Table 2. Demographics of the patients.

<table>
<thead>
<tr>
<th>Code</th>
<th>Sex</th>
<th>Age (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Male</td>
<td>68</td>
</tr>
<tr>
<td>P2</td>
<td>Female</td>
<td>60</td>
</tr>
<tr>
<td>P3</td>
<td>Female</td>
<td>57</td>
</tr>
<tr>
<td>P4</td>
<td>Female</td>
<td>76</td>
</tr>
<tr>
<td>P5</td>
<td>Female</td>
<td>56</td>
</tr>
<tr>
<td>P6</td>
<td>Female</td>
<td>65</td>
</tr>
<tr>
<td>P7</td>
<td>Male</td>
<td>69</td>
</tr>
<tr>
<td>P8</td>
<td>Female</td>
<td>64</td>
</tr>
<tr>
<td>P9</td>
<td>Male</td>
<td>68</td>
</tr>
<tr>
<td>P10</td>
<td>Female</td>
<td>66</td>
</tr>
<tr>
<td>P11</td>
<td>Male</td>
<td>65</td>
</tr>
<tr>
<td>P12</td>
<td>Male</td>
<td>67</td>
</tr>
<tr>
<td>P13</td>
<td>Female</td>
<td>57</td>
</tr>
<tr>
<td>P14</td>
<td>Female</td>
<td>72</td>
</tr>
<tr>
<td>P15</td>
<td>Female</td>
<td>67</td>
</tr>
<tr>
<td>P16</td>
<td>Male</td>
<td>53</td>
</tr>
<tr>
<td>P17</td>
<td>Female</td>
<td>61</td>
</tr>
<tr>
<td>P18</td>
<td>Female</td>
<td>69</td>
</tr>
<tr>
<td>P19</td>
<td>Female</td>
<td>61</td>
</tr>
<tr>
<td>P20</td>
<td>Male</td>
<td>63</td>
</tr>
<tr>
<td>P21</td>
<td>Male</td>
<td>65</td>
</tr>
<tr>
<td>P22</td>
<td>Female</td>
<td>68</td>
</tr>
<tr>
<td>P23</td>
<td>Female</td>
<td>67</td>
</tr>
<tr>
<td>P24</td>
<td>Female</td>
<td>61</td>
</tr>
<tr>
<td>P25</td>
<td>Female</td>
<td>67</td>
</tr>
<tr>
<td>P26</td>
<td>Female</td>
<td>59</td>
</tr>
<tr>
<td>P27</td>
<td>Male</td>
<td>64</td>
</tr>
<tr>
<td>P28</td>
<td>Female</td>
<td>69</td>
</tr>
<tr>
<td>P29</td>
<td>Female</td>
<td>64</td>
</tr>
<tr>
<td>P30</td>
<td>Female</td>
<td>68</td>
</tr>
</tbody>
</table>

Recruitment
We recruited physicians by telephone from a community-based physician network in Germany and via social media. For the interviews with patients, we promoted our study to doctors and approached medical centers and clinics. To identify suitable interview partners (ie, potential patients), we presented the research project to local sports, communication, and civic groups; promoted the study on various social media platforms; and spoke to the local press. In this way, we motivated suitable candidates to contact us. The participation of all study participants, both physicians and patients, was voluntary.

Data Collection
All interviews were semistructured and led by a list of questions and general topics that the interviewers were supposed to address. The semistructured interview guidelines are provided in Multimedia Appendix 1. As a structure and topic basis for creating the questions, the constructs from well-known technology adoption research models (unified theory of acceptance and use of technology [UTAUT] [9]) and models for the analysis of health behavior (health belief model [HBM] [16]) were considered. The questions were primarily open and allowed the interviewees to explore their experiences and views. Supported by a systematic and comprehensive interview process, the interviewers had high degrees of freedom to conduct the interview in order to gain deeper insights. The questions were adjusted correspondingly for the following interviews to gain...
more profound knowledge for each interview. Interviews lasted around 20-45 minutes and were conducted face-to-face or over the phone by 1 researcher (TS). The interviews were conducted in German and were recorded, transcribed, and translated into English for further analysis [17].

**Data Analysis**

For the coding process, we used the NVivo10 software (QSR International). The research was conducted using an interpretative phenomenological analysis (IPA) [18]. IPA aims to examine the world perspective of the participants and, if possible, take an “insider perspective” [18] of the phenomenon under investigation. At the same time, IPA recognizes that research is a dynamic process. In parallel with the data collection, we scanned and coded the data from the first round of analysis. One researcher (TS) conducted a thematic analysis to identify patterns and themes. After analyzing the first 3 interview transcripts, 2 researchers (MH and TS) developed a coding framework using an inductive approach that allowed for identification of predominant themes. All emerging themes were cross-checked and discussed within the whole research team and developed iteratively to ensure definition and reliability. In the process, commonalities and differences between the respective perspectives were identified. This led to 9 different key themes concerning benefits and barriers, influencing factors on the intention to adopt a DiGA, and the outcome expectation from different perspectives as the main areas. Figure 1 lists the corresponding coding scheme according to Gioia et al [19]. The aim was to understand the specific properties of these areas and the influence of these factors in the context of the introduction of DiGAs.

![Figure 1. Coding scheme using the methodology by Gioia et al. DiGA: Digitale Gesundheitsanwendungen (digital health application).](https://mhealth.jmir.org/2024/1/e48345)
Ethical Considerations

This study did not require ethical approval according to the guideline of the applicable Ethics Committee of the Bavarian Universities [20], as no risks or harm was brought forward to the participants. All participants received an information and consent form explaining the requirements for participation. This included the opportunity to have the form explained to them if needed. All interviewees gave written or verbal consent before the interview started. Data collection, storage, and analysis were conducted in adherence to European Union General Data Protection Regulation (EU-GDPR). None of the participants were compensated.

Results

Overview

Our data demonstrate that only 4 of the interviewed physicians already prescribed DiGAs and none of the patients had used a DiGA before (Table S1 in Multimedia Appendix 1). Nevertheless, most physicians have experienced mHealth apps themselves. These were mostly used for medical information in their professional routine or privately for personal fitness and nutrition goals.

On the patient side, 20 participants had already had experience with mHealth apps or were actively using them (Table S2 in Multimedia Appendix 1).

Factors Enabling or Hindering Physicians to Prescribe DiGAs

Overview

Our results indicated that DiGAs are not yet widely known and used at this early stage by the physicians we interviewed. They indicated that the level of information is still insufficient and that very few DiGAs are prescribed compared with the prescription of drugs. DiGAs represent a completely new and innovative approach, so entry challenges are not considered unusual. Nevertheless, these challenges are initially barriers to prescribing DiGAs according to our participants. We were able to identify the following factors that influence physicians to prescribe a DiGA or not.

Lack of Information

Another reason often mentioned in connection with rejection was a lack of information. Physicians often do not know which DiGAs are available and where they could get the necessary information. As a result, physicians are often reluctant and skeptical about prescribing DiGAs (“No, I have not been educated on what DiGAs are available, how to prescribe, and how to tell if a DiGA is effective. I would not know where to look for this information” [E2, E3, E8, E12, E16, E20, E21, E25, and E28]; “I have little time to search for all the information I need to prescribe a DiGA in good conscience” [E1, E4, E8, E15, E16, E20, and E21]). They tend to be negative out of concern for malpractice and the resulting liability risk. Likewise, the physicians described the requirements for evidence of a DiGA (requirements for security, functionality, data protection, information security and quality, and positive effects on care) in the assessment procedure of a DiGA as insufficient.

The DiGA Adoption Process of Physicians and Patients

The analysis indicates that for DiGAs in particular, additional steps are added to the traditional prescribing process. Because of the special role of the doctor as a gatekeeper, the doctor is the first to decide whether the patient is suitable for a DiGA. First, the familiar steps such as the patient’s trust in the doctor and the treating doctor’s determination of the patient’s medical condition represent the keystone of the process. Following these steps, when prescribing a DiGA, the next steps are the doctor’s consent to prescribe a DiGA under certain conditions and the assessment of the patient’s ability to use a DiGA. Thereafter, the doctor prescribes a DiGA, and the patient is given the opportunity to adopt it. The process ends with the expectation that the patient will continuously use the prescribed DiGA and report the results to the treating physician. This process is visualized in Figure 2, with the doctor and the patient being influenced by different factors that determine whether the DiGA is prescribed (doctor’s perspective) and accepted (patient’s perspective).

Figure 2. The adoption process for DiGAs by the physician and patient. DiGA: Digitale Gesundheitsanwendungen (digital health application).

Digitalization of Processes and Information

They see the benefit of a DiGA for their work as low as long as the patient files and the exchange between other actors in the health care system are not fully digitized. The introduction of an EHR in Germany has failed so far because of technical challenges.

Lack of IT Resources

The lack of IT resources was mentioned by the interviewed doctors as a barrier to prescribing DiGAs within the framework of the DVG. Numerous efforts have been made toward digitalization. These include networking with various players in the German health system, such as doctors, hospitals, pharmacies, and health insurance companies. So far, only limited resources are available for the implementation of these plans. The doctors interviewed describe that, for example, the provision of services for an electronic health record (EHR) could help to digitalize various processes and information. They see the benefit of a DiGA for their work as low as long as the patient files and the exchange between other actors in the health care system are not fully digitized.
app are not yet available, and the apps are therefore only provisionally included in the directory. A comprehensive explanation of the BfArM authorization process of a DiGA can counteract these problems. However, participants reported receiving little information from insurers, DiGA manufacturers, or the BfArM. Physician respondents stated that proactive communication from DiGA providers was limited and possible involvement in the development process was unknown (“These apps have often been developed without the support of physicians and are now being ‘fast-tracked’ to market—that is not exactly building trust” [E25]). This statement reflects a high level of mistrust. In contrast, one participant stated that he worked with an app development company to develop an app in his field and described that this approach was extremely helpful to reduce the information deficit, mistrust, and gain more confidence. This shows that involvement is possible, but that an exchange of information is necessary.

Lack of Certification

In medical circles, a lack of certification has been widely mentioned. Half of all interviewed physicians agree that DiGAs, similar to other medical products (eg, drugs and medical aids such as wheelchairs or bandages), should be provided with a known official certification to ensure more trust in the product (“I think a well-known and recognized certification in medicine could help to build confidence in DiGAs” [E2, E13, E15, E17, E26, and E28]). Certification would make it easy to recognize that the DiGAs are safe as a medical device and are also medically or technically suitable within the scope of the intended purpose stated by the manufacturer. A certification equivalent to that for medicines could be conceivable in this context.

Lack of Involvement

Another aspect that our interviewees criticized was the lack of involvement after the prescription of a DiGA. They described that they are little involved in the procedure after prescribing a DiGA and in many cases currently only take on the role of a “prescriber” (“I would like to be involved in the whole process from the manufacturing to the evaluation of the data with the patient” [E51]). This shows that the participating physicians apparently make a great distinction between a traditional medicine in the form of medication administration and the use of digital components. After prescribing a traditional medicine, the physician only gets feedback when the patient comes back after some time and tells them how the medicine works. But with digital options such as DiGAs, the requirement is now higher: participants demanded an adjustment of the involvement in the postprescription process. A preview of future digitization plans shows that the involvement of physicians in the digital feedback process will be considered.

Lack of Financial Incentives

Participants further stated that counseling for a DiGA is much more time-consuming than for medicines, but the monetary incentive is not there. As a result, we identified a lack of financial incentive. Financial pressure weighs on the physicians in this regard, which is not compensated for by health insurance companies. As a result, the incentive for prescribing (€2 [US $2.16] per prescription) and treatment support (eg, successful monitoring) is currently considered too low (“2 euros for prescribing or 7 euros for treatment support are in no way an incentive to prescribe a DiGA” [E2]).

Finally, the study also found differences in physicians’ skills, knowledge, and attitudes toward digital technology. We define this factor as digital literacy, which has 2 specific characteristics. Some of the surveyed physicians stated that they have the impression that a physician is either digitally interested and very open-minded or completely ignorant of new innovations, so that even educational conversations might fail (“Either you find it good as a physician and have dealt with it once or recommended it to your patients. Or you ignore it at first” [E6]). Along with this result, the physicians interviewed expressed the fear that they might lack knowledge, for example, when it comes to patients’ technical questions or that they would have to become a kind of “technical support” in the event of problems in this area or in the evaluation of DiGA analyses. Others, however, considered the introduction of DiGAs to be an advantage for their work and less of a hindrance or a problem.

Factors Enabling or Hindering Older Users’ Adoption of DiGAs

Overview

As indicated above, a process of influencing factors leads to the adoption of DiGAs by patients. Thus, if the physician determines a medical condition, is willing to prescribe a DiGA, and considers the patient’s ability to use a DiGA to be positive, the physician will eventually prescribe a DiGA so that the patient will have the opportunity to adopt it.

Lack of Information

Similar to the physicians, we also identified a lack of information on the patient side. The DiGA concept is rather rarely known by the patients, and experiences were only described in 3 of the 30 interviewed participants. Nevertheless, from the patient perspective, DiGAs are recognized as an innovative and profitable treatment option that can be carried out independently of the time and place of the doctor’s visit (“I want to have the flexibility to do my therapy when it suits me” [P1, P6, P9, P12, P18, P26, P29, and P30]). Patients described DiGAs as a helpful “bridge” and a refresher or repetition of therapy content, especially for patients waiting for an appointment with a specialist (eg, psychotherapist). Some patients pointed to a long history of illness and low chances of success of conventional therapies and considered DiGAs as another treatment option. The interviews thus confirmed that the population’s willingness for DiGAs is high. However, many patients lack further information (“My physician or my health insurance company haven’t informed me about it yet—how should I know?” [P2, P3, P12, P16, P19, P22, and P29]).

The factors influencing the adoption of DiGAs are discussed in more detail below. In this context, we are oriented toward the most well-known models in adoption research: the UTAUT [9], a model from information systems that measures the acceptance of a technology by users to gain access to individual user behavior, and the HBM [16], a theoretical model from health psychology that analyzes and predicts health-related behavior.
**Self-Efficacy**

Self-efficacy refers to a sense of control over one’s environment and behavior. Participants considered self-efficacy important but also taken for granted. A distinction was made between technology self-efficacy about the DiGA and health self-efficacy with a health aspect. We define technology self-efficacy as the patient’s belief in their competence to use the DiGA. Health self-efficacy here describes the patient being confident in managing their health. Participants expressed confidence in their technology self-efficacy. For all participants, everyday use of smartphones and various apps was normal and regular. Some indicated that they still prefer paper calendars or dislike apps for health prevention but saw digital documentation and treatment support as an advantage in case of possible illness (“I’m more used to pen and paper, but if you get used to it, it certainly has its advantages” [P1, P4, P12, P14, P21, and P29]). In terms of health self-efficacy, participants were mostly self-confident. No participant stated that they did not want to deal with illnesses or would rather not know how healthy or ill their body was. Most participants seemed to have a high level of health literacy and wanted to actively deal with their health (“Yes, I would rather be the person who would then say, I would like to know this to know my enemy and then fight against it” [P19]). Participants indicated that a personal sense of control facilitates health behavior change (“Then an app like that would be great because you can calm down a bit more and have more security for yourself” [P10]). Participants confirmed that they focus on the opportunities rather than the obstacles (“Yes, I usually don’t go right away, but after a short scare I always face all the problems, so I’m more for problem solving rather than suppression” [P3]).

**Perceived Threat**

Perceived threat was mentioned as a crucial central factor for the use of a DiGA in the context of an impending or existing chronic disease. According to the HBM, perceived threat stems from beliefs about perceived susceptibility to disease and the perceived severity of disease consequences [16]. Susceptibility here refers to a person’s risk of contracting a disease. Severity refers not only to the medical consequences but also to the potential impact of an illness on a person’s daily life, family life, and social relationships. Participants indicated that a threat has a strong influence on health behavior. They explained that if they were seriously ill, they would use any means to support treatment, both digital and analogue (“So, I would say if I got sick now, I would be very interested in a DiGA already” [P29]; “Yes, when I get sick, I try everything possible to get better, whether digital or not” [P3]). Nevertheless, participants indicated that they generally felt very body and health conscious (“I eat healthy and exercise as much as I can, that’s part of my everyday life” [P9, P11, P15, P16, P21, P23, P26, and P29]). Only 1 respondent indicated that they felt an increased fear of possible illness. All participants indicated that they were primarily concerned with their health and well-being and that preventive measures were a natural part of their lives (“I prefer to focus positively on my health instead of worrying daily about illness and negative health” [P23, P29, and P31]).

**Trust in the Physician, Attitudes Toward Privacy, and Safety Concerns**

Partly different from classical adoption research, trust in the physician, attitudes toward privacy, and safety concerns were mentioned as key characteristics. Due to the medical field, the existing models require expansion. Trust does not describe a direct influencing factor but rather a precondition. For the participants, a deep trust relationship with their physician was crucial to consult a physician in case of a health problem and to receive good medical treatment. If this precondition is not given, an exchange about digital treatment options does not take place (“If I feel that I cannot trust him, then I would change” [P1, P7, P10, P13, and P22]; “The most important thing is the trust relationship. If I don’t trust my physician, I change physicians” [P27]). Subordinated are the characteristics of privacy and security. All participants considered it important to be able to set privacy settings themselves. The vast majority said that they did not particularly care about the content of privacy settings, but wanted to decide for themselves who could access which health data (“I would like to know who knows what about me” [P6, P7, P8, P15, and P26]; “There should be settings options. I don’t want to transmit everyday occurrences” [P9, P10, P11, P13, and P27]). Based on the accreditation of a DiGA as a medical device and the assessment by the BfArM, they have no concerns about the safety of DiGAs and feel confident in using DiGAs (“I trust that our federal system is highly secure” [P11]). On the basis of these correlations, we categorized the 3 concepts together. We rank the trust factor as the most important, as the privacy and security factors can be mitigated by a high level of trust.

**Social Influence**

The concept of social influence reflects the effect of environmental factors, for example, the opinion of friends and family, and is a significant factor in traditional adoption research [21]. However, in this study, we could not find any relevant results. Here, the medical context seems to have an important role. Most participants indicated that they did not discuss their health behaviors, personal diagnosis, or treatment plans with their social circle (“I don’t want to share all health data, including the fact that I use a health app, with other people” [P9]). We suspect that social influence may be a factor that is difficult to capture due to the sensitive nature of the data.

**Outcome Expectations**

Outcome expectations are defined as the expected consequences of a certain health behavior, which can be negative or positive [22], that is, what does the patient expect from using a DiGA. Participants stated that they would be able to reduce the frequency of visits to the physician, thus saving travel and time; that an existing illness would be better monitored by the physician; that they would feel safe and well cared for as a result; and that they would receive health-promoting treatment in the form of the DiGA. However, these expectations were tied to the continued use of features of the DiGA. Participants agreed that a DiGA must be simple to use, regardless of age. Many participants expect the app to provide more detailed information about the disease, symptoms, medication, and contraindications. However, this information needs to be understandable,
meaningful, and informative for every patient, regardless of age, education level, and professional background (“It should be easy for me, and I should be able to understand and comprehend it. The ease of use.” [P1, P5, P16, P19, and P21]).

Facilitating Conditions

Eventually, these expectations also cross over to the facilitating conditions, which includes the impact of the patient’s knowledge, skills, and resources. Here, the participants considered technical support to be particularly important. This was not because they felt unsure about using the technology but rather because it was a new type of health intervention. Some participants considered the physician to be an appropriate point of contact when difficulties arose in a few cases. Other participants stated that if they had difficulties or questions, they would ask their family and friends for help, as they do with other technical matters (“I think it would be good to have a number that I can call and that can help me” [P2]; “Well, I trust my son, he knows me anyway. And he would also know the diseases I have” [P18]). In summary, this study distinguishes between technical infrastructure and health support, with the health aspect (eg, knowledge and understanding of the diseases and their treatment) being more important here.

Post Hoc Analysis: Physician Assessment of the Patient’s Ability to Use a DiGA

During our interviews with the physicians, we discovered a phenomenon that we did not initially anticipate and that has not yet been described in the relevant technology adoption literature: the pre–user adoption decision of another instance, whether the user will be able and willing to voluntary use a technology. Typically, in technology adoption research analyses, this means whether a user is willing to use a technology in an either voluntary or mandatory environment. In any case, the decision remains with the user.

Now we see a new phenomenon: the assessment of one instance (physician), whether a subsequent instance (patient) would be able to use the DiGA. Only if the assessment is positive would the physician offer the DiGA to the patient, which will then trigger the traditional adoption questions and corresponding behavior of the user, as described in the well-researched technology adoption models such as the technology acceptance model, UTAUT, and HBM [9, 16, 23].

There are no guidelines under which circumstances a patient should be assumed to be able to use a DiGA. Thus, each physician needs to do this assessment individually. If they conclude that the patient will likely not be able to install, maintain, and use the app as foreseen by its developers, there is no point in prescribing the DiGA. As there are no objective guidelines, the assessment is either done explicitly, by asking the patient, or implicitly, by assuming what the patient is capable of.

Interestingly, the physicians interviewed were very consistent regarding the assessment of digital literacy of their patients. They indicated that prioritization of certain patient groups is facilitated by anticipating the digital literacy of their patients. Unfortunately, this often leads to a negative bias toward older users; the physicians described the typical DiGA user as a young and tech-savvy patient (“I would not consider my older patients for the use of DiGAs” [E1, E2, E4, E7, E8, E11, E12, E14, E15, E16, E17, E20, E21, E22, and E28]; “There are certainly exceptions, but most of my older patients are totally overwhelmed with a tablet or a smartphone, because the interest would not even be there” [E11]). Consistently, physicians expressed that they would not even consider an older patient as a DiGA user.

As mentioned before, these findings arose from the data and were not anticipated before. Thus, both issues, the second-order technology adoption process and the (possible) systematic disadvantage of older users, need deeper investigation in further research.

Discussion

Overview

In this research, we identified the salient factors that were either beneficial or hindering the adoption of DiGAs from the physician and the patient perspective. Furthermore, the results of our study suggest that the adoption process for a DiGA does not only depend on patient behavior but also on the physician’s behavior.

Most informants have a positive attitude toward the digitalization in general. Nevertheless, physicians’ demands on DiGAs are high, and their perceptions can be affected by a lack of facilitating conditions, trust, and digital competence. Certain influencing factors for the adoption of DiGAs by patients are consistent with the literature on established adoption research [9, 24–28].

Principal Implications

Our study contributes to the field by investigating factors influencing the adoption of DiGAs to inform future research and guide strategies and efforts for this user group. DiGAs represent a wide range of assistive apps that aim to support disease behaviors, manage various health conditions, and maintain the well-being of those with chronic diseases. There are very few empirical studies addressing the factors influencing users’ adoption of DiGAs [7, 29, 30]; hence, there is limited knowledge and guidance from the existing literature.

First, it is important to demonstrate that existing technology acceptance models reach their limits when used in the context of DiGAs. In contrast to Davis [23] and Venkatesh et al [9], our interviews with physicians and potential patients led to the assumption that, in addition to usefulness and ease of use, there are more constructs that play a significant role. We found that technology and health aspects such as technology- and health-related self-efficacy, trust, and a trustful doctor-patient relationship play a major role in the intention to use DiGAs. So far, these aspects have rarely been brought together. A study by Uncovska et al [30] confirms this finding. However, we note that there are few studies on the adoption process of DiGA. Other studies regarding mHealth app adoption have highlighted that health consciousness of individuals is a factor that directly influences both the intention to use mHealth apps and the actual use behavior [31, 32]. Public trust in the health care system [33] and a strong doctor-patient relationship can empower patients to contribute to treatment decision-making [34].
Second, we found that our interviewees on the patient side distinguished between technology-related and health-related self-efficacy. The consideration of a health component is not integrated into traditional technology adoption research. To date, it has not seemed necessary to consider the health domain in adopting general technologies. However, this is an important difference in the adoption of DiGAs. With DiGAs, the focus is on the health aspect for both the physician and the patient. Patients who may have low self-efficacy with technology do not simultaneously have to have low self-efficacy with their health. Older people, in particular, may have very sophisticated health self-efficacy while lacking technology self-efficacy [35-38]. Distinguishing these forms of self-efficacy provides a more detailed explanation of adoption behavior, increasing our understanding in the context of DiGAs. Personal beliefs, such as outcome expectations and self-efficacy expectations, are among the most critical variables in terms of intention formation and bridging the gap between intention and behavior, according to existing literature [39-41]. Nevertheless, a division into different areas of self-efficacy has not yet been made in information systems research but also in research on health adoption. As technology and health self-efficacy positively impact the adoption of DiGAs, we believe that it is important to consider both factors.

Third, related to the previous aspect, is the construct of perceived threat. Previous research shows that people are concerned about adopting technology in different areas, such as privacy, effort, or performance [7,42,43]. However, when using DiGAs, health is firmly in focus from a medical perspective. Therefore, the perceived threat of diseases and the need to use a DiGA strongly influenced the adoption of a DiGA. In our interviews, this construct was strongly emphasized, and we suspect moderation effects on other constructs, such as technology and health self-efficacy. A high perceived threat can increase the influence of the perceived health self-efficacy on adopting a DiGA because a threat can be better assessed by someone with high health self-efficacy and is, therefore, more likely to act. As a result, it is also possible that DiGAs will have a higher use rate, especially for hazardous diseases. A recent study by Pourhaji et al [44] investigated the perceived threat and stress response to the COVID-19 pandemic and found that the Iranian population’s health behavior was influenced by the perceived severity and susceptibility of the infection, which meant that preventive interventions were more likely to be accepted. Further studies related to COVID-19 found that risk severity also tends to increase with age, but the perception of susceptibility to contracting COVID-19 decreases [45-47]. Thus, risk perception does not seem to increase with age, but vulnerability and severity show opposite patterns [48]. The HBM postulates that individual beliefs about risk can be influenced by various factors such as sociodemographic and sociopsychological variables as well as knowledge, experience, and awareness [16]. However, patient awareness can also become an important issue, as these patients may not perceive a threat and, therefore, not adopt a DiGA.

Fourth, in addition to the patient, this study involves another important stakeholder, the physician. This stakeholder is not considered in the technology adoption models as they do not provide a specific gatekeeper for the technology or basically consider different stakeholders. But in the case of DiGA adoption by the patient, the first step requires the physician’s adoption of a DiGA. This observation has also been noted in previous studies [7,49-52]. Subsequently, the physician’s positive assessment of the patient’s competence to perform a certain behavior is one of the essential conditions for a patient to consistently perform a health intervention. This result relates to previous research without reference to DiGAs as well as with reference to DiGAs [7,53]. Similarly, this view can be developed in a negative direction when doctors decide that the patient is not capable of adopting and using a DiGA, which could be justified by digital ageism. Ageism is a societal bias conceptualized as (1) prejudicial attitudes toward older adults, (2) discriminatory practices toward older adults, or (3) institutionalized policies and social practices that promote these attitudes [54]. Ball et al [55] show that both the development and use of technology have excluded older adults, resulting in a “physical-digital divide,” which exists when a group feels excluded because they are unable to engage with the technologies used around them. Some studies suggest that ageism is widespread in the health care system [56-58]. For example, Walter et al [59] showed that physicians promote less preventive care for older patients. Chu et al [60] emphasized that the exclusion of older people from technology development leads to a broader cycle of inequality and ageist social attitudes, widening the digital divide. In contrast, we noticed that the trust factor impacts a patient’s health behavior, which is in line with Wildenbos et al [61]. Beyond this, we also found that the physician’s trust in the DiGA is equally important for their prescription of a DiGA. A physician needs a strong relationship of trust with the patient to convince the patient of the treatment methods. By motivating their patients to adopt a DiGA and use it to support their therapy, physicians focus on their social influence on the patient [7,62]. We provide justification for the incorporation of the physician as an important influence on adoption behavior in this context. After all, DiGAs live and die with physicians’ willingness to prescribe to their patients and influence them to understand the technology’s necessity. We argue for an adoption model that does not only incorporate a human-technology interaction but also a human-human-technology interaction.

**Implication for Practice**

The results of our study demonstrated that there is insufficient information available and published for both physicians and patients, as well as a lack of comprehensive technical support. Some statements (eg, the involvement of medical professionals in the development process of a DiGA, DiGA list unknown, and lack of evidence-based sources) confirm the knowledge deficit. It is important for health policy makers and public authorities such as the Ministry of Health, the Medical Association, and insurers to address these issues. Extensive information and source references are needed to take into account the needs of physicians and to enable DiGAs to get started more effectively. Including the consideration of conflicting goals in technology development from the beginning seems necessary. Codevelopment can improve app use and effectiveness in the long term by using a user-centered design.
to develop DiGAs that are effective in chronic disease self-management [63,64]. Likewise, we demonstrated that physicians distinguish clearly between digital and traditional treatment options and have significantly higher expectations of digital resources. Despite a previously excluded position of the physician in the development and introduction process of medicines, the physician now anticipates more involvement in digital developments. In order to counteract the feeling of exclusion and disconnection from the supposedly nonmedically focused DiGA development, appropriate education seems to be required. One way to achieve this is through promotion within public networks, local authorities, and medical associations.

Insufficient participation in the follow-up to a prescription was also a concern for medical practitioners. In this context, it is important to fully communicate to physicians the many opportunities DiGAs offer. DiGAs have higher potential than other treatment interventions such as medicines to maintain a meaningful exchange of information and stay in touch with the patient even after prescription. We found that physicians feel a loss of control when the DiGA is prescribed and then used by the patient. A high degree of self-management is demanded of the patient, leaving the physician feeling incapable of action. Digital monitoring with the help of a DiGA also results in an advantage after the prescription compared with the conventional prescription of medical devices and medicines. Furthermore, negative attitudes and lack of digital competence among physicians are major barriers to physician prescription of DiGAs and, thus, patient adoption. At the same time, the lack of facilitating conditions and the high demands regarding the introduction of various digital changes (eg, EHR) exert excessive pressure on physicians. In this case, further information and education of physicians would be useful. In addition, a trial period of the DiGA can demonstrate digital connectivity to the physician.

Limitations and Future Research
Because of the exploratory character of the research design, our findings naturally lack generalizability and should be regarded as a first starting point on the investigation of a new phenomenon. As DiGAs are a new option for physicians, several interview partners have not really experienced them yet. We found that some participants, especially those who had not yet heard of DiGAs, found it difficult to properly understand the use and benefits of DiGAs.

Another issue corresponding to the novelty of the phenomenon is that informants’ perceptions change quickly. Therefore, our findings reflect the perceptions of interview partners in the early phase of introductions of DiGAs into the market. It is likely that some of the issues raised will not be present in a couple of years, when DiGAs are more common to the market and perceived as natural to prescribe as all other medicines today.

This poses interesting questions for further research. It would be interesting to conduct longitudinal studies to gain a better understanding about the diffusion of such innovations in the medical space from the legislative setting into the physicians’ toolkit and finally to the patients. This could generate valuable insights for future management of digital innovations in the medical area. In close conjunction to these questions, a cross-national comparison could generate advice for policy to smoothen the introduction phases of digital medical innovations in new other countries.

Finally, the study of the second-order adoption mechanisms—highlighted earlier in the document—could lead to interesting theoretical insights and valuable advice for practitioners to enhance the prescription and adoption of DiGAs and comparable digital innovations.

Conclusions
DiGAs provide an opportunity to support people with severe (often chronic) diseases, to live independently with greater confidence and understanding of their condition, better symptom management, and ultimately enhanced quality of life. Our study provides deep insights into the needs and circumstantial evidence that enables a better understanding of the perspectives and preferences for adopting DiGAs by physicians and potential patients. We found that there is a considerable lack of information on both physicians’ and patients’ sides, resulting in poor trust and digital competence. Furthermore, we identified several factors influencing the adoption of DiGAs, which led to a new understanding of adoption research concerning digital health technologies.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Interview guidelines and descriptive results.
[DOCX File, 42 KB - mhealth_v12i1e48345_app1.docx]

References
1. Albrecht UV. Chances and Risks of Mobile Health Apps (CHARISMHA). Hannover Medical School. 2016. URL: https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/A/App-Studie/charismha_abr_v.01.1e-20160606.pdf [accessed 2023-12-07]


5. The fast-track process for digital health applications (DiGA) according to Section 139e SGB V. A guide for manufacturers, service providers and users. Federal Institute for Drugs and Medical Devices. 2020. URL: https://www.bfarm.de/SharedDocs/Downloads/EN/MedicalDevices/DiGA_Guide.html [accessed 2023-12-07]


https://mhealth.jmir.org/2024/11/e48345


42. Schroeder T, Haug M, Gewald H. Data privacy concerns using mHealth apps and smart speakers: comparative interview study among mature adults. JMIR Form Res 2022;6(6):e28025 [FREE Full text] [doi: 10.2196/jmir.3283] [Medline: 24824062]


57. Pritchard KI. Have we been guilty of ageism in the primary treatment of breast cancer? Br J Cancer 2007;96(7):1011-1012 [FREE Full text] [doi: 10.1038/sj.bjc.6603697] [Medline: 17406347]


Abbreviations

BfArM: Bundesinstitut für Arzneimittel und Medizinprodukte (Federal Institute for Drugs and Medical Devices)
DIGA: Digitale Gesundheitsanwendungen (digital health applications)
DIGAV: Digitale-Gesundheitsanwendungen-Verordnung (Digital Health Applications Ordinance)
DVG: Digitale-Versorgungs-Gesetz (Digital Healthcare Act)
EHR: electronic health record
EU-GDPR: European Union General Data Protection Regulation
HBM: health belief model
IPA: interpretative phenomenological analysis
mHealth: mobile health
UN: United Nations
UTAUT: unified theory of acceptance and use of technology
WHO: World Health Organization
Evidence of How Physicians and Their Patients Adopt mHealth Apps in Germany: Exploratory Qualitative Study

Please cite as:
Schroeder T, Haug M, Georgiou A, Seaman K, Gewald H
Evidence of How Physicians and Their Patients Adopt mHealth Apps in Germany: Exploratory Qualitative Study
JMIR Mhealth Uhealth 2024;12:e48345
URL: https://mhealth.jmir.org/2024/1/e48345
doi: 10.2196/48345
PMID: 38231550

©Tanja Schroeder, Maximilian Haug, Andrew Georgiou, Karla Seaman, Heiko Gewald. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 17.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Engagement With a Remote Symptom-Tracking Platform Among Participants With Major Depressive Disorder: Randomized Controlled Trial

Katie M White1, PhD; Ewan Carr2, PhD; Daniel Leightley1, PhD; Faith Matcham1,3, PhD; Pauline Conde2, BSc; Yatharth Ranjan2, BSc; Sara Simblett1, PhD; Erin Dawe-Lane4, MSc; Laura Williams5, BA, MSc; Claire Henderson6,7, PhD; Matthew Hotopf1,7, PhD

1Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom
2Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom
3School of Psychology, University of Sussex, Falmer, United Kingdom
4Department of Psychology, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom
5NIHR MindTech MedTech Co-operative, Institute of Mental Health and Clinical Neurosciences, University of Nottingham, Nottingham, United Kingdom
6Health Services & Population Research Department, King’s College London, London, United Kingdom
7South London and Maudsley NHS Foundation Trust, London, United Kingdom

Corresponding Author:
Katie M White, PhD
Department of Psychological Medicine
Institute of Psychiatry, Psychology and Neuroscience
King’s College London
16 De Crespigny Park
London, SE5 8AF
United Kingdom
Phone: 44 020 7848 0002
Email: katie.white@kcl.ac.uk

Abstract

Background: Multiparametric remote measurement technologies (RMTs), which comprise smartphones and wearable devices, have the potential to revolutionize understanding of the etiology and trajectory of major depressive disorder (MDD). Engagement with RMTs in MDD research is of the utmost importance for the validity of predictive analytical methods and long-term use and can be conceptualized as both objective engagement (data availability) and subjective engagement (system usability and experiential factors). Positioning the design of user interfaces within the theoretical framework of the Behavior Change Wheel can help maximize effectiveness. In-app components containing information from credible sources, visual feedback, and access to support provide an opportunity to promote engagement with RMTs while minimizing team resources. Randomized controlled trials are the gold standard in quantifying the effects of in-app components on engagement with RMTs in patients with MDD.

Objective: This study aims to evaluate whether a multiparametric RMT system with theoretically informed notifications, visual progress tracking, and access to research team contact details could promote engagement with remote symptom tracking over and above the system as usual. We hypothesized that participants using the adapted app (intervention group) would have higher engagement in symptom monitoring, as measured by objective and subjective engagement.

Methods: A 2-arm, parallel-group randomized controlled trial (participant-blinded) with 1:1 randomization was conducted with 100 participants with MDD over 12 weeks. Participants in both arms used the RADAR-base system, comprising a smartphone app for weekly symptom assessments and a wearable Fitbit device for continuous passive tracking. Participants in the intervention arm (n=50, 50%) also had access to additional in-app components. The primary outcome was objective engagement, measured as the percentage of weekly questionnaires completed during follow-up. The secondary outcomes measured subjective engagement (system engagement, system usability, and emotional self-awareness).

Results: The levels of completion of the Patient Health Questionnaire-8 (PHQ-8) were similar between the control (67/97, 69%) and intervention (66/97, 68%) arms (P value for the difference between the arms=.83, 95% CI −9.32 to 11.65). The intervention
group participants reported slightly higher user engagement (1.93, 95% CI −1.91 to 5.78), emotional self-awareness (1.13, 95% CI −2.93 to 5.19), and system usability (2.29, 95% CI −5.93 to 10.52) scores than the control group participants at follow-up; however, all CIs were wide and included 0. Process evaluation suggested that participants saw the in-app components as helpful in increasing task completion.

**Conclusions:** The adapted system did not increase objective or subjective engagement in remote symptom tracking in our research cohort. This study provides an important foundation for understanding engagement with RMTs for research and the methodologies by which this work can be replicated in both community and clinical settings.

**Trial Registration:** ClinicalTrials.gov NCT04972474; https://clinicaltrials.gov/ct2/show/NCT04972474

**International Registered Report Identifier (IRRID):** RR2-10.2196/32653

(JMIR Mhealth Uhealth 2024;12:e44214) doi:10.2196/44214

**KEYWORDS** remote measurement; technology; engagement; app; depression; smartphones; wearable devices; engagement; symptom tracking; self-awareness; community; mobile phone

**Introduction**

**Background**

Multiparametric remote measurement technologies (RMTs), which comprise smartphone apps and wearable devices, have the potential to revolutionize the clinical care of people with chronic, episodic health conditions [1]. Major depressive disorder (MDD) is one such condition, characterized by the relapse and remission of low mood and anhedonia over time [2]. Continuously measured longitudinal RMT data on the symptoms of MDD (mood variability, activity, cognition, and sleep) can capture a less biased picture of clinical state than retrospective self-report data [3]. Research using multiparametric sources might identify signals that could potentially predict future depressive episodes [4]. Such data could be ultimately implemented in patient self-management and shared decision-making in clinical practice [5].

It is important to understand how users engage with RMTs for depression symptom tracking. A recent systematic review found that engagement with RMTs can be measured objectively, for example, as the number of app-based symptom-tracking assessments completed, and subjectively, for example, as the perceived usability of and experience of using the RMT system [6]. Higher levels of objective engagement result in increased data availability, which, in turn, increases the validity of the machine learning approaches used for relapse prediction [7]. Objective engagement can also be used as an indicator of real-world uptake [8,9]. Further evidence suggests that increased satisfaction with mobile health apps is positively associated with the intention to continually use the tools [10]. Therefore, understanding engagement with RMTs is key to realizing their potential for relapse prediction.

Previous studies have reported inconsistent levels of engagement with RMT systems. Data collection, based on the total data expected, ranges from 42% to 82% for app-based symptom reporting and from 50% to 75% for device wear time [11]. The largest, multisite study of multiparametric RMTs for tracking depression to date, Remote Assessment of Disease and Relapse–Major Depressive Disorder (RADAR-MDD) [12], tracked 623 participants for 2 years using a smartphone app for mood tracking and a wrist-worn wearable for continuous passive data collection. The study has recently reported data availability metrics; 55.4% (345/623) of the sample completed >50% of the self-reported mood questionnaires expected to be completed, and 70.1% (437/623) had wearable heart rate data for >50% of the study days. Qualitative analyses from RADAR-MDD have revealed that the presence of a physical research team providing technological support and planned task reminders was a fundamental facilitator of long-term engagement in the study [13]. To ensure the scalability and real-world implementation of RMT systems, it is important to investigate methods that maximize engagement with RMTs while minimizing the human resources needed.

Focusing on the user interface (UI) of RMT systems is the logical first step for promoting engagement. Positioning the design of system UI within a theoretical framework of behavior change could help maximize effectiveness [14]. The Behavior Change Wheel [15] posits that researchers should begin by identifying a target behavior before considering the barriers to and facilitators of this behavior in terms of capability, opportunity, and motivation (the capability, opportunity, motivation, and behavior [COM-B] model). In the case of RMTs, the target behavior can be defined as objective engagement with symptom monitoring tasks. A series of published studies have evaluated both perceived [11,16] and experienced [13] barriers to RMT use in MDD research. Factors such as the knowledge of the utility of the research (capability), motivation linked to mood (motivation), and confirmation of logged data (opportunity) have been suggested to be prominent. The Behavior Change Wheel further provides a series of “intervention functions” best suited to address these factors, each with its own related behavior change techniques. With regard to RMTs, these have been suggested to be the provision of information from credible sources, visual feedback on behavior, and access to support.

The app design literature provides several options for incorporating behavior change techniques into RMT system design. First, following the Fogg behavioral model [17], push notifications can provide a trigger to perform a behavior, such as completing a monitoring task. Notifications can include tailored content, such as insights into the benefits of self-monitoring, which serves to simultaneously motivate the
user to respond to the notification and engage them in future tasks [18]. Second, visual incentives, such as graphs, can be embedded into the app to reflect on patterns in user progress and spark intrinsic motivation to complete future tasks [19]. Visualization can also help users manage uncertainty by attending to information about themselves [20]. A combination of qualitative and single-arm evaluation studies supports the perceived value of data visualization [21,22] and progress viewing [19] in encouraging symptom-tracking completion. Provision of contact details directly within an app can allow the user to directly and immediately access support, if required.

Without a control group, it is difficult to quantify the effect of in-app components on engagement [23]. A randomized controlled trial (RCT) of a substance abuse tracking app [20] suggested that users were 5% more likely to self-report on a day if they received a prior notification with an inspirational quote, although these results were not statistically significant. Conversely, users were 2% less likely to self-report following the provision of personalized visual data summaries; however, this main effect was significantly moderated by the prior day task completion such that those who had not completed the previous task were 36% more likely to self-report after receiving data visualization [20]. Users receiving prompts with tailored health messages, such as those highlighting the beneficial effects of symptom monitoring, were 4% more likely to engage in self-monitoring via another app for mental well-being [24]. It is important to replicate this work with multiparametric symptom monitoring systems, as it is currently unclear which combination of in-app features best promotes engagement with these technologies.

**This Study**

This study aimed to evaluate whether in-app components could promote engagement with a multiparametric RMT system for symptom tracking in depression. We conducted a 2-arm RCT to compare the system as usual with an adapted system that contained informative notifications, a visual progress report, and access to the research team contact details as a substitute for planned research team contact. We measured engagement as both objective and subjective concepts. This study had the following four specific objectives: (1) to describe data availability in an RCT of a multiparametric RMT system for tracking depression, (2) to test whether in-app components increase the rates of objective data completion, (3) to explore how in-app components influence the subjective experience of using the app, and (4) to understand how the components of the system are used by participants via process evaluation measures.

For objectives 2 and 3, we hypothesized that participants using the adapted app (intervention group) would have higher engagement in symptom monitoring, as measured by both objective engagement (completion of mood questionnaires) and subjective engagement (usability, utility, and emotional self-awareness).

### Methods

#### Ethical Considerations

This study was approved by the Psychiatry, Nursing, and Midwifery Research Ethics Subcommittee at King’s College London (reference number: RESCM-20/21-21083) and registered as a clinical trial (reference number: NCT04972474). A trial protocol has been previously published [25].

#### Trial Design

This was a single-center, 2-arm, parallel-group RCT (participant-blinded) with 1:1 randomization conducted in London, United Kingdom. We compared a remote symptom-tracking system (RADAR-base [26]; the control arm) with a system that contained additional in-app components (the intervention arm). Both the control and intervention arms were delivered via the RADAR-base system [26] using a smartphone app and a wearable Fitbit Charge (Fitbit Inc) device. Participants in the intervention arm had additional access to (1) theoretically informed notifications, (2) progress visualization, and (3) research team contact details through the study app. All participants were asked to use the system for 12 weeks.

Data were collected at baseline (0 weeks) and follow-up (12 weeks after randomization). Participants in both arms were sent 3 symptom-tracking tasks each week via the app; Fitbit data were collected continuously.

#### Participants

All participants were recruited from the RADAR-MDD study between April and May 2021. The inclusion criteria were as follows: (1) previous participation in the RADAR-MDD study at the London site (which required experiencing at least 1 episode of MDD in the 2 years before enrollment), (2) consent to be contacted, (3) willingness and ability to continue to use an Android (Google LLC) smartphone (provided for use by RADAR-MDD; see the study by Matcham et al [3] for the full study protocol), and (4) willingness and ability to complete a remote enrollment session owing to the COVID-19 pandemic. Participants were excluded if they were diagnosed with one of the following comorbid psychiatric disorders: bipolar disorder, schizophrenia, psychosis, schizoaffective disorder, or dementia.

Potential participants were invited to take part (up to 3 invitations were sent per participant, as per ethical considerations) and subsequently checked for eligibility, both via email. If eligible, contact details were entered into the REDCap (Research Electronic Data Capture [27]) system, which emailed an automated link to the informed consent form and baseline questionnaires. After participants provided consent and completed the baseline questionnaires, they were sent a link to book an enrollment session (via email, phone call, or video call).

On the day of the enrollment session, the principal investigator (KMW) initiated the REDCap randomization module and generated unique QR codes to link the study devices to the RADAR-base management portal. Each participant was sent a personalized set of instructions for downloading and logging...
into the system using the QR codes at the chosen enrollment time, accompanied by a phone or video call as requested.

Participants were purposefully not contacted by the research team during the follow-up period, aside from sending 1 check-in email at the 6-week time point. However, participants were able to initiate contact with the team if they had any queries during follow-up. The research team did not make withdrawals based on “lost to follow up,” given the fundamental aims of the study; however, participants were aware that they could withdraw at any point.

Suicidal ideation was assessed at baseline and follow-up using the Inventory of Depressive Symptomatology–Self-Report [28] item “thoughts of death or suicide.” Participants who reported suicidal ideation and intent at either time point were contacted via phone call by the principal investigator, advised to contact their treating physician, and emailed a list of signposting resources.

At the 12-week end point, participants were directed to debrief information that explained the aims of the study and provided instructions for logging out of the system.

RADAR-Base

The RADAR-base system is an open-source platform that supports data collection via remote devices [3,26]. It requires users to download and log into an Android smartphone app in addition to wearing and syncing a wearable device. All participants were asked to complete the following three validated symptom-tracking tasks per week via the study app: (1) Patient Health Questionnaire-8 (PHQ-8 [29]); (2) Rosenberg Self-Esteem Scale (RSES [30]); and (3) a speech task, during which the user records themselves reading aloud a short paragraph (Multimedia Appendix 1). All tasks became available on the same day each week, 1 hour apart, beginning at the point of enrollment. All tasks had to be completed within 24 hours.

Interventions

Control Arm

Participants were sent 3 tasks per week via the RADAR-base study app, as outlined in the previous section. For each task, they received a notification on the day that the task was due that read, “Questionnaire Time. Won’t usually take longer than 3 minutes.” They were unable to view any data other than those available on the Fitbit app.

Intervention Arm

The design of the additional in-app components was grounded in behavioral theory and user research on the barriers to and facilitators of RMT use in patients with MDD [11,13,16]. The COM-B [31] framework of behavior change highlighted education, incentivization, and enablement as the most suitable forms of intervention function. Findings from research with users of the RADAR-base system allowed for the translation of these functions into tangible components tailored specifically to the needs and preferences of the target cohort [32]. It was decided that an engaging app should include notifications with information on symptom tracking from a credible source, behavioral feedback via progress visualization, and instant access to researcher contact details (see the study by White et al [25] and Multimedia Appendix 2 for a detailed overview of this process).

Participants in the intervention arm received notifications and tasks at the same time as those in the control arm but with the following additional content (Figure 1):

1. Theoretically informed notifications: the notifications included additional sentences that described the potential benefits of symptom monitoring for emotional self-awareness, clinical practice, and research. Participants were also reminded that they could complete the task “any time today.”

Figure 1. Screenshots of the in-app components included in the intervention arm.
2. Progress visualization: participants were provided with a graph in the app that tracked the completion of the tasks. This graph could be viewed at any time from the main app home page.

3. Researcher contact details: the main app home page included a phone number, an email address, and contact hours of the research team for the reporting of technical issues or requests for support.

Measures
A summary of measures and data collection time points is presented in Table 1. The measures were identical between the intervention and control arms.
Table 1. A summary of measures and data collection points across the 12-week follow-up period.

<table>
<thead>
<tr>
<th>Measures</th>
<th>Baseline</th>
<th>End point</th>
<th>Weekly</th>
<th>Continuously</th>
</tr>
</thead>
<tbody>
<tr>
<td>REDCap(^a) survey</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consent</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contact information</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study devices</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sociodemographics</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social environment</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical history</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIDAS(^b)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDS-SR(^c)</td>
<td>✓✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The World Health Organization CIDI-SF(^d)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAD-7(^e)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WSAS(^f)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIPQ(^g)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Life events</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSRI(^h)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UES(^i)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESQ(^j)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAUQ(^k)</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Active app measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHQ-8(^l)</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>RSES(^m)</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Speech task</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>Fitbit</strong></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Heart rate, step count, and GPS</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td><strong>Process evaluation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>App use metrics</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Qualitative interviews</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

\(^a\)REDCap: Research Electronic Data Capture.  
\(^b\)LIDAS: Lifetime Depression Assessment Self-Report.  
\(^c\)IDS-SR: Inventory of Depressive Symptomatology–Self-Report.  
\(^d\)CIDI-SF: Composite International Diagnostic Interview-Short Form.  
\(^e\)GAD-7: Generalized Anxiety Disorder-7.  
\(^f\)WSAS: Work and Social Adjustment Scale.  
\(^g\)BIPQ: Brief Illness Perception Questionnaire.  
\(^h\)CSRI: Client Service Receipt Inventory.  
\(^i\)UES: User Engagement Scale.  
\(^j\)ESQ: Emotional Self-Awareness Questionnaire.  
\(^k\)MAUQ: mHealth App Usability Questionnaire.  
\(^l\)PHQ-8: Patient Health Questionnaire-8.  
\(^m\)RSES: Rosenberg Self-Esteem Scale.
Questionnaires

After registration for the study, participants completed web-based baseline questionnaires via REDCap, providing information on sociodemographics and physical and mental health history, including the presence of depression, recent life events, and service use. The principal investigator also manually extracted data pertaining to previous participation in the RADAR-MDD study, including participation length and technology use. At the 12-week time point, participants repeated these questionnaires.

Outcome Measures

The primary outcome was objective engagement with the system, measured as the number and percentage of weekly PHQ-8 questionnaires completed during follow-up (compared with the total of 12 questionnaires that were sent). Completion of 1 PHQ-8 questionnaire was defined as the completion of all 8 questions.

There were four secondary outcomes, three of which measured subjective engagement with the system:

1. User engagement: this was measured using the User Engagement Scale (UES) [33] adapted to mobile health use [34], a 30-item questionnaire measuring focused attention, perceived usability, esthetic appeal, and reward. All items are scored on a Likert scale ranging from 1 (“strongly disagree”) to 5 (“strongly agree”). Total scores are calculated by summing the scores for each item in each of the 4 subscales and dividing the resultant value by the number of items in each subscale. An overall engagement score can be calculated using the average of each subscale. A higher average score indicates higher user engagement. The UES has been widely adopted and shows good reliability and construct validity [35].

2. Emotional self-awareness: this was measured using the Emotional Self-Awareness Questionnaire (ESQ) [36], a 33-item scale measuring recognition, contextualization, and decision-making in relation to self-emotion. All items are scored on a 5-point Likert scale ranging from 0 (“never”) to 4 (“a lot”). The total score is calculated as a continuous variable that ranges from 0 to 132, with a higher score reflecting higher emotional self-awareness. The ESQ has a reliability of 0.92 and shows significant positive correlations with the Emotional Intelligence Test [36].

3. System usability: this was measured using the mHealth App Usability Questionnaire (MAUQ) for stand-alone apps [37]. The MAUQ is an 18-item scale that measures the immediate and long-term self-reported usability of an app, including its ease of use and utility for self-management (overall Cronbach α=.914). All items are scored on a 7-point Likert scale ranging from 1 (“disagree”) to 7 (“agree”). The app usability score is calculated as the sum of scores across the items for each participant, ranging from 18 to 126, with a higher score reflecting higher reported usability of the app.

4. Overall adherence to the RADAR-base system: a participant was considered to have adhered to the system if they (1) responded to at least 50% of the 3 weekly tasks and (2) >2 heart rate data points were recorded by the Fitbit device on at least 50% of the days during the 12-week intervention period. This measure was chosen to align with previous data availability reporting [12] and other studies [38].

Process evaluation measures were collected to evaluate the use of the in-app components. Quantitative measures covered app engagement, in-app interactions, and notification engagement. A total of 20 participants, split evenly across the 2 arms, were also invited at the study endpoint to qualitatively discuss their experiences with the components through a 1:1 interview (Multimedia Appendix 3).

Sample Size

Power calculations were performed based on data availability from the RADAR-MDD study [12]. To detect a difference of 25% completion of PHQ-8 tasks between the control and intervention arms, with 80% power and 95% CIs, 132 participants were required (66 per arm). We decided on 25% as the minimum difference that would be practically useful for analyses.

Randomization and Blinding

Participants were randomly allocated in a 1:1 ratio to either the control or intervention arm using simple randomization via the REDCap randomization module.

The principal investigator was unblinded to allocation to ensure that remote enrollments had been carried out successfully and had access to incoming data throughout the study. The trial data manager (DL) was blinded to arm allocation, as this information was stored elsewhere. Participants had previously used the RADAR-base system and, therefore, could not be fully blinded to arm allocation. However, the explicit aims and arm assignments of the study were not revealed until study debrief.

Statistical Methods

Sociodemographic and clinical variables at baseline were described by arm using appropriate summary statistics (counts and percentages for categorical variables and mean and SD or median and IQR for continuous variables). We reported data availability for all outcomes. Data availability for each study app task (PHQ-8, RSES, and speech task) was summarized as the median (IQR) number of weekly tasks completed. Fitbit wear time was summarized as mean (SD) days with >2 heart rate data points. The overall completion of all 4 data sources was also reported, calculated as a percentage of the total expected count (n=12) for the study app tasks and the total expected days of wear time (n=84) for the Fitbit.

The primary outcome, objective engagement, was analyzed using 2-sample 2-tailed t tests, which tested the difference in the mean percentage of PHQ-8 completion over 12 weeks between the study arms.

Three secondary outcomes (UES, ESQ, and MAUQ) were analyzed using separate linear regression models. Each model included the follow-up score as the dependent variable and arm allocation (0=control; 1=intervention) as the only covariate. Models for outcomes measured at baseline and follow-up (UES and ESQ) additionally included the baseline values of the outcomes. Differences in the combined adherence to the system (0=<50% total data completion; 1=>50% total data completion)
were tested using Pearson chi-square test. The threshold for statistical significance in all the analyses was \( P = .05 \).

All outcomes were analyzed under the intention-to-treat principle using R (version 4.1; R Foundation for Statistical Computing) [39]. All data, including those from withdrawn participants, were included in the analyses.

**Supplementary Analysis**

A supplementary analysis of all outcomes was conducted to estimate complier average causal effect (CACE) [40]. We defined compliers as participants in the intervention group who viewed the progress report page at least once in the 12 weeks. The CACE analyses were performed using a 2-stage least squares regression with arm allocation as the instrumental variable.

**Process Evaluation**

We examined quantitative app use measures from data retrieved from Google Analytics (Google LLC). These were reported under the following three categories: (1) app engagement (user-initiated app opening and active weeks), (2) in-app interactions (questionnaire initiation, progress report viewing, viewed progress report >1 time, and progress report viewing duration), and (3) notification engagement (notifications received, notifications opened, and the percentage of notifications opened out of notifications received). The number of active weeks was calculated as the number of weeks the participant was active out of the total 12 weeks, with at least 3 screen view or user engagement metrics recorded per participant. Each indicator was summarized by arm as mean (SD) and median (IQR). The qualitative experiences of the study were thematically analyzed and reported as a brief narrative synthesis. These data will be reported in more detail elsewhere.

**Results**

**Recruitment**

A total of 347 individuals were contacted between April and May 2021. Of them, 114 (32.9%) agreed to participate, and 100 (28.8%) completed an enrollment session and were enrolled in the study. Enrollment sessions took place via email (89/100, 89%), video call (9/100, 9%), or phone call (2/100, 2%). Figure 2 details the participation rate and reasons for nonparticipation.

**Sample Characteristics**

All (100/100, 100%) participants completed the baseline outcome assessment, and 87 (87%) participants completed the 12-week follow-up assessment. Among the total 100 participants, 1 (1%) participant in the intervention group withdrew from the study before the 12-week point, citing technological issues with the study apps as the main reason for withdrawal. The follow-up period was from April to September 2021.

Baseline characteristics were similar between the 2 groups (Table 2). The groups contained an equal number of participants (n=50). The mean age of the sample was 53.3 (SD 14.3) years, and 76 (76%) of the 100 participants were female. Most participants reported mild (36/100, 36%) or moderate (29/100, 29%) symptoms of depression at enrollment, as measured by the Inventory of Depressive Symptomatology–Self-Report. Overall, 12 (12%) participants reported suicidal ideation at baseline. Among the 100 participants, 59 (59%) “strongly agreed” that they were confident in using the smartphone they were using for the study, and 51 (51%) “strongly agreed” that they were confident in using the Fitbit device.
Figure 2. Participant flowchart following CONSORT (Consolidated Standards of Reporting Trials) guidelines.
Table 2. Baseline characteristics and demographics of the study sample.

<table>
<thead>
<tr>
<th></th>
<th>Intervention (n=50)</th>
<th>Control (n=50)</th>
<th>Overall (N=100)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age (y), mean (SD)</strong></td>
<td>55.3 (12.7)</td>
<td>51.2 (15.7)</td>
<td>53.3 (14.3)</td>
</tr>
<tr>
<td><strong>Gender, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>40 (80)</td>
<td>36 (72)</td>
<td>76 (76)</td>
</tr>
<tr>
<td>Men</td>
<td>10 (20)</td>
<td>14 (28)</td>
<td>24 (24)</td>
</tr>
<tr>
<td><strong>Ethnicity, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black or mixed ethnicity</td>
<td>3 (6)</td>
<td>3 (6)</td>
<td>6 (6)</td>
</tr>
<tr>
<td>White British</td>
<td>40 (80)</td>
<td>41 (82)</td>
<td>81 (81)</td>
</tr>
<tr>
<td>White other</td>
<td>5 (10)</td>
<td>4 (8)</td>
<td>9 (9)</td>
</tr>
<tr>
<td>Other</td>
<td>2 (4)</td>
<td>2 (4)</td>
<td>4 (4)</td>
</tr>
<tr>
<td>Total time in education (y), mean (SD)</td>
<td>20.2 (3.34)</td>
<td>20.5 (3.71)</td>
<td>20.4 (3.51)</td>
</tr>
<tr>
<td><strong>Benefit receipt, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>24 (48)</td>
<td>23 (46)</td>
<td>47 (47)</td>
</tr>
<tr>
<td>No</td>
<td>26 (52)</td>
<td>27 (54)</td>
<td>53 (53)</td>
</tr>
<tr>
<td><strong>Income (£; US $), n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;15,000 (US $18,828.67)</td>
<td>9 (18)</td>
<td>12 (24)</td>
<td>21 (21)</td>
</tr>
<tr>
<td>15,000-24,000 (US $18,828.67-$30,125.88)</td>
<td>8 (16)</td>
<td>9 (18)</td>
<td>17 (17)</td>
</tr>
<tr>
<td>24,000-40,000 (US $30,125.88-$50,209.8)</td>
<td>15 (30)</td>
<td>10 (20)</td>
<td>25 (25)</td>
</tr>
<tr>
<td>40,000-55,000 (US $50,209.8-$69,038.47)</td>
<td>11 (22)</td>
<td>7 (14)</td>
<td>18 (18)</td>
</tr>
<tr>
<td>&gt;55,000 (US $69,038.47)</td>
<td>7 (14)</td>
<td>12 (24)</td>
<td>19 (19)</td>
</tr>
<tr>
<td><strong>Employment status, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employed</td>
<td>24 (48)</td>
<td>25 (50)</td>
<td>49 (49)</td>
</tr>
<tr>
<td>Sick leave</td>
<td>0 (0)</td>
<td>3 (6)</td>
<td>3 (3)</td>
</tr>
<tr>
<td>Retired</td>
<td>20 (40)</td>
<td>14 (28)</td>
<td>34 (34)</td>
</tr>
<tr>
<td>Unemployed</td>
<td>2 (4)</td>
<td>4 (8)</td>
<td>6 (6)</td>
</tr>
<tr>
<td>Other</td>
<td>4 (8)</td>
<td>4 (8)</td>
<td>8 (8)</td>
</tr>
<tr>
<td>Current depression (continuous), mean (SD)(^a)</td>
<td>24.8 (13.7)</td>
<td>26.5 (13.3)</td>
<td>25.7 (13.5)</td>
</tr>
<tr>
<td><strong>Current depression (categorical), n (%) (^a)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>10 (20)</td>
<td>8 (16)</td>
<td>18 (18)</td>
</tr>
<tr>
<td>Mild</td>
<td>20 (40)</td>
<td>16 (32)</td>
<td>36 (36)</td>
</tr>
<tr>
<td>Moderate</td>
<td>12 (24)</td>
<td>17 (34)</td>
<td>29 (29)</td>
</tr>
<tr>
<td>Severe</td>
<td>5 (10)</td>
<td>7 (14)</td>
<td>12 (12)</td>
</tr>
<tr>
<td>Very severe</td>
<td>3 (6)</td>
<td>2 (4)</td>
<td>5 (5)</td>
</tr>
<tr>
<td><strong>Suicidal ideation, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>3 (6)</td>
<td>9 (18)</td>
<td>12 (12)</td>
</tr>
<tr>
<td>No</td>
<td>47 (94)</td>
<td>41 (82)</td>
<td>88 (88)</td>
</tr>
<tr>
<td>Current anxiety (continuous), mean (SD)(^b)</td>
<td>6.34 (4.62)</td>
<td>7.10 (5.21)</td>
<td>6.72 (4.92)</td>
</tr>
<tr>
<td><strong>Current anxiety (categories), n (%) (^b)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>22 (44)</td>
<td>17 (34)</td>
<td>39 (39)</td>
</tr>
<tr>
<td>Mild</td>
<td>16 (32)</td>
<td>20 (40)</td>
<td>36 (36)</td>
</tr>
<tr>
<td>Moderate</td>
<td>9 (18)</td>
<td>6 (12)</td>
<td>15 (15)</td>
</tr>
<tr>
<td>Severe</td>
<td>3 (6)</td>
<td>7 (14)</td>
<td>10 (10)</td>
</tr>
<tr>
<td></td>
<td>Intervention (n=50)</td>
<td>Control (n=50)</td>
<td>Overall (N=100)</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------------</td>
<td>----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Medical comorbidity, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>25 (50)</td>
<td>34 (68)</td>
<td>59 (59)</td>
</tr>
<tr>
<td>No</td>
<td>25 (50)</td>
<td>16 (32)</td>
<td>41 (41)</td>
</tr>
<tr>
<td>Functional disability, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No impairment</td>
<td>14 (28)</td>
<td>17 (34)</td>
<td>31 (31)</td>
</tr>
<tr>
<td>Some impairment</td>
<td>17 (34)</td>
<td>17 (34)</td>
<td>34 (34)</td>
</tr>
<tr>
<td>Significant impairment</td>
<td>19 (38)</td>
<td>16 (32)</td>
<td>35 (35)</td>
</tr>
<tr>
<td>Life events in the past year, mean (SD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.680 (1.04)</td>
<td>0.920 (1.07)</td>
<td>0.800 (1.05)</td>
</tr>
<tr>
<td>Confidence in smartphone use, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strongly agree</td>
<td>29 (58)</td>
<td>30 (60)</td>
<td>59 (59)</td>
</tr>
<tr>
<td>Agree</td>
<td>15 (30)</td>
<td>16 (32)</td>
<td>31 (31)</td>
</tr>
<tr>
<td>Neither agree nor disagree</td>
<td>5 (10)</td>
<td>2 (4)</td>
<td>7 (7)</td>
</tr>
<tr>
<td>Disagree</td>
<td>0 (0)</td>
<td>2 (4)</td>
<td>2 (2)</td>
</tr>
<tr>
<td>Strongly disagree</td>
<td>1 (2)</td>
<td>0 (0)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Confidence in Fitbit (Fitbit Inc) use, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strongly agree</td>
<td>27 (54)</td>
<td>24 (48)</td>
<td>51 (51)</td>
</tr>
<tr>
<td>Agree</td>
<td>16 (32)</td>
<td>22 (44)</td>
<td>38 (38)</td>
</tr>
<tr>
<td>Neither agree nor disagree</td>
<td>6 (12)</td>
<td>3 (6)</td>
<td>9 (9)</td>
</tr>
<tr>
<td>Disagree</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Strongly disagree</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Not using Fitbit</td>
<td>1 (2)</td>
<td>1 (2)</td>
<td>2 (2)</td>
</tr>
<tr>
<td>Existing RADAR-MDD(^c) status, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finished 2 years</td>
<td>0 (0)</td>
<td>2 (4)</td>
<td>2 (2)</td>
</tr>
<tr>
<td>Continuing past 2 years</td>
<td>13 (26)</td>
<td>17 (34)</td>
<td>30 (30)</td>
</tr>
<tr>
<td>Not reached 2 years</td>
<td>36 (72)</td>
<td>31 (62)</td>
<td>67 (67)</td>
</tr>
<tr>
<td>Withdrawn</td>
<td>1 (2)</td>
<td>0 (0)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Existing phone status, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Existing Android (Google LLC)</td>
<td>30 (60)</td>
<td>27 (54)</td>
<td>57 (57)</td>
</tr>
<tr>
<td>Switching from iPhone (Apple Inc)</td>
<td>13 (26)</td>
<td>14 (28)</td>
<td>27 (27)</td>
</tr>
<tr>
<td>Switching from nonsmartphone</td>
<td>2 (4)</td>
<td>4 (8)</td>
<td>6 (6)</td>
</tr>
<tr>
<td>Upgrading existing Android</td>
<td>5 (10)</td>
<td>5 (10)</td>
<td>10 (10)</td>
</tr>
</tbody>
</table>

\(^a^\)Measured using the Inventory of Depressive Symptomatology–Self-Report. The maximum score possible is 84. The scores are categorized as follows: none=0-13, mild=14-25, moderate=26-38, severe=39-48, and very severe=49-84.

\(^b^\)Measured using the Generalized Anxiety Disorder-7 item scale. The maximum score possible is 21. The scores are categorized as follows: none=0-5, mild=6-10, moderate=11-15, and severe=16-21.

\(^c^\)RADAR-MDD: Remote Assessment of Disease and Relapse–Major Depressive Disorder.

**Data Availability**

Of the 100 participants, 97 (97%) provided any data via the study app, and 93 (93%) had any recorded Fitbit data. A total of 2 (2%) participants were unable to use either the study app or Fitbit immediately following enrollment, 1 (1%) received no notifications from the study app during follow-up, and 4 (4%) were unable to sync the Fitbit with their smartphone. Moreover, 1 (1%) participant opted out of using a Fitbit for the study at enrollment. As data from the sources were unavailable owing to technical limitations, rather than nonengagement, these participants were excluded from the respective analyses (3/100, 3% for primary analysis and 8/100, 8% for the secondary combined adherence analysis).

Each app task had a maximum count of 12 (1 per week). Overall, participants completed a median of 9 (IQR 6-10) PHQ-8 tasks, 9 (IQR 6-10) RSES tasks, and 8 (IQR 6-9) speech tasks. Among the 100 participants, 2 (2%) completed all available tasks, and 7 (7%) completed all available PHQ-8 tasks. A total of 35 (35%)
participants completed all 3 tasks at each point when they completed a PHQ-8 task. The participants provided sufficient Fitbit data (at least 2 recordings per day) on a mean average of 74 (SD 19.7) days per participant during the 12-week (84-day) intervention period. **Figure 3.** Percentage of completion for the Fitbit and the 3 active tasks (Patient Health Questionnaire-8 [PHQ-8], Rosenberg Self-Esteem Scale [RSES], and speech task).

**Primary and Secondary Outcomes**

The primary analysis included 97 participants. The levels of completion of the PHQ-8 task were similar between the control (67/97, 69%) and intervention (66/97, 68%) arms (P value for the difference between the arms=.83, 95% CI −9.32 to 11.65). For the secondary outcomes (Table 3), we found that those in the intervention group reported slightly higher UES (1.93, 95% CI −1.91 to 5.78), ESQ (1.13, 95% CI −2.93 to 5.19), and MAUQ (2.29, 95% CI −5.93 to 10.52) scores than those in the control group at follow-up. However, all CIs were wide and included 0.
Table 3. Linear regression model coefficients for each of the 3 secondary outcomes.

<table>
<thead>
<tr>
<th>Subjective engagement outcome</th>
<th>Treatment effect (95% CI)</th>
<th>Participant (N=100), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UES&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1.93 (−1.91 to 5.78)</td>
<td>89 (89)</td>
</tr>
<tr>
<td>ESQ&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1.13 (−2.93 to 5.19)</td>
<td>89 (89)</td>
</tr>
<tr>
<td>MAUQ&lt;sup&gt;d,e&lt;/sup&gt;</td>
<td>2.29 (−5.93 to 10.52)</td>
<td>87 (87)</td>
</tr>
</tbody>
</table>

<sup>a</sup>For end point measures only.
<sup>b</sup>UES: User Engagement Scale.
<sup>c</sup>ESQ: Emotional Self-Awareness Questionnaire.
<sup>d</sup>MAUQ: mHealth App Usability Questionnaire.
<sup>e</sup>Only includes end point measure.

The combined adherence secondary analysis included 92 participants. The proportion of participants adhering to the system was similar between both arms (control=32/48, 67%; intervention=35/44, 80%; P value for difference between arms=.98).

For the CACE analyses, of the 48 participants in the intervention group, 29 (60%) met the complier definition of viewing the progress report at least once during the intervention period (Table 4). Table 5 presents the CACE treatment estimates for the primary and secondary outcomes. The between-arm difference in PHQ-8 completion was −1.92 (95% CI −19.93 to 15.54; P=.83), showing no evidence of a statistical difference. The treatment effect estimates for the UES, ESQ, and MAUQ were larger than the intention-to-treat estimates in favor of the intervention, but the effect sizes remained small.

Table 4. Number of participants and mean percentage of completion among the control group, intervention group compliers, and intervention group noncompliers (n=97).

<table>
<thead>
<tr>
<th>Study arm</th>
<th>Participant, n (%)</th>
<th>Completion (%), mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention group compliers&lt;sup&gt;a&lt;/sup&gt;</td>
<td>29 (30)</td>
<td>75.3 (23.9)</td>
</tr>
<tr>
<td>Intervention group noncompliers&lt;sup&gt;b&lt;/sup&gt;</td>
<td>19 (20)</td>
<td>57.0 (28.0)</td>
</tr>
<tr>
<td>Control group</td>
<td>49 (50)</td>
<td>69.2 (25.1)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Viewed the progress report module at least once during the intervention period.
<sup>b</sup>Did not view the progress report module during the intervention period.

Table 5. Intention-to-treat (ITT) and complier average causal effect (CACE) treatment estimates for the primary and secondary outcomes.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Participant, n (%)</th>
<th>ITT Treatment effect (95% CI)</th>
<th>CACE&lt;sup&gt;a&lt;/sup&gt; P value</th>
<th>CACE&lt;sup&gt;a&lt;/sup&gt; Treatment effect (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHQ-8&lt;sup&gt;b&lt;/sup&gt;</td>
<td>97 (100)</td>
<td>−1.16 (−11.65 to 9.32)</td>
<td>.83</td>
<td>−1.92 (−19.39 to 15.54)</td>
<td>.83</td>
</tr>
<tr>
<td>UES&lt;sup&gt;c&lt;/sup&gt;</td>
<td>89 (92)</td>
<td>1.93 (−1.91 to 5.78)</td>
<td>.32</td>
<td>3.49 (−3.75 to 10.73)</td>
<td>.34</td>
</tr>
<tr>
<td>ESQ&lt;sup&gt;d&lt;/sup&gt;</td>
<td>89 (92)</td>
<td>1.13 (−2.93 to 5.19)</td>
<td>.58</td>
<td>2.03 (−5.32 to 9.38)</td>
<td>.58</td>
</tr>
<tr>
<td>MAUQ&lt;sup&gt;e&lt;/sup&gt;</td>
<td>87 (90)</td>
<td>2.29 (−5.93 to 10.52)</td>
<td>.58</td>
<td>4.21 (−10.87 to 19.28)</td>
<td>.58</td>
</tr>
</tbody>
</table>

<sup>a</sup>Complier average causal effect estimates of intervention group compliers, defined as those who viewed the progress report module at least once during the intervention period.
<sup>b</sup>PHQ-8: Patient Health Questionnaire-8.
<sup>c</sup>UES: User Engagement Scale.
<sup>d</sup>ESQ: Emotional Self-Awareness Questionnaire.
<sup>e</sup>MAUQ: mHealth App Usability Questionnaire.

Process Evaluation

Table 6 presents the quantitative process evaluation measures collected throughout the study. Over the entire study period, participants in the intervention arm opened the app a mean average of 21.2 (SD 13.5) times, whereas the participants in the control group opened the app a mean average of 19.0 (SD 9.10) times. In total, 60% (35/58) of the participants who were able to view the progress report viewed it multiple times throughout the study, viewing for a mean average of 14.7 (SD 10.9) seconds per time. Participants in both groups received a similar number of notifications, although those in the control group opened a higher mean percentage (39.9, SD 25.9) of the notifications.
Table 6. Process evaluation use statistics by arm over the 12-week follow-up period.

<table>
<thead>
<tr>
<th>App engagement, mean (SD)</th>
<th>Intervention (n=48)</th>
<th>Control (n=49)</th>
</tr>
</thead>
<tbody>
<tr>
<td>User-initiated app opening</td>
<td>21.2 (13.5)</td>
<td>19.0 (9.10)</td>
</tr>
<tr>
<td>Active weeks(^a)</td>
<td>8.96 (3.14)</td>
<td>8.88 (2.60)</td>
</tr>
<tr>
<td>In-app interactions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Questionnaire initiation, mean (SD)</td>
<td>23.6 (10.3)</td>
<td>25.0 (10.5)</td>
</tr>
<tr>
<td>Progress report viewing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Values, mean (SD)</td>
<td>3.60 (7.64)</td>
<td>N/A(^b)</td>
</tr>
<tr>
<td>Values, median (IQR)</td>
<td>1.00 (0.25)</td>
<td>N/A</td>
</tr>
<tr>
<td>Viewed progress report &gt;1 time, n (%)</td>
<td>19 (40)</td>
<td>49 (100)</td>
</tr>
<tr>
<td>Yes</td>
<td>29 (60)</td>
<td>N/A</td>
</tr>
<tr>
<td>Progress report viewing duration (seconds), mean (SD)</td>
<td>14.7 (10.9)</td>
<td>N/A</td>
</tr>
<tr>
<td>Notification engagement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Notifications received, mean (SD)</td>
<td>22.0 (13.3)</td>
<td>22.6 (9.82)</td>
</tr>
<tr>
<td>Notifications opened, mean (SD)</td>
<td>6.58 (6.45)</td>
<td>8.69 (6.36)</td>
</tr>
<tr>
<td>Percentage of notifications opened, mean (SD)(^c)</td>
<td>34.3 (31.8)</td>
<td>39.9 (25.9)</td>
</tr>
<tr>
<td>None received, n (%)</td>
<td>2 (4)</td>
<td>1 (2)</td>
</tr>
</tbody>
</table>

\(^a\)Calculated as the number of weeks over the 12-week period with at least 3 screen view or user engagement metrics recorded, as per Google Analytics (Google LLC) data.

\(^b\)N/A: not applicable; participants in the control arm were unable to view the progress report.

\(^c\)Percentage of notifications opened based on the total notifications received for each participant.

Participants generally liked the new in-app components. They felt that the progress report could motivate task completion by providing clarity that previous tasks had been successfully logged:

> [It allowed me] certainly to feel more engaged and understand...how it contributes, but also to gain an overview of my own input into it, so not just I enter the data and it disappears. [P99]

Most participants found the notifications somewhat informative:

> The pop up things with little quotes about “doing this helps you”...yeah I liked those, I thought that was really good. You’re doing it for a purpose. [P29]

However, many participants were unsure whether they had seen all the notifications that their phone had received. Some highlighted the potentially demotivating effects of the progress report, depending on previous completion:

> I think it depends what mood you're in...if I hadn’t completed everything and I wasn’t in a good mood space I could be thinking “ooh I’ve failed.” [P40]

Whereas most participants agreed that the components might motivate others, the impact of the components on participants’ own task completion was more nuanced. Instead, participating in the research study seemed to be the strongest motivation for task completion:

> Because I had committed to do the study it meant that I said I am going to do it so I can’t be half-hearted about it...I want to do the best I could because it was for somebody else’s use.” [P99]

Many participants discussed the beneficial effects of taking part in symptom monitoring generally, such as increased awareness of their depression and communication with others. Several additional in-app components were suggested, including a direct communication channel between the app and research team.

**Harms and Protocol Violations**

No adverse or serious adverse events were reported. Among the 100 participants, 1 (1%) withdrew owing to technological issues.

**Discussion**

**Principal Findings**

This study conducted the first, fully remote RCT of the RADAR-base symptom-tracking system to test the effect of additional in-app components, based on behavioral change theory, on objective and subjective engagement. Overall, objective engagement was high across the sample. We found that the participants who received the adapted system (incorporating theoretically informed notifications, real-time progress reports, and researcher contact details) did not show higher levels of engagement with the system than the participants who received the system as usual. Although
subjective engagement (emotional self-awareness, system utility, and usability) was slightly higher in those who received the adapted app, the difference was small and did not reach statistical significance.

**Implications and Links With Previous Work**

Previous research, both usability studies [19,21,22] and RCTs [20,24], has suggested that providing notifications and progress visualization can prompt objective engagement in remote symptom monitoring. We propose several explanations as to why our results did not reflect past findings.

First, our findings may reflect the sample used. Participants were recruited from a previous study that used the RADAR-base system. This meant that they had prior experience of and interest in symptom monitoring. Previous work has also highlighted the impact of the academic setting on engagement through altruistic motivations [41]. It is possible that our results reflect a ceiling effect, whereby participants in both groups were motivated to participate in the research and complete symptom monitoring regardless of the changes to the app. This is particularly apparent given that 2 of the in-app components were designed to reflect individual achievement and benefits, aspects that might not have been as relevant in this research context.

Second, the combination of in-app components used in the adapted system might not have been sufficiently tailored to the user. The development of the app was grounded in both behavioral theory [31] and user involvement [21], which suggested that viewing real-time progress and being reminded of the proposed benefits of symptom monitoring might combat the barriers to engagement. However, although these components are proposed to encourage future tracking behavior, in practice, it is unclear how they interact with the motivation section of the COM-B model, in this case, the low motivation linked to low mood in our cohort of people with depression. Previous work has focused on symptom tracking for substance abuse [20] or general population [18,24] cohorts, both of which might react to incentivization in different ways from those with depression. Our qualitative discussions indeed suggested that the impact of viewing data progress might be affected by individual mood and motivational fluctuations. The addition of other virtual incentives, such as gamification [42], might have been more effective in promoting engagement with the tasks here, alongside the ability to personalize which components are seen and when they are seen.

Moreover, our components were static in that they were accessible to all the participants in the intervention group at the same time and frequency. Previous work has suggested that several factors can significantly moderate the relationship between in-app components and engagement. For example, Nahum-Shani and colleagues [20] found that receiving data insights only increased the likelihood of future self-reporting in those who were not frequent users of the app, suggesting that visualizing progress is not incentivizing (or is even perhaps “irritating”) for those who are actively engaged in the task from the beginning. Several studies have found a link between notification timing and engagement [20,24], although attempts at sensor-driven notification sending based on location have so far been unsuccessful in improving data availability [43]. Taken together, this suggests that future work is needed to understand the process of interacting with in-app components in this cohort.

Third, with regard to subjective engagement, the measures used in this study might not have reflected the experience of self-monitoring in the most nuanced manner. We used previous findings [6,41] to inform our operationalization of subjective engagement with RMTs as usability (UES), emotional self-awareness (ESQ), and utility (MAUQ). Our qualitative evaluation suggested that participants generally saw the in-app components as helpful in increasing task completion, which, in turn, might have promoted the feelings of emotional self-awareness they gained from monitoring their symptoms. We also saw that participants who viewed the progress report did so for around 15 seconds at a time, often repeatedly, which suggests a sustained interest in viewing progress. Although we did not see significant differences in either objective or subjective engagement, we did see slight treatment effects for all 3 subjective measures, which were higher still when adjusting for those who viewed the progress report. It is possible that different measures might have revealed a more significant change. For example, the UES is a tool designed primarily for digital health interventions and measures concepts such as focused attention, which are not as relevant to RMTs [6]. Measures tapping into other aspects of the experience of symptom monitoring, such as being seen as an individual [44] or the provision of a safety net [45], might provide a more detailed understanding of the interaction among the in-app components, objective engagement, and subjective engagement in the study; however, to our knowledge, these have yet to be developed.

**Strengths and Limitations**

To our knowledge, this was the first study to attempt to quantify the effect of in-app components on objective and subjective engagement with a multiparametric symptom-tracking system for depression. We used an established system that was previously used to conduct the largest, longitudinal study on RMT in MDD to date [12] and demonstrated the successful transference of the system to a remote RCT design. Methodologically, this study laid the foundation for future work to measure both objective and subjective engagements with symptom-tracking devices. We used an adapted system with in-app components, which allowed for an active control group (the system as usual) and embedded data collection to reduce confounding factors associated with the delivery medium [46]. In reference to our first aim, we have shown good data availability in the first fully remote trial of the system, with 87% (87/100) of the participants completing follow-up data collection, a median of 75% completion of symptom-tracking tasks, and a mean of 74 of 84 days of wear time data without planned researcher contact.

There are several limitations to this study. First, as mentioned, the sample was previously engaged in remote symptom tracking and driven by research altruism. This allowed for the recruitment of a large sample from an established group, obtaining results quickly and efficiently. However, it is unclear how far these results might generalize to community cohorts using symptom tracking in their daily life. Second, the study was conducted...
during the COVID-19 lockdown periods in the United Kingdom. A combination of increased free time and interest in health tracking could have resulted in increased engagement rates. Third, despite the large sample size, the study did not reach the intended number of participants needed to achieve the optimum statistical power. Fourth, although the app design was grounded in previous research, working within the confines of an established system gave way to certain design constraints. Some additional facilitators that arose from the COM-B analysis, such as the in-app reporting of technological malfunctions, could not be included or assessed in terms of their impact on engagement.

**Avenues for Future Work**

Future work should use these findings as a basis for further RCTs quantifying the effects of RMT system design on objective and subjective engagement with remote symptom tracking. Context-specific, dynamic tailoring of notifications and data insights could be key here. Although in-app components reduce the need for human resources, the impact of external factors should not be dismissed. Our system amendments did not promote engagement over and above the system as usual; future work could seek to understand how incentives such as research team support could interact with in-app components to increase engagement, such as the use of supportive chatbots [23]. Of major importance is replicating this work with different cohorts. Using the adapted system with non–help-seeking participants or those with lower technological literacy might affect the impact of the components that we tested. For example, the impact of the theoretically informed notifications might be greater in those who are less aware of the proposed benefits of symptom monitoring. Similarly, engagement with the app is likely to vary if the app is implemented in clinical practice; progress tracking and notification content might be more impactful for those who use the system for their own direct benefit. This work could also seek to complement the RCT design with additional analysis manuscripts for increased insight into the impact of UI features. For example, this could include correlational analyses of in-app component use with the measures of objective and subjective engagement or exploring whether baseline demographics are predictive of engagement in such trials. Another area for exploration is the measurement of the subjective experience of remote symptom tracking. The development of a suitable instrument that encapsulates experiential engagement would propel the understanding of the promotion of engagement across the field.

**Conclusions**

This study found that a combination of informative notifications, progress visualization, and research team contact details did not increase engagement in remote symptom tracking in our research cohort. However, the system provided good data availability, and the process evaluation measures suggested that participants saw benefits in using the adapted system. We have provided the methodology and scope for future exploration in this area, as well as opportunities to replicate this work in both community and clinical cohorts to further the promotion of engagement in remote health symptom tracking for both data collection and clinical management.

**Acknowledgments**

This paper represents independent research partly funded by the National Institute for Health and Care Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley National Health Service (NHS) Foundation Trust and King’s College London. The views expressed are those of the authors and not necessarily those of the NHS, NIHR, or Department of Health and Social Care.

The Remote Assessment of Disease and Relapse–Central Nervous System project received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement number 115902. This joint undertaking receives support from the European Union’s Horizon 2020 research and innovation program and European Federation of Pharmaceutical Industries and Associations (EFPIA). This communication reflects the views of the RADAR-CNS consortium, and neither Innovative Medicines Initiative nor the European Union and EFPIA are liable for any use that may be made of the information contained herein. The funding body was not involved in the design of the study or the collection, analysis, or interpretation of the data. This research was reviewed by a team with experience of mental health problems and their carers who were specially trained to advise on research proposals and documentation through the Feasibility and Acceptability Support Team for Researchers: a free, confidential service in England provided by the NIHR Maudsley Biomedical Research Centre via King’s College London and South London and Maudsley NHS Foundation Trust. Finally, the authors would like to thank all members of the RADAR-CNS patient advisory board who have experience of living with or supporting those who are living with depression, epilepsy, or multiple sclerosis.

**Authors’ Contributions**

KMW designed, coordinated, and conducted the study as well as led data processing, data coding, data analysis, and the writing of this manuscript. EC contributed to the design of the study and advised on data analysis. DL contributed to the design of the study and aided with data extraction. FM contributed to the design of the study. PC contributed to in-app component development and data extraction. YR contributed to app development and data extraction. SS contributed to service user involvement research and in-app component development. ED-L contributed to service user involvement research and in-app component development. LW contributed to the coding and thematic analysis of the interview data. CH contributed to the development and design of the study, reviewed data analysis, and supervised the first author. MH contributed to the development and design of the study, reviewed data analysis, and supervised the first author.
Conflicts of Interest

MH is the principal investigator of the RADAR-CNS consortium, a private-public precompetitive consortium with research funding from Janssen, Union Chimique Belge, Merck Sharp & Dohme, Biogen, and Lundbeck. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Speech task paragraph.

[DOCX File, 10 KB - mhealth_v12i1e44214_app1.docx]

Multimedia Appendix 2
Development of the adapted system based on the capability, opportunity, motivation, and behavior framework and service user research.

[DOCX File, 243 KB - mhealth_v12i1e44214_app2.docx]

Multimedia Appendix 3
Semistructured interview schedules.

[DOCX File, 21 KB - mhealth_v12i1e44214_app3.docx]

Multimedia Appendix 4
CONSORT checklist.

[PDF File (Adobe PDF File), 65 KB - mhealth_v12i1e44214_app4.pdf]

References


**Abbreviations**

- **CACE**: complier average causal effect
- **COM-B**: capability, opportunity, motivation, and behavior
- **ESQ**: Emotional Self-Awareness Questionnaire
- **MAUQ**: mHealth App Usability Questionnaire
- **MDD**: major depressive disorder
- **PHQ-8**: Patient Health Questionnaire-8
- **RADAR-MDD**: Remote Assessment of Disease and Relapse–Major Depressive Disorder
- **RCT**: randomized controlled trial
- **REDCap**: Research Electronic Data Capture
- **RMT**: remote measurement technology
- **RSES**: Rosenberg Self-Esteem Scale
- **UES**: User Engagement Scale
- **UI**: user interface
Edited by L Buis; submitted 30.11.22; peer-reviewed by VRD Herbuela, J Meyer, A Daros; comments to author 07.03.23; revised version received 21.05.23; accepted 09.06.23; published 19.01.24.

Please cite as:
JMIR Mhealth Uhealth 2024;12:e44214
URL: https://mhealth.jmir.org/2024/1/e44214
doi:10.2196/44214
PMID:38241070

©Katie M White, Ewan Carr, Daniel Leightley, Faith Matcham, Pauline Conde, Yatharth Ranjan, Sara Simblett, Erin Dawe-Lane, Laura Williams, Claire Henderson, Matthew Hotopf. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org/), 19.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Effectiveness of a Nurse-Led Mobile-Based Health Coaching Program for Patients With Prostate Cancer at High Risk of Metabolic Syndrome: Randomized Waitlist Controlled Trial

Kyoungjin Lee1,2, PhD; Jeongok Park3, PhD; Eui Geum Oh3, PhD; JuHee Lee3, PhD; Chang Park4, PhD; Young Deuk Choi5, PhD

1College of Nursing and Brain Korea 21 FOUR Project, Yonsei University, Seoul, Republic of Korea
2College of Nursing, Kyungbok University, Namyangju, Republic of Korea
3Mo-Im Kim Nursing Research Institute, College of Nursing, Yonsei University, Seoul, Republic of Korea
4Department of Population Health Nursing Science, College of Nursing, University of Illinois at Chicago, Chicago, IL, United States
5College of Medicine, Yonsei University, Seoul, Republic of Korea

Corresponding Author:
Jeongok Park, PhD
Mo-Im Kim Nursing Research Institute, College of Nursing
Yonsei University
Seodaemun-gu, Seoul, Korea
Seoul, 03722
Republic of Korea
Phone: 82 2 2228 3390
Fax: 82 2 2227 8303
Email: JOPARK02@yuhs.ac

Abstract

Background: Androgen deprivation therapy (ADT), a standard treatment for prostate cancer (PC), causes many physical side effects. In particular, it causes metabolic changes such as fasting glucose abnormalities or accumulation of body fat, and its continuation can lead to metabolic syndrome (MetS), which is closely related to diabetes and cardiovascular disease. Therefore, it is important to maintain and practice a healthy lifestyle in patients with PC.

Objective: This study aims to evaluate the effectiveness of a nurse-led mobile-based program that aims to promote a healthy lifestyle in patients with PC undergoing ADT with MetS risk factors.

Methods: This was a single-blind, randomized, waitlist control interventional study. A total of 48 patients were randomly assigned to the experimental and waitlist control groups at the urology cancer clinic of a tertiary general hospital in South Korea. The inclusion criteria were patients who had undergone ADT for >6 months, had at least 1 of the 5 MetS components in the abnormal range, and could access a mobile-based education program. The experimental group attended a 4-week mobile-based program on exercise and diet that included counseling and encouragement to maintain a healthy lifestyle, whereas the control group was placed on a waitlist and received usual care during the follow-up period, followed by the intervention. The primary outcome was a change in the healthy lifestyle score. The secondary outcomes were changes in 5 MetS components, body composition, and health-related quality of life. The outcomes were measured at 6 weeks and 12 weeks after the initiation of the intervention. Each participant was assigned to each group in a sequential order of enrollment in a 4×4 permuted block design randomization table generated in the SAS (SAS Institute) statistical program. A linear mixed model was used for statistical analysis.

Results: A total of 24 participants were randomly assigned to each group; however, 2 participants in the experimental group dropped out for personal reasons before starting the intervention. Finally, 46 participants were included in the intention-to-treat analysis. The experimental group showed more positive changes in the healthy lifestyle score (β=29.23; P≤.001), level of each MetS component (fasting blood sugar: β=−12.0; P=.05 and abdominal circumference: β=−2.49; P=.049), body composition (body weight: β=−1.52; P<.001 and BMI: β=−0.55; P<.001), and the urinary irritative and obstructive domain of health-related quality of life (β=14.63; P<.001) over time than the waitlist control group.
Conclusions: Lifestyle changes through nurse-led education can improve level of each MetS components, body composition, and ADT side effects. Nurses can induce positive changes in patients’ lifestyles and improve the self-management of patients starting ADT through this program.

Trial Registration: Clinical Research Information Service KCT0006560; http://tinyurl.com/yhvj4vwh

KEYWORDS
nurses; prostate neoplasms; healthy lifestyle; metabolic syndrome; exercise; diet; mobile phone

Introduction

Adverse Effects of Androgen Deprivation Therapy in Prostate Cancer

Prostate cancer (PC) is a commonly occurring cancer in men worldwide, and it is the fourth most common cancer among adult men in South Korea [1]. The incidence of PC in South Korea has increased rapidly in recent years, from 2.2% in 2000 to 14.3% in 2018 [1], owing to changes in dietary patterns and the development of new diagnostic technologies. PC is affected by androgenic activity in the body [2]. Therefore, the goal of treatment is to either remove androgens using castration or neutralize the effects of androgens. Androgen deprivation therapy (ADT), which disrupts the mechanisms that create androgens, is a standard treatment for PC [3]. Initially, ADT was used to treat patients with metastatic PC or older persons with cancer with limited treatment options. However, in recent years, there has been a gradual expansion in the use of ADT to include treating patients in the early stages of PC [4,5].

Repeated ADT can lead to climacteric symptoms in male individuals [6], including sexual dysfunction, hot flashes [7], gynecomastia [8], depression, fatigue, changes in sleep patterns [9], loss of muscle strength [10,11], osteoporosis [12], metabolic syndrome (MetS) [13], and cardiovascular disease [14,15]. These issues can reduce the quality of life of patients and may even be life threatening [16,17]. Specifically, evidence suggests that ADT leads to metabolic changes [13,18]. MetS, also called insulin resistance syndrome, is the clustering of several risk factors associated with obesity. It is closely related to diabetes [14,15,19] and has also been identified as a major cause of cardiovascular disease owing to its association with dyslipidemia, diabetes, and hypertension [14,15]. PC and MetS have a close relationship: patients with PC who received ADT were found to be 2.5 times more likely to have MetS than those who did not receive ADT [18]. A cross-sectional analysis using Korean National Health Insurance Service data found that the prevalence rate of MetS in patients with PC was 40.1%, which is higher than the 34.5% prevalence rate in older Korean male adults [20]. In addition, as age increases, MetS has been found to increase by 8 times in patients with PC [21]. Given that most patients with PC are older adults, ADT-induced metabolic problems coupled with aging can lead to deterioration of health, resulting in cardiovascular disease or death [22]. Despite the necessity of severe side effects management for ADT, >50% of patients with PC are unaware of these problems [23], and most health care providers tend to focus on cancer treatment. Moreover, no protocols have been developed to manage ADT-induced MetS in patients with PC [24,25].

Importance of Improving Self-Management Skill for ADT-Induced MetS

ADT-induced MetS causes metabolic changes that are different from those caused by classic MetS including changes in body composition [26]. Androgens are a group of male sex hormones, and androgen level reduction leads to an increase in body fat percentage and a decrease in fat-free mass, such as muscle loss and decreased bone density. Although classic MetS is characterized by an increase in visceral fat, there is insufficient evidence to determine whether or not this also occurs in patients with PC [27,28]. Implementing a classic MetS management program is not suitable for patients with PC because ADT-induced MetS has different characteristics than classic MetS, and each patient experiences different side effects, including physical and emotional problems. For example, 80% of patients with PC experience hot flashes and 12% to 14% experience breast tenderness and depression [29].

MetS is a representative chronic lifestyle disease that requires individual self-management. The World Health Organization emphasizes that lifestyle is the most critical factor affecting health status [30]. The current health care system appears to focus its attention and investments toward the discovery and treatment of the causes of disease rather than lifestyle. However, lifestyle modification is the most effective way to reduce the incidence of cardiovascular complications caused by MetS [31]. In general, lifestyle modifications, including exercise and a nutritional diet, are known to result in lower occurrence rates of MetS and a lower risk of cardiovascular disease [31]. Such lifestyle modifications require improving self-management skills. Self-management is a lifelong task that requires engagement in activities that promote good health [32]. Improving self-management allows patients to maintain active lives, leading to a better quality of life [32].

Although the optimal duration of ADT remains undefined, patients with PC typically receive ADT for 2 to 3 years [33]. Furthermore, as metabolic changes begin at least 3 months after ADT initiation, it is very important for patients with PCs to improve their self-management skills from the beginning of ADT. In addition, health care providers should be aware of ADT-related side effects from an early stage and mediate metabolic changes of PC through education. Therefore, newly developed health coaching programs for patients with PC should be configured differently from previous iterations [34-36]. In other words, the focus should be on improving self-management skills to encourage lifestyle changes that take into consideration the side effects of ADT from the beginning of ADT administration. However, effective self-management can be
challenging to maintain. In addition, the COVID-19 pandemic has diminished physical activity and nutritional quality worldwide [37]. The World Health Organization has emphasized the importance of exercise and maintaining a healthy lifestyle during the pandemic [38]. Web-based education has relatively few limitations in terms of time and space, and it has the advantage of being able to feature various types of media and teaching and learning materials compared with offline education programs [39].

Aims
This study aimed to evaluate the effectiveness of a nurse-led mobile-based health coaching program that promotes healthy lifestyle changes, normal range of MetS components, and health-related quality of life (HRQoL) for patients with PC receiving ADT. The primary aim was to identify the changes in (1) healthy lifestyle through this program, and the secondary aims were to identify the changes in (1) the levels of each MetS component, including blood pressure, fasting blood sugar (FBS), high-density lipoprotein (HDL) cholesterol, triglyceride, and abdominal circumference (AC); (2) body composition, such as body weight, BMI, skeletal muscle mass, fat mass, and fat percentage; and (3) quality of life.

Methods
Participants
The study population consisted of men who were diagnosed with PC at the urology cancer center of a single tertiary general hospital in South Korea. The inclusion criteria were (1) patients with PC who had been receiving ADT for >6 months at the time of enrollment in this study, (2) those with an abnormal range of at least 1 of the 5 MetS components, (3) those who were fully aware of the object and contents of the study and voluntarily participated, (4) those who understood spoken and written Korean and who could communicate without cognitive impairment, and (5) those who had a smartphone and were able to access the mobile-based education program. The exclusion criteria were (1) patients diagnosed with and treated for cancers other than PC; (2) those who had undergone surgery or chemotherapy for <3 months before the start of the study; (3) those who answered “yes” to ≥1 question of the Physical Activity Readiness Questionnaires; (4) those with cardiovascular diseases such as unstable angina pectoris, uncontrolled blood pressure, myocardial infarction, or comorbidities such as a musculoskeletal or nervous system disease; (5) those who were participating in other research programs; (6) those who had a change in medication to control their blood pressure, cholesterol, or blood sugar <3 months before the start of the study; and (7) those who had difficulty with typical daily activities.

Ethical Considerations
This study was approved by the Institutional Review Board of the Severance Hospital, Yonsei University Health System (4-2020-0889). We informed all participants of the purpose of this study, process, methods, voluntary participation terms, and the possible risks and benefits of participation. Furthermore, we gave the participants 1 week to consider their participation in the study before deciding whether to sign the informed consent form. All study participants were compensated with KRW 50,000 (approximately US $43) as a gift at both the beginning and end of the study.

Enrollment Process
To recruit participants, we posted recruitment announcements on the bulletin board of a urology cancer center. On the day of outpatient visits, we met face-to-face with patients who indicated an interest in participating in the study. We provided potential participants with the opportune time to ask questions. Then, we fully informed the potential participants about all facets of the study and invited them to voluntarily sign the informed consent form. Finally, we registered the participants after screening for the inclusion and exclusion criteria (Figure 1). The allocation process for this study was conducted jointly with a research assistant to ensure that all enrollments were transparent.
Using $F$ test in the G*Power program version 3.1 (Heinrich Heine University) [40], we calculated the sample size to identify the variance difference of repeated data between the 2 groups. The minimum number of participants was calculated to be 41 based on an effect size of 0.51, which was derived from the study by Bourke et al [41], with a significance level of .05, a power of 0.08, and 3 repeat measures. Considering a projected dropout rate of 15% [41], the total sample size was set at 48. Therefore, 24 participants were registered in each group. We assigned the participants randomly into experimental and waitlist control groups using a pregenerated sequence in SAS 9.4 (SAS Institute) based on the $2 \times 2$ permuted randomized block method. Using this method, 4 people are grouped as 1 block, and the 1 block is then divided into 2 groups. Then, we assigned the groups at a 1:1 ratio. The participants did not know which group they were assigned to. They received the intervention individually, which minimized the risk of contamination between groups.

**Research Design and Setting**

This was a single-blind, single-center randomized waitlist controlled trial. It used a pre- and posttest design and a $2 \times 2$ permuted randomization block method. We prepared an allocation table for each group before participants were enrolled. Each participant was assigned to an experimental or a waitlist control group in the order of enrollment using a $2 \times 2$ randomization table generated in the SAS program (SAS Institute). The protocol for this study was registered with the Clinical Research Information Service (registration no. KCT0006560). We recruited the participants from a tertiary general hospital located in Seoul, South Korea.

**Intervention**

To improve the positive health behaviors of patients with PC receiving ADT, a nurse-led mobile-based health coaching program based on the analyze, design, develop, implement, and evaluate model and the information-motivation-behavioral (IMB) skills model was developed. The analyze, design, develop, implement, and evaluate model, known as an instructional design model, is a representative framework used by teaching and learning methods. The IMB model has been widely used as a theoretical basis for developing interventions that aim to encourage patients with chronic diseases to maintain positive health behaviors. According to the IMB model, acquiring sufficient information, enhancing motivation, acquiring the skills required to perform a behavior, and improving self-efficacy are the factors that lead to long-term behavioral changes that improve subjective and objective health outcomes [42]. The nurse-led mobile-based health coaching program used an individualized approach to ensure that the IMB...
skills addressed in the program were best suited to each patient to improve their specific lifestyle patterns. In this program, self-management information included personalized recommendations for diet and exercise and strategies to manage both individual ADT-induced side effects and common side effects from other prescribed drugs. Motivation strategies included health contracting, individual goal setting, encouragement and continuous counseling, and rewarding. In addition, we encouraged the patients’ main caregivers to become involved in their patients’ behavioral self-management. Behavioral self-management skills included sharing vicarious experiences (eg, sharing success stories), self-monitoring through exercises, and maintaining a nutrition diary.

The program consisted of a 4-week intensive program conducted via one-on-one Zoom (Zoom Video Communication, Inc) meetings, followed by an 8-week maintenance program conducted via individual contact through telephone calls and KakaoTalk (Kakao Corporation), which is a popular Korean SMS text messaging app (Figure 2). Participants were provided with relevant information and a to-do list regarding exercise and diet once a week for 4 weeks in an intensive program. We provided an educational package that included a booklet containing PowerPoint (Microsoft Corporation) slides, exercise and nutritional diaries, exercise video files, TheraBand resistance bands (THERABAND), and a pedometer. We focused on overcoming barriers to healthy diet and exercise. For example, we recommended alternative exercises for participants with knee pain, which included a higher proportion of movements performed in a sitting position. Furthermore, the participants received feedback to address problems at the end of every web-based meeting during the intensive program and via SMS text messages and telephone calls during the maintenance program. When participants felt uncertain about an exercise motion, they recorded a video of themselves performing the exercise and sent it to us. We then provided feedback describing the correct motion, using the video as a reference. Regarding diet, we focused on changing unhealthy eating habits, controlling participants’ daily calorie intake, encouraging a low-fat and low-carbohydrate diet, and ensuring adequate protein intake. Participants were educated on the proportions of food from each food group they required per day according to their individual daily recommended calories and the diabetic exchange diet. Feedback and questions related to the diet program were addressed via KakaoTalk. We recommended implementing the following strategies for maintaining healthy behaviors and enhancing self-efficacy once a week or more during the maintenance period: vicarious experiences (eg, sharing success stories), emotional support, encouragement, keeping an exercise and nutrition diary, and goal reminders. To minimize the expected bias that can occur in an intervention study, we conducted the intervention according to the study protocol and used a checklist to ensure consistency between the experimental and waitlist control groups. The control group was placed on a waitlist and only received usual care during the intervention period. After completing data collection, we provided them with the same mobile-based program and educational materials as the experimental group (time point 3; T3). A group of experts evaluated this program to confirm its content and construct validity.

Figure 2. Educational material: (A) education on information and management of metabolic syndrome via the internet using Zoom (Zoom Video Communications, Inc), (B) personalized diet coaching, and (C) personalized exercise coaching: providing feedback describing the correct motion.
Data Collection

Data collection happened at 3 different time points: baseline (time point 1; T1), 6 weeks after the beginning of the intervention (time point 2; T2), and 12 weeks after the beginning of the intervention (T3). We collected T1 and T3 data on a face-to-face basis on the day the patient visited the urology cancer center. Variables that required physical contact, such as AC and grip strength, were measured together using a web-based survey at T1 and T3 but not at T2. At T2, we collected data collection on variables, including lifestyle score and HRQoL, on a non–face-to-face basis using a web-based survey. Clinical data that were obtained included treatment type, ADT exposure time, cancer stage, biochemical data, and Gleason score from electronic medical records (EMRs) at T1.

Outcomes

Lifestyle Score

Lifestyle score was the primary outcome of this study. Lifestyle changes were evaluated using the lifestyle evaluation tool by Kang [43] to assess the health behaviors of patients with MetS. A higher score indicates better self-management behaviors. This assessment consists of 36 items across 6 domains: physical activity and weight control, dietary habits, drinking and smoking, sleep and rest, stress, and drug and health management [43].

We recorded responses on a 4-point scale with the options not at all, sometimes, often, and always. Total scores could range from 36 to 144 points, with a higher score indicating a healthier lifestyle. The Cronbach α of the original study was .92 (physical activity and weight control=.87, dietary habits=.87, drinking and smoking=.87, sleep and rest=.86, stress=.74, and drug and health management=.70). In this study, Cronbach α was .82 (physical activity and weight control=.90, dietary habits=.86, drinking and smoking=.41, sleep and rest=.67, stress=.63, and drug and health management=.54).

MetS Components

The components of MetS consist of FBS, AC, blood pressure (both systolic and diastolic), fasting triglyceride level, and fasting HDL cholesterol. We instructed the participants not to have a meal and not to take antihypertensive drugs on the test day. Then, at the urology cancer center, we first measured AC using a tape at the umbilicus between the highest point of the iliac crest and the lower edge of the 12th rib with an error range of 0.5 cm while the participants fasted. Second, we measured the blood pressure twice on the participant’s nondominant arm using a TM-2657P device (A&D Company Limited) after they had rested for at least 10 minutes. The average systolic blood pressure (SBP) and diastolic blood pressure (DBP) were calculated.

Biochemical Data

Biochemical data included the levels of the following MetS components: fasting triglyceride, fasting HDL cholesterol, and FBS. These variables were measured using blood samples collected while participants had fasted for >6 hours, and the data were obtained from EMRs.

Body Composition

Body composition data were obtained using a body composition analyzer (InBody H20B [Biospace]). We instructed the participants to stand upright and hold the handle attached to the measurement device, which put them into contact with 8 electrodes (2 each on both hands and both feet). Body composition measurements included height (cm), body weight (kg), body fat mass (kg), body fat percentage (%), skeletal muscle mass (kg), and BMI (kg/m²).

HRQoL Measurement Tool

We measured HRQoL using the Korean version of the 26-item Expanded Prostate Cancer Index Composite (EPIC-26). EPIC-26 is a short-form version of the original expanded PC index composite (EPIC) tool, which contains 50 items. The EPIC tool was developed to understand treatment-related symptoms with a higher degree of sensitivity than previous diagnostic tools and the impact of PC treatment on patients’ lives [44]. Higher scores indicated a better HRQoL, with possible scores ranging from 0 to 100. EPIC-26 consists of 5 symptom domains: urinary incontinence, urinary irritation and obstruction, sexual, bowel, and hormonal. There is no Korean version of EPIC-26, but there is a Korean version of the original 50-item EPIC tool. Therefore, the 26 items from EPIC-26 were extracted from the Korean version of the original 50-item EPIC tool [45], and the survey was conducted using this tool. Permission to use both EPIC-26 and the original 50-item EPIC was granted by the original author.

The Cronbach α of EPIC-26 ranged from .70 to .90 in all domains except for the hormonal domain (Cronbach α=.62). The Cronbach α of the Korean version of EPIC was .83 [45]. The Cronbach α of the tool used in this study was .63 (urinary incontinence=.88, urinary irritation and obstruction=.64, sexual=.84, bowel=.13, and vitality or hormonal=.46).

Clinical Data

Disease-related patient information, treatment type, ADT exposure time, cancer stage, and Gleason score were obtained from EMRs.

Data Analysis

Demographic data, disease-related characteristics, and main outcome variables were analyzed using mean, SD, frequency, and percentage. An independent 1-tailed t test and a chi-square test were performed to identify differences between the groups. The equality of variance was assessed before using the pooled variance estimator for the t test. Fisher exact test was performed as appropriate. According to the International Conference on Harmonization E9 guideline [46], which provides guidance on statistical principles for clinical trials, a modified intention-to-treat analysis was conducted. As 2 participants withdrew from the study for personal reasons before initiating the intervention, there was a lack of data that could evaluate the effect of the main outcomes. We judged that there would be no difference in the intention-to-treat analysis owing to the low dropout rate (2/48, 4%) and high compliance rate (22/22, 81%). Hypothesis testing was conducted using a 1-tailed test and the PROC MIXED procedure in SAS (SAS Institute). SE estimates
were obtained as a result of the PROC MIXED procedure using the empirical option to adjust for skewed data from potentially different covariance structures. This method is based on the sandwich estimation approach [47]. It improves variance and covariance with robust and consistent estimates, irrespective of the covariance structure in the actual data. As a follow-up analysis, we determined statistically significant time points within the groups by calculating the difference in the least square means from the baseline at each time point.

**Results**

**Overview**

The data for this study were collected across 7 months, from March 24 to September 15, 2021. We took approximately 30 minutes per participant to complete data collection. A total of 48 participants were recruited. Two participants in the experimental group declined to participate in the program before starting the intervention, citing personal reasons. A total of 46 participants were finally included in the analysis, with an attrition rate of 4% (2/48).

**Results of General and Disease-Related Characteristics**

The general and disease-related characteristics are presented in Table 1. The mean age of the participants was 68.83 (SD 7.09) years. Most participants (41/46, 89%) lived with their spouses or families. Approximately 46% (21/46) of the participants were unemployed, and 35% (16/46) had jobs requiring relatively less physical activity, such as office workers, taxi drivers, and service workers. Most participants (27/46, 59%) were exsmokers, and the mean smoking duration was 16.13 (SD 21.38) pack-years.
<table>
<thead>
<tr>
<th>Variables</th>
<th>Total</th>
<th>Experimental group (n=22)</th>
<th>Control group (n=22)</th>
<th>$\chi^2$ (df)</th>
<th>t (df)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (y), mean (SD)</td>
<td>68.83 (7.09)</td>
<td>67.59 (6.59)</td>
<td>69.96 (7.48)</td>
<td>N/A</td>
<td>−1.13 (44)</td>
<td>.26</td>
</tr>
<tr>
<td>Monthly income (US $), mean (SD)</td>
<td>3679.77 (4603.72)</td>
<td>4813.0 (6188.7)</td>
<td>2640.9 (2035.5)</td>
<td>N/A</td>
<td>1.57 (44)</td>
<td>.13</td>
</tr>
<tr>
<td>Religion, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.99</td>
</tr>
<tr>
<td>No</td>
<td>21 (46)</td>
<td>10 (22)</td>
<td>11 (24)</td>
<td>0.0 (1)</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>25 (54)</td>
<td>12 (26)</td>
<td>13 (28)</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Education, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.40</td>
</tr>
<tr>
<td>Less than or equal to middle</td>
<td>27 (59)</td>
<td>11 (24)</td>
<td>16 (35)</td>
<td>0.7 (1)</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>school</td>
<td>Greater than or equal to college</td>
<td>19 (41)</td>
<td>11 (24)</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Living, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.77b</td>
</tr>
<tr>
<td>Alone</td>
<td>5 (11)</td>
<td>3 (7)</td>
<td>2 (4)</td>
<td>0.5 (2)</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>With spouse only</td>
<td>21 (46)</td>
<td>9 (20)</td>
<td>12 (26)</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>With their family</td>
<td>20 (43)</td>
<td>10 (22)</td>
<td>10 (22)</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Job-related physical activities, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.93b</td>
</tr>
<tr>
<td>Unemployed</td>
<td>21 (46)</td>
<td>10 (22)</td>
<td>11 (24)</td>
<td>0.3 (2)</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Less active</td>
<td>16 (35)</td>
<td>7 (15)</td>
<td>9 (35)</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Highly active</td>
<td>9 (20)</td>
<td>5 (11)</td>
<td>4 (9)</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Smoking history</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 pack-year, mean (SD)</td>
<td>16.13 (21.38)</td>
<td>15.78 (21.58)</td>
<td>16.44 (21.65)</td>
<td>N/A</td>
<td>−0.1 (44)</td>
<td>.92</td>
</tr>
<tr>
<td>Nonsmoker, n (%)</td>
<td>16 (35)</td>
<td>6 (13)</td>
<td>10 (22)</td>
<td>1.6 (2)</td>
<td>N/A</td>
<td>.45b</td>
</tr>
<tr>
<td>Exsmoker, n (%)</td>
<td>27 (59)</td>
<td>15 (33)</td>
<td>12 (26)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Current smoker, n (%)</td>
<td>3 (7)</td>
<td>1 (2)</td>
<td>2 (4)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Disease characteristics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of comorbidities, mean (SD)</td>
<td>1.61 (1.04)</td>
<td>1.64 (1.05)</td>
<td>1.58 (2.03)</td>
<td>N/A</td>
<td>0.17 (44)</td>
<td>.87</td>
</tr>
<tr>
<td>0, n (%)</td>
<td>6 (13)</td>
<td>3 (7)</td>
<td>3 (7)</td>
<td>0.8 (3)</td>
<td>N/A</td>
<td>.87b</td>
</tr>
<tr>
<td>1, n (%)</td>
<td>17 (37)</td>
<td>7 (15)</td>
<td>10 (22)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>2, n (%)</td>
<td>14 (30)</td>
<td>8 (17)</td>
<td>6 (13)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>≥3, n (%)</td>
<td>9 (20)</td>
<td>4 (9)</td>
<td>5 (11)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Treatment type, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.10</td>
</tr>
<tr>
<td>Operation</td>
<td>16 (35)</td>
<td>5 (11)</td>
<td>11 (24)</td>
<td>2.7 (1)</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Operation+radiation</td>
<td>30 (65)</td>
<td>17 (37)</td>
<td>13 (28)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>ADT</strong> type, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antiandrogen</td>
<td>30 (65)</td>
<td>14 (30)</td>
<td>16 (35)</td>
<td>0.0 (1)</td>
<td>N/A</td>
<td>.99</td>
</tr>
<tr>
<td>Antiandrogen+LHRHd</td>
<td>16 (35)</td>
<td>8 (17)</td>
<td>8 (17)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>ADT duration (month), mean (SD)</td>
<td>40.63 (24.71)</td>
<td>40.64 (22.61)</td>
<td>40.63 (26.98)</td>
<td>N/A</td>
<td>0.0 (44)</td>
<td>.99</td>
</tr>
<tr>
<td>PSA, mean (SD)</td>
<td>0.07 (0.19)</td>
<td>0.09 (0.27)</td>
<td>0.05 (0.08)</td>
<td>N/A</td>
<td>0.59 (44)</td>
<td>.56</td>
</tr>
<tr>
<td>Gleason score, mean (SD)</td>
<td>7.80 (0.98)</td>
<td>8.00 (1.07)</td>
<td>7.63 (0.88)</td>
<td>N/A</td>
<td>1.31 (44)</td>
<td>.20</td>
</tr>
</tbody>
</table>

Main outcomes
<table>
<thead>
<tr>
<th>Variables</th>
<th>Total (n=44)</th>
<th>Experimental group (n=22)</th>
<th>Control group (n=22)</th>
<th>$\chi^2$ (df)</th>
<th>$t$ (df)</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lifestyle score</strong>, mean (SD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total score</td>
<td>99.07 (13.56)</td>
<td>100.50 (12.53)</td>
<td>97.71 (14.57)</td>
<td>N/A</td>
<td>0.70</td>
<td>.48</td>
</tr>
<tr>
<td>Exercise and weight loss</td>
<td>19.67 (5.88)</td>
<td>20.23 (5.45)</td>
<td>19.17 (6.32)</td>
<td>N/A</td>
<td>0.61</td>
<td>.55</td>
</tr>
<tr>
<td>Diet</td>
<td>42.17 (6.72)</td>
<td>42.77 (6.05)</td>
<td>41.63 (7.38)</td>
<td>N/A</td>
<td>0.57</td>
<td>.57</td>
</tr>
<tr>
<td>Alcohol and smoking</td>
<td>9.54 (3.40)</td>
<td>10.09 (3.25)</td>
<td>9.04 (3.53)</td>
<td>N/A</td>
<td>1.05</td>
<td>.30</td>
</tr>
<tr>
<td>Stress management</td>
<td>9.82 (1.79)</td>
<td>9.82 (1.79)</td>
<td>9.33 (2.41)</td>
<td>N/A</td>
<td>0.77</td>
<td>.45</td>
</tr>
<tr>
<td>Sleep and rest</td>
<td>6.45 (1.60)</td>
<td>6.45 (1.60)</td>
<td>6.71 (1.76)</td>
<td>N/A</td>
<td>-0.51</td>
<td>.61</td>
</tr>
<tr>
<td>Medication adherence and physical examination</td>
<td>11.52 (2.04)</td>
<td>11.18 (1.84)</td>
<td>11.83 (2.20)</td>
<td>N/A</td>
<td>-1.08</td>
<td>.28</td>
</tr>
<tr>
<td><strong>MetS</strong>, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>21 (46)</td>
<td>10 (22)</td>
<td>11 (24)</td>
<td>0.0 (1)</td>
<td>N/A</td>
<td>.98</td>
</tr>
<tr>
<td>Yes</td>
<td>25 (54)</td>
<td>12 (26)</td>
<td>13 (28)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>MetS component, mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBP(^b) (mm/Hg)</td>
<td>136.42 (13.01)</td>
<td>139.7 (11.44)</td>
<td>133.4 (13.83)</td>
<td>N/A</td>
<td>1.69</td>
<td>.10</td>
</tr>
<tr>
<td>DBP(^b) (mm/Hg)</td>
<td>83.51 (7.46)</td>
<td>85.18 (6.11)</td>
<td>81.98 (8.35)</td>
<td>N/A</td>
<td>1.47</td>
<td>.15</td>
</tr>
<tr>
<td>AC(^b) (cm)</td>
<td>95.73 (5.97)</td>
<td>94.30 (5.29)</td>
<td>97.05 (6.36)</td>
<td>N/A</td>
<td>-1.59</td>
<td>.12</td>
</tr>
<tr>
<td>FBS(^k) (mg/dl)</td>
<td>111.17 (22.22)</td>
<td>112.6 (27.64)</td>
<td>109.8 (16.27)</td>
<td>N/A</td>
<td>0.41</td>
<td>.68</td>
</tr>
<tr>
<td>HDL(^l) (mg/dl)</td>
<td>52.89 (13.61)</td>
<td>55.50 (13.32)</td>
<td>50.50 (13.71)</td>
<td>N/A</td>
<td>1.25</td>
<td>.22</td>
</tr>
<tr>
<td>Triglyceride (mg/dl)</td>
<td>114.46 (49.05)</td>
<td>116.5 (44.45)</td>
<td>112.5 (53.82)</td>
<td>N/A</td>
<td>0.27</td>
<td>.79</td>
</tr>
<tr>
<td><strong>Body composition, mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body weight (kg)</td>
<td>73.12 (5.99)</td>
<td>72.53 (4.54)</td>
<td>73.66 (7.13)</td>
<td>N/A</td>
<td>-0.64</td>
<td>.52</td>
</tr>
<tr>
<td>Body fat mass (kg)</td>
<td>22.18 (4.59)</td>
<td>21.75 (3.51)</td>
<td>22.57 (5.44)</td>
<td>N/A</td>
<td>-0.61</td>
<td>.54</td>
</tr>
<tr>
<td>Body fat percent (%)</td>
<td>30.38 (5.31)</td>
<td>30.50 (4.74)</td>
<td>30.28 (5.88)</td>
<td>N/A</td>
<td>0.14</td>
<td>.89</td>
</tr>
<tr>
<td>Skeletal muscle mass (kg)</td>
<td>28.22 (3.08)</td>
<td>28.03 (2.96)</td>
<td>28.39 (3.23)</td>
<td>N/A</td>
<td>-0.39</td>
<td>.70</td>
</tr>
<tr>
<td>BMI (kg/m(^2))</td>
<td>25.93 (2.04)</td>
<td>25.65 (1.70)</td>
<td>26.19 (2.32)</td>
<td>N/A</td>
<td>-0.89</td>
<td>.38</td>
</tr>
<tr>
<td><strong>HRQoL(^m) domains, mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urinary incontinence</td>
<td>67.61 (25.99)</td>
<td>69.17 (26.56)</td>
<td>66.17 (25.93)</td>
<td>N/A</td>
<td>0.39</td>
<td>.70</td>
</tr>
<tr>
<td>Urinary irritation and obstruction</td>
<td>88.72 (10.59)</td>
<td>87.78 (10.11)</td>
<td>89.58 (11.16)</td>
<td>N/A</td>
<td>-0.57</td>
<td>.57</td>
</tr>
<tr>
<td>Bowel problem</td>
<td>95.11 (9.31)</td>
<td>95.83 (8.33)</td>
<td>94.44 (10.26)</td>
<td>N/A</td>
<td>0.50</td>
<td>.62</td>
</tr>
<tr>
<td>Sexual problem</td>
<td>19.28 (19.54)</td>
<td>21.89 (19.18)</td>
<td>16.89 (19.97)</td>
<td>N/A</td>
<td>0.86</td>
<td>.39</td>
</tr>
</tbody>
</table>
The average number of comorbidities was 1.61 (SD 1.04), and the most common comorbid diseases were cardiovascular diseases such as hypertension or dyslipidemia (31/46, 67%). There was no difference in comorbidities between the experimental and waitlist control groups ($P$=.87). Of the 65 participants, 30 (65%) received antiandrogen monotherapy, and the rest received antiandrogen therapy with luteinizing hormone–releasing hormone. The average duration of ADT was 40.63 (SD 24.71) months, the mean prostate-specific antigen level was 0.07 (SD 0.19) ng/ml, and the mean Gleason score, which determines the prognosis and pathological status of cancer, was 7.80 (SD 0.98). The experimental and waitlist control groups showed no statistical differences between general and disease-related characteristics. The mean healthy lifestyle score was 99.07 (SD 13.56). The MetS components with mean values in an abnormal range were SBP (mean 136.42, SD 13.01 mm Hg), AC (mean 95.73, SD 5.97 cm), and FBS (mean 111.17, SD 22.22 mg/dL). There were no statistical differences in MetS components between the groups. The mean body weight and BMI were approximately 73 (SD 5.99) kg and 26 (SD 2.04) kg/m$^2$, respectively. The participants responded that they had problems in the urinary incontinence (mean 67.61, SD 25.99) and sexual (mean 19.28, SD 19.54) domains of HRQoL. The experimental and waitlist control groups were homogeneous in terms of their main outcome variables.

### Results of Primary Outcome Variables Between Groups Over Time

Table 2 and Figure 3 show the results of the lifestyle score variables between the groups over time. The study found that there were no group differences in lifestyle scores at baseline (Table 1). However, over time, the experimental group’s lifestyle scores consistently increased ($T_1=100.55$, $T_2=125.82$, and $T_3=130.27$), whereas the waitlist control group’s lifestyle scores showed no consistent increase ($T_1=97.71$, $T_2=95.92$, and $T_3=98.21$). Over time, the lifestyle scores of the experimental group significantly increased ($\beta=29.23; P\leq.001$) compared with the waitlist control group. The experimental group developed a healthier lifestyle with time, and remarkable changes were observed during the intervention period.
Table 2. Group comparison in lifestyle score, MetS components, body composition, and HRQoL domain parameters (N=46).

<table>
<thead>
<tr>
<th>Outcome and group</th>
<th>T1c, mean (SD)</th>
<th>T2d, mean (SD)</th>
<th>T3e, mean (SD)</th>
<th>Difference of changes between groups over timef</th>
<th>Estimate (SE; 95% CI)</th>
<th>F test (df)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lifestyle score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>100.55 (12.53)</td>
<td>125.82 (11.73)</td>
<td>130.27 (10.15)</td>
<td>29.23 (3.50; 22.26 to 36.19)</td>
<td>38.49 (88)</td>
<td>≤.001</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>97.71 (14.57)</td>
<td>95.92 (18.64)</td>
<td>98.21 (15.07)</td>
<td>29.23 (3.50; 22.26 to 36.19)</td>
<td>38.49 (88)</td>
<td>≤.001</td>
<td></td>
</tr>
<tr>
<td><strong>MetS components</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>139.73 (11.44)</td>
<td>N/Ah</td>
<td>126.25 (14.08)</td>
<td>−5.64 (3.72; −13.12 to 1.85)</td>
<td>2.30 (44)</td>
<td>.14</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>133.38 (13.83)</td>
<td>N/Ah</td>
<td>125.54 (12.58)</td>
<td>−5.64 (3.72; −13.12 to 1.85)</td>
<td>2.30 (44)</td>
<td>.14</td>
<td></td>
</tr>
<tr>
<td>DBP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>85.18 (6.11)</td>
<td>N/Ah</td>
<td>79.36 (9.35)</td>
<td>−0.94 (2.21; −5.40 to 3.52)</td>
<td>0.18 (44)</td>
<td>.67</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>81.98 (8.35)</td>
<td>N/Ah</td>
<td>77.10 (7.36)</td>
<td>−0.94 (2.21; −5.40 to 3.52)</td>
<td>0.18 (44)</td>
<td>.67</td>
<td></td>
</tr>
<tr>
<td>FBS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>112.64 (27.64)</td>
<td>N/Ah</td>
<td>102.59 (12.42)</td>
<td>−12.0 (6.02; −24.14 to 0.13)</td>
<td>3.98 (44)</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>109.83 (16.27)</td>
<td>N/Ah</td>
<td>111.79 (16.88)</td>
<td>−12.0 (6.02; −24.14 to 0.13)</td>
<td>3.98 (44)</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>HDLk cholesterol</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>55.50 (13.33)</td>
<td>N/Ah</td>
<td>55.32 (12.46)</td>
<td>−0.27 (1.48; −3.25 to 2.72)</td>
<td>0.03 (44)</td>
<td>.86</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>50.50 (13.71)</td>
<td>N/Ah</td>
<td>50.58 (12.34)</td>
<td>−0.27 (1.48; −3.25 to 2.72)</td>
<td>0.03 (44)</td>
<td>.86</td>
<td></td>
</tr>
<tr>
<td>Triglyceride</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>116.55 (44.45)</td>
<td>N/Ah</td>
<td>107.73 (43.19)</td>
<td>−11.65 (12.04; −35.91 to 12.61)</td>
<td>0.94 (44)</td>
<td>.34</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>112.54 (53.82)</td>
<td>N/Ah</td>
<td>115.38 (62.71)</td>
<td>−11.65 (12.04; −35.91 to 12.61)</td>
<td>0.94 (44)</td>
<td>.34</td>
<td></td>
</tr>
<tr>
<td>ACd</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>94.30 (5.29)</td>
<td>N/Ah</td>
<td>91.02 (3.94)</td>
<td>−2.49 (1.23; −4.98 to −0.01)</td>
<td>4.09 (44)</td>
<td>.049</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>97.05 (6.36)</td>
<td>N/Ah</td>
<td>96.27 (6.83)</td>
<td>−2.49 (1.23; −4.98 to −0.01)</td>
<td>4.09 (44)</td>
<td>.049</td>
<td></td>
</tr>
<tr>
<td><strong>Body composition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Body weight (kg)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>72.53 (4.54)</td>
<td>N/Ah</td>
<td>70.04 (4.53)</td>
<td>−1.52 (0.46; −2.45 to −0.58)</td>
<td>10.71 (44)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>73.66 (7.13)</td>
<td>N/Ah</td>
<td>72.68 (7.04)</td>
<td>−1.52 (0.46; −2.45 to −0.58)</td>
<td>10.71 (44)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td><strong>BMI (kg/m²)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>25.65 (1.70)</td>
<td>N/Ah</td>
<td>24.76 (1.49)</td>
<td>−0.55 (0.17; −0.88 to −0.21)</td>
<td>10.54 (44)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>26.19 (2.32)</td>
<td>N/Ah</td>
<td>25.84 (2.27)</td>
<td>−0.55 (0.17; −0.88 to −0.21)</td>
<td>10.54 (44)</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td><strong>Skeletal muscle mass (kg)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>28.03 (2.96)</td>
<td>N/Ah</td>
<td>27.84 (4.19)</td>
<td>0.53 (0.83; −1.15 to 2.20)</td>
<td>0.40 (44)</td>
<td>.53</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>28.39 (3.23)</td>
<td>N/Ah</td>
<td>27.67 (2.84)</td>
<td>0.53 (0.83; −1.15 to 2.20)</td>
<td>0.40 (44)</td>
<td>.53</td>
<td></td>
</tr>
<tr>
<td><strong>Fat mass (kg)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>21.75 (3.51)</td>
<td>N/Ah</td>
<td>19.76 (5.01)</td>
<td>−1.93 (1.35; −4.65 to 0.78)</td>
<td>2.06 (44)</td>
<td>.16</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>22.57 (5.44)</td>
<td>N/Ah</td>
<td>22.52 (4.66)</td>
<td>−1.93 (1.35; −4.65 to 0.78)</td>
<td>2.06 (44)</td>
<td>.16</td>
<td></td>
</tr>
<tr>
<td><strong>Fat percentage</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>30.50 (4.74)</td>
<td>N/Ah</td>
<td>28.32 (7.27)</td>
<td>−2.75 (1.90; −6.58 to 1.09)</td>
<td>2.08 (44)</td>
<td>.16</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>30.28 (5.88)</td>
<td>N/Ah</td>
<td>30.84 (4.61)</td>
<td>−2.75 (1.90; −6.58 to 1.09)</td>
<td>2.08 (44)</td>
<td>.16</td>
<td></td>
</tr>
<tr>
<td><strong>HRQoL domains</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Outcome and group

<table>
<thead>
<tr>
<th>Outcome and group</th>
<th>T1a, mean (SD)</th>
<th>T2b, mean (SD)</th>
<th>T3c, mean (SD)</th>
<th>Difference of changes between groups over timef</th>
<th>Estimate (SE; 95% CI)</th>
<th>F test (df)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urinary incontinence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>69.17 (26.56)</td>
<td>81.19 (18.87)</td>
<td>83.38 (18.67)</td>
<td></td>
<td>7.70 (6.55; −5.31 to 20.72)</td>
<td>0.70 (88)</td>
<td>.50</td>
</tr>
<tr>
<td>Control</td>
<td>66.18 (25.93)</td>
<td>72.29 (23.00)</td>
<td>72.68 (29.53)</td>
<td></td>
<td>7.70 (6.55; −5.31 to 20.72)</td>
<td>0.70 (88)</td>
<td>.50</td>
</tr>
<tr>
<td>Urinary irritation and obstruction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>87.78 (10.11)</td>
<td>94.60 (7.54)</td>
<td>97.73 (4.93)</td>
<td></td>
<td>14.63 (4.05; 6.57 to 22.69)</td>
<td>7.01 (88)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Control</td>
<td>89.58 (11.16)</td>
<td>89.32 (11.43)</td>
<td>84.90 (18.14)</td>
<td></td>
<td>14.63 (4.05; 6.57 to 22.69)</td>
<td>7.01 (88)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Bowel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>95.83 (8.33)</td>
<td>95.64 (7.21)</td>
<td>98.48 (4.37)</td>
<td></td>
<td>2.30 (2.98; −3.61 to 8.22)</td>
<td>0.34 (88)</td>
<td>.71</td>
</tr>
<tr>
<td>Control</td>
<td>94.44 (10.26)</td>
<td>91.67 (14.33)</td>
<td>94.79 (10.15)</td>
<td></td>
<td>2.30 (2.98; −3.61 to 8.22)</td>
<td>0.34 (88)</td>
<td>.71</td>
</tr>
<tr>
<td>Sexual</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>21.89 (19.18)</td>
<td>18.86 (16.02)</td>
<td>25.49 (16.71)</td>
<td></td>
<td>1.52 (4.50; −7.43 to 10.47)</td>
<td>0.08 (88)</td>
<td>.92</td>
</tr>
<tr>
<td>Control</td>
<td>16.89 (19.97)</td>
<td>12.15 (12.18)</td>
<td>18.97 (15.03)</td>
<td></td>
<td>1.52 (4.50; −7.43 to 10.47)</td>
<td>0.08 (88)</td>
<td>.92</td>
</tr>
<tr>
<td>Hormonal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimental</td>
<td>85.00 (11.65)</td>
<td>89.32 (12.94)</td>
<td>92.73 (7.03)</td>
<td></td>
<td>6.27 (3.48; −0.64 to 13.18)</td>
<td>1.85 (88)</td>
<td>.16</td>
</tr>
<tr>
<td>Control</td>
<td>85.00 (15.88)</td>
<td>86.88 (12.58)</td>
<td>86.46 (13.87)</td>
<td></td>
<td>6.27 (3.48; −0.64 to 13.18)</td>
<td>1.85 (88)</td>
<td>.16</td>
</tr>
</tbody>
</table>

aMetS: metabolic syndrome.
bHRQoL: health-related quality of life.
cT1: time point 1.
dT2: time point 2.
eT3: time point 3.
fReference: interaction between control group and T1.
gSBP: systolic blood pressure.
hN/A: not applicable.
iDBP: diastolic blood pressure.
jFBS: fasting blood sugar.
kHDL cholesterol: high-density lipoprotein cholesterol.
lAC: abdominal circumference.
Results of Secondary Outcome Variables Between Groups Over Time

The difference in the prevalence of MetS before and after the intervention between the experimental group and the waitlist control group was not statistically significant ($\chi^2 = 1.1; P = .31$ at T3 [not presented in the tables]). Among the MetS components, the parameters for FBS ($\beta = -12.0; F_{44} = 3.98; P = .05$) and AC ($\beta = -2.49; F_{44} = 4.09; P = .049$) showed significant interactions between group and time (Table 2 and Figures 4 and 5). Regarding body composition, the mean body weight and BMI in the experimental group decreased significantly by 1.52 kg and 0.55 kg/m$^2$, respectively ($P < .001$) compared with the baseline values. Group, time, and group and time interactions were also statistically significant between these variables ($P < .001$). Over time, the mean body weight ($P < .001$) and BMI ($P < .001$) decreased more in the experimental group than in the waitlist control group (Table 2 and Figures 6 and 7). Regarding HRQoL domains (Table 2 and Figure 8), a more significant improvement was observed in the experimental group than in the waitlist control group for the urinary irritative and obstructive domain. The mean changes in the urinary irritative and obstructive domain of HRQoL were statistically significant at each time point from the baseline, whereas the mean changes in the control group for this domain were not statistically significant. Group and time interactions were not significant, except in the urinary irritative and obstructive domain of HRQoL in the experimental group ($\beta = 14.63; F_{8,8} = 7.01; P < .001$).
Figure 4. Group comparison among the main outcome variables (fasting blood sugar) over time. T1: time point 1; T3: time point 3.

Figure 5. Group comparison among the main outcome variables (abdominal circumference) over time. T1: time point 1; T3: time point 3.
Figure 6. Group comparison among the main outcome variables (body weight) over time. T1: time point 1; T3: time point 3.

Figure 7. Group comparison among the main outcome variables (BMI) over time. T1: time point 1; T3: time point 3.
Discussion

Principal Findings

This study was conducted to determine the effectiveness of a nurse-led mobile-based education program for patients with PC who are at risk of MetS. This study showed that a nurse-led mobile-based health coaching program promoted a healthy lifestyle in patients with PC receiving ADT, which ultimately improved the levels of some MetS components (eg, reduction in FBS and AC), body composition (eg, reduction in weight and BMI), and HRQoL (eg, the urinary irritative and obstructive domain).

In this study, with a 3-month intervention, the results indicated that each variable required a different amount of time to show changes. In a previous study, a period of at least 3 to 8 weeks was required to confirm significant changes in weight, BMI, and the levels of each MetS component following lifestyle intervention programs [48-50]. Specifically, AC, FBS, body weight, and BMI have been found to decrease significantly over short periods [49,51]. However, in studies with intervention periods of ≥6 months, significant results were confirmed for MetS components including HDL cholesterol, SBP, and DBP [52,53]. Even in weight control programs that included strict diet control and exercise intervention guidelines, MetS components, including HDL cholesterol, did not change over a period of 8 weeks and only showed a significant change in both men and women 10 months after the end of the intervention [54].

The study by Dawson et al [55] observed decreases in body weight and AC but not in HDL cholesterol level among the MetS components in patients with PC. Focht and colleagues [56] also reported decreased bodyweight in patients with PC. These studies included lifestyle interventions lasting 3 months for patients with PC who had received ADT. The mean duration of ADT was 14 months in the study by Focht et al [56] and 30 months in the study by Dawson et al [55]. Specifically, Focht et al [56] confirmed that the body composition of body weight and body fat decreased significantly at 3 months compared with the control group. Furthermore, Dawson et al [55] conducted a program that emphasized exercise and protein supplementation for 3 months in patients with PC receiving ADT. As a result, the MetS components, the AC of participants decreased most significantly in the experimental group than in the waitlist control group.

Reductions in body weight, AC, and FBS levels were closely related. When beginning to lose weight, the body temporarily lowers its metabolic function to maintain homeostasis and first metabolizes glucose, which is a basic energy source. When stored glycogen is broken down, the insulin mechanism is activated for additional energy consumption [57]. Weight loss causes a decrease in FBS level along with the action of insulin, which reactivates FBS, stored in the form of excess fat in the liver or abdomen. This fat is continuously used to generate energy, consequently, the fat accumulated in the liver or abdomen is consumed, leading to a reduced volume [58]. In this study, although no significant decrease in body fat was observed, a decrease in AC was observed. Similarly, in previous
studies, decreases in body weight [56], AC [55], and FBS levels [49] were confirmed 3 months after the intervention. In this study, there were no significant changes in body composition related to body fat and skeletal muscle mass or in MetS components related to blood pressure (SBP and DBP) and lipids (HDL cholesterol and triglyceride). Insulin resistance develops over a long period, which increases the risk of obesity, diabetes, and MetS. In addition, it affects the lipid ratio, leading to an increase in low-density lipoprotein cholesterol and a decrease in HDL cholesterol [58], and causes inflammatory changes, resulting in changes in blood pressure owing to an increase in the residual amount of sodium in the blood [59]. Moreover, in a study that examined prediabetic patients diagnosed with impaired fasting glucose levels over the course of a 10-year follow-up period, 37% of the patients developed diabetes [60]. This finding indicates that the disease mechanism does not change over a short period but rather progresses slowly. Therefore, a long-term follow-up study is required to more accurately confirm the effectiveness of a lifestyle intervention program for MetS.

Among the HRQoL domains, an effect was observed only in the urinary irritative and obstructive domain, which may be caused by the education on how to manage the side effects of ADT treatment in the third and fourth weeks of the intervention. Patients with PC who have undergone multiple surgeries or frequent radiation therapy complained of side effects [61], such as urinary irritation and frequency. During the intervention in this study, information about appropriate water intake and pelvic floor muscle exercises to relieve urinary irritative and obstructive symptoms such as urinary irritation and frequency were included in an educational brochure [57], and appropriate water intake and pelvic floor muscle exercise were recommended for the participants depending on the presence or absence of symptoms. Thus, participants with these symptoms may have experienced relief through this intervention. Of the remaining HRQoL domains, the urinary incontinence, bowel, sexual, and hormonal domains did not show statistically significant changes. HRQoL was assessed using a questionnaire on treatment side effects and symptoms during the preceding 4 weeks. Most participants in this study were receiving long-term treatment after surgery, with an average of 40 (SD 24.71) months of ADT treatment. The bowel domain contains questions about whether the participants had diarrhea and bloody stools, which are acute side effects of participants who have undergone PC resection surgery. This domain had a higher mean score than the other HRQoL domains, with an average of ≥95 beginning at the baseline; therefore, these questions might not be relevant for participants of this study who were not in the acute stage after surgery. In addition, most questions in the hormonal domain examined the side effects that appear toward the start of ADT such as breast tenderness, bloody stools, and weight loss. Therefore, the insignificant changes in this domain might be because the participants in this study had already been taking medication for these side effects. At 19 points, the sexual domain had the lowest mean score of all the HRQoL domains. Sexual function might be limited in recovery as a result of a short-term health coaching program [62,63]. Although a previous study found that exercise can improve sexual function [64], sexual function requires interaction with a partner [65]. To restore sexual function, it is necessary to combine both psychological intervention and drug treatment [66]. This study found lifestyle changes to be critical in reducing the risk of MetS and that improved exercise and nutritional regimens should be implemented consistently for at least 3 months. Furthermore, a patient-centered, individualized approach that considers the side effects of ADT is needed to increase adherence.

Limitations
This study had some limitations. First, the criteria for participant selection included patients with at least 1 MetS component. To more thoroughly examine the effectiveness of this program on MetS management, stricter inclusion criteria should have been applied, such as only including those who were being treated with luteinizing hormone–releasing hormone, which is a type of ADT that causes the occurrence of more MetS components MetS components; those with ≥3 MetS components; those with a sedentary lifestyle; or those with a specific time after radical prostatectomy. Owing to data unavailability, this study only collected the Gleason score but not the information on TNM classification as the patient’s pathological data. It is important to establish accurate patient pathological criteria; hence, future studies should incorporate both the TNM classification and the Gleason score with accurate criteria. Second, the 3-month intervention period planned in this study was limited to evaluating the continuation of self-management and improvement in MetS risk factors. It is necessary to extend the intervention period to at least 6 months to further evaluate the persistence of self-management. As older men with PC receiving ADT are at high risk of osteoporosis, measuring bone health–related indicators in addition to changes in body composition is highly recommended for future studies. Third, it is difficult to generalize the results of this study because the participants were recruited from a single hospital and not from a multicenter. Fourth, obtaining low Cronbach α values for some items in the lifestyle score may indicate a mismatch between the measurement and the participant's lifestyle. This is because the lifestyle score was created for the general adult population; however, the participants in this study were older adults who had PC, had undergone surgery, and were taking hormone-suppressing medications.

Conclusions
A nurse-led mobile-based health coaching program was developed to promote a healthy lifestyle among patients with PC receiving ADT. The research findings confirmed that lifestyle changes through nurse-led education can improve the components of MetS, body composition, and ADT’s side effects. Therefore, by participating in this program, nurses were capable of creating changes in patients’ lifestyles and improving the self-management of patients who were beginning ADT for the first time. In addition, this study can provide a basis for the development of other mobile-based education programs and tools.
Acknowledgments

This work was supported by the Korea Medical Device Development Fund grant funded by the Korean Government (the Ministry of Science and Information and Communications Technology; the Ministry of Trade, Industry and Energy; the Ministry of Health and Welfare; and the Ministry of Food and Drug Safety; project number: 1711174508, KMD PR_20200901_0077). This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R1A6A3A13060728). This work was supported by the NRF of Korea grant funded by the Korean Government (MSIT; Ministry of Science and Information Communications Technology number 2019R1F1A1062769 and number 2022R1A2C1092084). KL received a scholarship from Brain Korea 21 FOUR Project funded by National Research Foundation NRF of Korea, Yonsei University College of Nursing.

Authors’ Contributions

KL contributed to data curation, formal analysis, investigation, methodology, project administration, software work, statistical analysis, visualization, original draft, writing, review and editing, funding acquisition. JP contributed to funding acquisition, investigation, methodology, project administration, visualization, review and editing, and supervision. All authors contributed to the conceptualization and have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1218 KB - mhealth_v12i1e47102_app1.pdf]

References


Abbreviations

AC: abdominal circumference
ADT: androgen deprivation therapy
DBP: diastolic blood pressure
EMR: electronic medical record
EPIC: expanded prostate cancer index composite
EPIC-26: 26-item Expanded Prostate Cancer Index Composite
FBS: fasting blood sugar
HDL: high-density lipoprotein
HRQoL: health-related quality of life
IMB: information-motivation-behavioral
MetS: metabolic syndrome
PC: prostate cancer
SBP: systolic blood pressure
T1: time point 1
T2: time point 2
T3: time point 3
Patient Engagement With and Perspectives on a Mobile Health Home Spirometry Intervention: Mixed Methods Study

Andrew W Liu¹, BA; William Brown, III¹,²,³,⁴, MA, DrPH, PhD; Ndubuisi E Madu¹, BA; Ali R Maiorano¹, BA; Olivia Bigazzi¹, MS; Eli Medina¹, MBA; Christopher Sorric¹, MS; Steven R Hays², MD; Anobel Y Odisho¹,³,⁵, MD, MPH

¹Center for Digital Health Innovation, University of California, San Francisco, CA, United States
²Department of Medicine, University of California, San Francisco, CA, United States
³Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, United States
⁴Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, United States
⁵Department of Urology, University of California, San Francisco, CA, United States

Corresponding Author:
Anobel Y Odisho, MD, MPH
Center for Digital Health Innovation
University of California
1700 Owens St 541
San Francisco, CA, 94158
United States
Phone: 1 415 353 7171
Email: anobel.odisho@ucsf.edu

Abstract

Background: Patient engagement attrition in mobile health (mHealth) remote patient monitoring (RPM) programs decreases program benefits. Systemic disparities lead to inequities in RPM adoption and use. There is an urgent need to understand patients’ experiences with RPM in the real world, especially for patients who have stopped using the programs, as addressing issues faced by patients can increase the value of mHealth for patients and subsequently decrease attrition.

Objective: This study sought to understand patient engagement and experiences in an RPM mHealth intervention in lung transplant recipients.

Methods: Between May 4, 2020, and November 1, 2022, a total of 601 lung transplant recipients were enrolled in an mHealth RPM intervention to monitor lung function. The predictors of patient engagement were evaluated using multivariable logistic and linear regression. Semistructured interviews were conducted with 6 of 39 patients who had engaged in the first month but stopped using the program, and common themes were identified.

Results: Patients who underwent transplant more than 1 year before enrollment in the program had 84% lower odds of engaging (odds ratio [OR] 0.16, 95% CI 0.07-0.35), 82% lower odds of submitting pulmonary function measurements (OR 0.18, 95% CI 0.09-0.33), and 78% lower odds of completing symptom checklists (OR 0.22, 95% CI 0.10-0.43). Patients whose primary language was not English had 78% lower odds of engaging compared to English speakers (OR 0.22, 95% CI 0.07-0.67). Interviews revealed 4 prominent themes: challenges with devices, communication breakdowns, a desire for more personal interactions and specific feedback with the care team about their results, understanding the purpose of the chat, and understanding how their data are used.

Conclusions: Care delivery and patient experiences with RPM in lung transplant mHealth can be improved and made more equitable by tailoring outreach and enhancements toward non-English speakers and patients with a longer time between transplant and enrollment. Attention to designing programs to provide personalization through supplementary provider contact, education, and information transparency may decrease attrition rates.

(JMIR Mhealth Uhealth 2024;12:e51236) doi:10.2196/51236

KEYWORDS
mobile health; mHealth; remote patient monitoring; interview; interviews; dropout; attrition; eHealth; methods; telemedicine; statistics; numerical data; patient-centered care; spirometry; lung transplant; lung; transplant; transplants; transplantation; organ; organs; engagement; monitor; monitoring; pulmonary; respiratory; lungs; experience; experiences; device; devices; complication; complications

https://mhealth.jmir.org/2024/11/e51236

JMIR Mhealth Uhealth 2024 | vol. 12 | e51236 | p.424
(page number not for citation purposes)
Introduction

Many large health systems have turned to remote patient monitoring (RPM) programs to improve population health outcomes, consistently engage with patients, and deliver care more efficiently at scale. RPM uses technology such as mobile devices (mHealth), wearables, or digital devices to communicate health information between patients and providers [1]. A 2022 systematic review found 268 published studies involving the use of RPM, reporting outcomes such as reduced health system costs, decreased hospitalizations, and improved patient quality of life [2,3]. The diversity of RPM clinical use cases includes dementia, diabetes, inflammatory bowel disease, sleep disorders, COVID-19, and cardiovascular disease, among many others [4-7].

However, RPM mHealth programs are most valuable when they maintain consistent patient engagement, allowing for adequate remote monitoring and increased data quality. High patient drop-off in RPM programs is common and seemingly inevitable upon implementation [8]. In a meta-analysis of 17 studies, 43% of patients stopped using the intervention before study completion [9]. In the same meta-analysis, observational real-world studies had even higher (49%) attrition compared to randomized controlled trials. Individual disengagement in RPM programs may arise from a wide range of factors, such as technological difficulties, lack of face-to-face encounters, or irrelevant content [10]. In addition, systemic disparities exist in terms of digital access and literacy, especially among vulnerable populations, leading to inequities in RPM adoption and use [10-13]. There is an urgent need to understand the patient experiences of RPM in real-world studies, especially for patients who have dropped off in engagement, as these patients may face untraditional and uncommunicated needs and challenges.

At the University of California, San Francisco (UCSF), a real-world RPM program for lung transplant recipients was launched in 2020 and is currently ongoing [14]. Patients used Bluetooth-enabled home spirometers to monitor pulmonary function and reported results and outcomes using a web-based chat interface. With the goal of engaging a higher proportion of patients and improving RPM compliance, we used a mixed methods approach to quantitatively identify the clinical and demographic predictors of patient engagement and conducted qualitative semistructured interviews with patients who had stopped using the home spirometry program after the first month to understand the nuances behind patient drop-off.

Methods

Home Spirometry Program

All patients who have had a lung transplant at UCSF are enrolled in an ongoing real-world RPM mHealth intervention that was implemented as a change in routine care [14]. The intervention is composed of an automated, English-only, chatbot-based symptom monitoring experience powered by a third-party vendor (Conversa Health, Inc), paired with a Bluetooth-enabled handheld home spirometer (Spirobank Smart or SmartOne, MIR Medical International Research) that allows patients to record their forced expiratory volume in the first second (FEV1) to assess and track their pulmonary function (Figure 1). The program was launched in May 2020, with all patients transplanted after May 1, 2020, automatically enrolled and given the chance to opt out. Additionally, all UCSF patients who had ever had a lung transplant were offered an opportunity to self-enroll (opt in) to the program at the time. In August 2020, all patients who had previously received a lung transplant and who had not self-enrolled were automatically bulk enrolled in the program, allowing them to opt out. Patients transplanted after May 2020 were onboarded in person during their posttransplant care, while patients transplanted before May 2020 were onboarded either digitally through mailed guides or during a routine outpatient follow-up visit.

Through the automated chat, patients can complete individual modules, in which they can report symptoms, manually input new FEV1 measurements recorded from their Bluetooth spirometer, and receive educational content embedded into chat modules (Figure 2). The goal of the program is early detection of acute and chronic allograft dysfunction and infections. Any abnormal drops in pulmonary function (>10% from baseline for each patient) or concerning patient symptoms generated an immediate alert to an electronic health record shared in-basket that was monitored by the lung transplant care team. Patients were also provided with the clinic contact information and instructed to reach out for additional advice. The transplant team managed clinical findings at their discretion. Additionally, patients are expected to engage with routine automated chat prompts as part of their posttransplant care indefinitely to provide the most up-to-date spirometry and symptom data to their providers, who review with patients during outpatient follow-up visits. Initially, all patients were on a weekly chat reminder cadence, with a reversion to a daily reminder cadence if their condition deteriorated (as defined by a 10% drop in FEV1 or the reporting of concerning symptoms). In May 2021, the chat reminder cadence was changed to allow patients 1-year posttransplant with stable conditions to opt into a monthly cadence, with the possibility to return to a weekly or daily cadence if their condition deteriorated, and the eventual readjudication to weekly or monthly once they began to recover. Patients could self-initiate chat sessions at will.
Figure 1. Home spirometry device and onboarding.

Figure 2. Screenshot of patient chat experience. FEV₁: forced expiratory volume in the first second.
Predictors of Patient Engagement

The 3 primary outcomes were patient engagement with the intervention, including (1) module engagement, defined as completing at least 1 module of any chat, (2) spirometry engagement, defined as patients who submitted at least 1 home spirometry FEV$_1$ value, or (3) symptom checklist engagement, defined as patients who responded to a symptom-reporting checklist at least once. Unengaged patients were defined as those who did not (1) complete any modules, (2) submit FEV$_1$ values, or (3) complete symptom checklists. As submitting a FEV$_1$ or a symptom checklist inherently resulted in the completion of a module, all patients classified as engaged by definitions 2 and 3 were also classified as engaged by definition 1, with definition 1 serving as the broadest outcome. It was possible for patients to have only engaged in educational modules without FEV$_1$ or symptom submission, resulting in only meeting the module engagement definition without meeting the other 2. Additionally, as secondary outcomes, the number of times patients completed the chat, submitted FEV$_1$ values, and symptom checklists were also calculated for each patient within the first year after enrollment. Since patients were asked to complete the chat at regularly scheduled repeated intervals, patients who died during their first year of enrollment were excluded from this secondary analysis.

Patient demographic and scheduling data were extracted from the electronic health record. This included patient age, transplant date, sex, race or ethnicity, address, insurance payor, primary language, and marital status. The primary diagnosis resulting in a need for transplantation was classified into 5 categories: restrictive disease, obstructive disease, cystic fibrosis or bronchiectasis, pulmonary hypertension, and other disease. Rural or urban status was assigned at the zip code level using the rural-urban community area codes classification [15]. Area Deprivation Index (ADI) national percentiles, based on a patient’s US census block group location, were used as a proxy measure of socioeconomic status [16].

Differences in the patient cohort conditioned on engagement status based on whether or not patients had completed any modules were compared using the chi-square test for categorical features and the 2-sample t test (2-tailed) for continuous features. Multivariate logistic and linear regression models were created to assess the predictors of engagement defined by chat module completion and longitudinal engagement, respectively. All analyses were performed using R (version 3.5.1; R Foundation for Statistical Computing), and a $P$ value $<$ .05 was considered significant.

Qualitative Interviews

A convenience sample of patients (n=6) who initially engaged with the mHealth program in the first month but subsequently stopped using it (n=39) were recruited to participate in 40-minute semistructured qualitative interviews. The interviews were led by 1 of 2 user experience designers (NEM and ARM) and conducted through Zoom videoconferencing (Zoom Technologies Inc). Patients were invited to include their significant others or caretakers in the call to help provide additional context and reassurance. Patients were compensated with a US $25 gift card. Interviews were audio-recorded and transcribed. NEM, ARM, AWL, and OB then collaboratively synthesized results to determine concepts and themes by pairing quotes and notes to relevant themes. Three individuals who did not participate in the interviews (AWL, WB III, and AYO) reviewed the selected themes with accompanying quotes and selected quotes for presentation.

The qualitative interview opened with topics concerning the patient’s background, including transplant experience, current condition, the frequency of home spirometry, and chat use. Patients were prompted with screenshots of the chat experience and asked about their perceptions, program expectations, reasons for opting out and nonuse, and any challenges they experienced. Patients were also asked about whether solving the challenges they brought up would cause them to reengage at a higher rate. Finally, patients were also given the opportunity to ask questions and provide comments on anything not already touched upon.

Ethical Considerations

The retrospective quantitative and prospective qualitative portions of this study, including patient compensation, were separately reviewed and approved by the University of California, San Francisco institutional review board (22-35950 and 22-35948).

Results

Overview

Between May 4, 2020, and November 1, 2022, a total of 626 patients were enrolled in the chat and 25 patients opted out, resulting in a total of 601 patients included in multivariate logistic regression analysis. Additionally, 33 patients died within their first year of enrollment, resulting in 568 patients included in multivariate linear regression modeling.

In total, 479 (79.7%) patients completed ≥1 module, 433 (72%) patients submitted ≥1 FEV$_1$ value, and 438 (72.9%) patients submitted ≥1 symptom assessment. The engaged or unengaged cohorts when compared by module completion status were composed of patients of similar sex, race or ethnicity, rural or urban zip code status, marital status, primary language, insurance type, and diagnosis. Engaged patients were marginally younger (66.1 vs 65.2 years; $P<$ .01), had more recent transplants (1.6 years since transplant for engaged patients vs 5.7 years for nonengaged patients; $P<$ .01), and lived in areas with lower ADI national percentiles (14 vs 16; $P<$ .01; Table 1).
Table 1. Patient demographics by module completion (engagement) status (n=601). Patients were considered engaged if they completed any chat modules, including the forced expiratory volume in the first second (FEV1) and symptom submission modules (definition 1).

<table>
<thead>
<tr>
<th></th>
<th>Not engaged participants (n=122)</th>
<th>Engaged participants (n=479)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient age (years), median (IQR)</td>
<td>66.1 (57.7-72.8)</td>
<td>65.2 (55.8-71.2)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Tim since transplant (years), median (IQR)</td>
<td>5.7 (2.7-9.3)</td>
<td>1.6 (0.2-5.1)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Transplant date ≥1 year of enrollment date, n (%)</td>
<td>109 (89.7)</td>
<td>279 (58.2)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Submitted FEV1 data within 1 year, n (%)</td>
<td>___a</td>
<td>433 (90.4)</td>
<td></td>
</tr>
<tr>
<td>Submitted symptom data within 1 year, n (%)</td>
<td>___</td>
<td>437 (91.4)</td>
<td></td>
</tr>
<tr>
<td>FEV1 submissions in the first year, median (IQR)</td>
<td>___</td>
<td>26 (8-52)</td>
<td></td>
</tr>
<tr>
<td>Symptom submissions in the first year, median (IQR)</td>
<td>___</td>
<td>24 (7-50)</td>
<td></td>
</tr>
<tr>
<td>Total number of modules completed in the first year, median (IQR)</td>
<td>0 (0-0)</td>
<td>23 (9-40)</td>
<td>.30</td>
</tr>
<tr>
<td>Male, n (%)</td>
<td>76 (62.3)</td>
<td>260 (57.6)</td>
<td>.30</td>
</tr>
<tr>
<td>Ethnicity, n (%)</td>
<td></td>
<td></td>
<td>.43</td>
</tr>
<tr>
<td>Asian, Native Hawaiian, or other Pacific Islander</td>
<td>6 (5.3)</td>
<td>39 (8.5)</td>
<td></td>
</tr>
<tr>
<td>Black or African American</td>
<td>7 (6.1)</td>
<td>37 (8)</td>
<td></td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>31 (27.2)</td>
<td>94 (20.4)</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>62 (54.4)</td>
<td>252 (54.7)</td>
<td></td>
</tr>
<tr>
<td>Other or unknown</td>
<td>8 (7)</td>
<td>39 (8.5)</td>
<td></td>
</tr>
<tr>
<td>Primary language, n (%)</td>
<td></td>
<td></td>
<td>.13</td>
</tr>
<tr>
<td>English</td>
<td>99 (86.8)</td>
<td>424 (92)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>15 (13.2)</td>
<td>37 (8)</td>
<td></td>
</tr>
<tr>
<td>Urban, n (%)</td>
<td>114 (95)</td>
<td>439 (95.6)</td>
<td>.96</td>
</tr>
<tr>
<td>ADIb national percentile, median (IQR)</td>
<td>16 (8-28)</td>
<td>14 (5-30)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Marital status, n (%)</td>
<td></td>
<td></td>
<td>.92</td>
</tr>
<tr>
<td>Married or partnered</td>
<td>79 (69.3)</td>
<td>321 (69.6)</td>
<td></td>
</tr>
<tr>
<td>Single, separated, or other</td>
<td>35 (30.7)</td>
<td>140 (30.4)</td>
<td></td>
</tr>
<tr>
<td>Insurance, n (%)</td>
<td></td>
<td></td>
<td>.09</td>
</tr>
<tr>
<td>Commercial</td>
<td>20 (18.2)</td>
<td>127 (28)</td>
<td></td>
</tr>
<tr>
<td>Medicare</td>
<td>73 (66.4)</td>
<td>280 (61.8)</td>
<td></td>
</tr>
<tr>
<td>Medicaid</td>
<td>13 (11.8)</td>
<td>39 (8.6)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>4 (3.6)</td>
<td>7 (1.5)</td>
<td></td>
</tr>
<tr>
<td>Diagnosis, n (%)</td>
<td></td>
<td></td>
<td>.77</td>
</tr>
<tr>
<td>Restrictive disease</td>
<td>77 (72)</td>
<td>291 (72.4)</td>
<td></td>
</tr>
<tr>
<td>Obstructive disease</td>
<td>11 (13.1)</td>
<td>31 (12.4)</td>
<td></td>
</tr>
<tr>
<td>Cystic fibrosis</td>
<td>14 (10.3)</td>
<td>50 (7.7)</td>
<td></td>
</tr>
<tr>
<td>Pulmonary hypertension</td>
<td>3 (2.8)</td>
<td>21 (5.2)</td>
<td></td>
</tr>
<tr>
<td>Other disease</td>
<td>2 (1.9)</td>
<td>9 (2.2)</td>
<td></td>
</tr>
</tbody>
</table>

aNot applicable.
bADI: Area Deprivation Index.

Predictors of Engagement

In a multivariate logistic regression model to identify predictors of engagement as defined by completion of a module (definition 1), patients who were enrolled ≥1 year from their transplant had 84% lower odds of engaging compared to those with more recent transplants (odds ratio [OR] 0.16, 95% CI 0.07-0.35; P<.01) when demographic factors (race or ethnicity, age, zip code status, insurance type, marital status, diagnosis, and socioeconomic status) were held constant. Similarly, patients
with ≥1 year between transplant and enrollment had 82% lower odds of submitting pulmonary function measurements (OR 0.18, 95% CI 0.09-0.33; P<.01) and 78% lower odds of completing symptom checklists (OR 0.22, 95% CI 0.10-0.43; P<.01). Patients whose primary language was not English had 78% lower odds of engaging compared to primarily English speakers (OR 0.22, 95% CI 0.07-0.67; P<.01). Patient age, race or ethnicity, marital status, insurance type, sex, rural or urban zip code status, ADI national percentile ranking, and diagnosis were not found to have significant associations with engagement in the multivariate logistic regression model (Table 2).

In multivariate linear regression evaluating the number of modules completed in the first year of enrollment, single patient completed 8.69 fewer modules (95% CI –13.75 to –3.64; P<.01), patients with non-English primary languages completed 11.89 fewer modules (95% CI –20.62 to –3.17; P<.01), and patients with longer duration since transplant completed 15.22 fewer modules (95% CI –19.69 to –10.75; P<.01; Table 3). Multivariate analysis conducted on the number of FEV1 and symptom submissions within the first year found that longer duration between transplant and enrollment and having single or separated or other marital status remained predictors for longitudinal engagement (Table S1 in Multimedia Appendix 1).

Table 2. Predictors of the degree of engagement.\(^a\)

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Number of FEV1 submissions</th>
<th>P value</th>
<th>Number of symptom submissions</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>0.35 (–0.13 to 0.82)</td>
<td>.16</td>
<td>0.21 (–0.20 to 0.62)</td>
<td>.32</td>
</tr>
<tr>
<td><strong>Race or ethnicity (vs White)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian, Native Hawaiian, or other Pacific Islander</td>
<td>0.90 (–13.53 to 7.62)</td>
<td>.92</td>
<td>–2.94 (–17.31 to 11.44)</td>
<td>.69</td>
</tr>
<tr>
<td>Black or African American</td>
<td>–8.32 (–0.13 to 0.82)</td>
<td>.23</td>
<td>–8.42 (–20.83 to 3.98)</td>
<td>.18</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>–2.96 (–21.82 to 5.19)</td>
<td>.58</td>
<td>–3.49 (–13.16 to 6.18)</td>
<td>.48</td>
</tr>
<tr>
<td>Other</td>
<td>5.91 (–6.73 to 18.56)</td>
<td>.36</td>
<td>5.60 (–6.23 to 17.42)</td>
<td>.35</td>
</tr>
<tr>
<td><strong>Marital status (vs married or partnered)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single or separated or other</td>
<td>–13.86 (–22.13 to –5.59)</td>
<td>&lt;.01</td>
<td>–11.59 (–19.23 to –3.95)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td><strong>Insurance (vs commercial)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicare</td>
<td>2.43 (–7.63 to 12.48)</td>
<td>.64</td>
<td>5.40 (–3.62 to 14.42)</td>
<td>.24</td>
</tr>
<tr>
<td>Medicaid</td>
<td>1.08 (–13.77 to 15.92)</td>
<td>.89</td>
<td>3.42 (–10.38 to 17.22)</td>
<td>.63</td>
</tr>
<tr>
<td>Other</td>
<td>–9.10 (–38.03 to 19.82)</td>
<td>.54</td>
<td>–11.27 (–34.86 to 12.31)</td>
<td>.35</td>
</tr>
<tr>
<td><strong>Sex (vs male)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>0.42 (–7.09 to 7.94)</td>
<td>.91</td>
<td>0.07 (–6.82 to 6.96)</td>
<td>.98</td>
</tr>
<tr>
<td><strong>Primary language (vs English)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-English</td>
<td>–11.38 (–25.65 to 2.90)</td>
<td>.12</td>
<td>–10.51 (–23.95 to 2.93)</td>
<td>.12</td>
</tr>
<tr>
<td>Transplant date ≥1 year of enrollment date</td>
<td>–16.42 (–23.75 to –9.10)</td>
<td>&lt;.01</td>
<td>–14.71 (–21.53 to –7.88)</td>
<td>.01</td>
</tr>
<tr>
<td>Rural</td>
<td>–1.55 (–19.04 to 15.94)</td>
<td>.86</td>
<td>0.88 (–15.77 to 17.54)</td>
<td>.92</td>
</tr>
<tr>
<td>ADI(^b) national percentile</td>
<td>–0.07 (–0.26 to 0.13)</td>
<td>.49</td>
<td>–0.09 (–0.27 to 0.09)</td>
<td>.32</td>
</tr>
<tr>
<td><strong>Diagnosis (vs restrictive disease)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cystic fibrosis</td>
<td>–3.15 (–19.65 to 13.36)</td>
<td>.71</td>
<td>–7.07 (–21.38 to 7.24)</td>
<td>.33</td>
</tr>
<tr>
<td>Obstructive disease</td>
<td>0.39 (–10.56 to 11.34)</td>
<td>.94</td>
<td>2.09 (–8.18 to 12.36)</td>
<td>.69</td>
</tr>
<tr>
<td>Pulmonary hypertension</td>
<td>–9.78 (–26.75 to 7.19)</td>
<td>.26</td>
<td>–8.49 (–24.10 to 7.12)</td>
<td>.29</td>
</tr>
<tr>
<td>Other disease</td>
<td>8.97 (–16.77 to 34.70)</td>
<td>.49</td>
<td>9.64 (–14.92 to 34.20)</td>
<td>.44</td>
</tr>
</tbody>
</table>

\(^a\)Results from a multivariate linear regression model (n=568).

\(^b\)ADI: Area Deprivation Index.
Table 3. Predictors of module completion engagement. Results from a multivariate logistic regression model (n=601). Patients were considered engaged if they completed any chat modules, including the forced expiratory volume in 1 second and symptom submission modules (definition 1).

<table>
<thead>
<tr>
<th>Predictors</th>
<th>OR(^a) (95% CI)</th>
<th>(P) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>0.99 (0.94-1.03)</td>
<td>.78</td>
</tr>
<tr>
<td>Race or ethnicity (vs White)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian, Native Hawaiian, or other Pacific Islander</td>
<td>1.94 (0.51-10.05)</td>
<td>.37</td>
</tr>
<tr>
<td>Black or African American</td>
<td>1.19 (0.38-4.63)</td>
<td>.78</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>1.20 (0.50-3.10)</td>
<td>.69</td>
</tr>
<tr>
<td>Other</td>
<td>1.20 (0.44-3.88)</td>
<td>.74</td>
</tr>
<tr>
<td>Marital status (vs married or partnered)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single or separated or other</td>
<td>0.95 (0.48-1.94)</td>
<td>.89</td>
</tr>
<tr>
<td>Insurance (vs commercial)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicare</td>
<td>0.78 (0.32-1.82)</td>
<td>.58</td>
</tr>
<tr>
<td>Medicaid</td>
<td>0.38 (0.12-1.26)</td>
<td>.11</td>
</tr>
<tr>
<td>Other</td>
<td>0.20 (0.03-1.28)</td>
<td>.08</td>
</tr>
<tr>
<td>Sex (vs male)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>1.25 (0.68-2.34)</td>
<td>.48</td>
</tr>
<tr>
<td>Primary language (vs English)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-English</td>
<td>0.22 (0.07-0.67)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Transplant date ≥1 year of enrollment date</td>
<td>0.16 (0.07-0.35)</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>Rural</td>
<td>0.53 (0.13-2.62)</td>
<td>.38</td>
</tr>
<tr>
<td>ADI(^b) national percentile</td>
<td>1.00 (0.98-1.02)</td>
<td>.95</td>
</tr>
<tr>
<td>Diagnosis (vs restrictive disease)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cystic fibrosis</td>
<td>1.04 (0.28-4.43)</td>
<td>.96</td>
</tr>
<tr>
<td>Obstructive disease</td>
<td>0.85 (0.35-2.23)</td>
<td>.74</td>
</tr>
<tr>
<td>Pulmonary hypertension</td>
<td>0.75 (0.19-3.92)</td>
<td>.71</td>
</tr>
<tr>
<td>Other disease</td>
<td>1.09 (0.15-22.20)</td>
<td>.94</td>
</tr>
</tbody>
</table>

\(a\)OR: odds ratio.  
\(b\)ADI: Area Deprivation Index.

Qualitative Interviews

Overview

Interviews were conducted with 2 male and 4 female patients, aged 35-70 years, and White and Latinx. Patients ranged from 1 to 18 years posttransplantation. One patient had his primary caretaker present as his proxy (Table 4). From the interviews, 4 key thematic concepts emerged: challenges with devices, communication breakdowns, desire for more personal interactions and specific feedback with the care team about data and concerns, and understanding the purpose of the care chat.
Communication Breakdowns

The transition for patients from using the automated care chat to reaching out to their care team via phone call or secure portal message led to communication breakdowns. Phone calls would occur back and forth between patients and different clinic staff before they were able to reach the right individual and receive relevant answers. Sometimes if you have a question, I guess you have to call your nurse coordinator. And at times you have to wait for them to call you back.... So, I'm at work, so sometimes I miss their phone calls. Sometimes we play phone tag... [PR1]

Patients had challenges with devices

Patients reported a diverse set of difficulties associated with the use of their spirometer and mobile devices at home. Challenges encompassed all facets of the experience, such as difficulties with installing the spirometer app on their mobile device and connecting their spirometers through Bluetooth.

My machine will not connect to my phone or anybody else's phone. I tried unloading it, reloading it, even took it to AT&T because I'm not very tech savvy. [Participant 3]

Low technological literacy was a barrier to engagement, and further product development is needed to make it easier to troubleshoot difficulties. One patient reported their preference to carry around an oximeter instead of using their spirometer due to a lack of confidence in their spirometer from experiencing device issues.

Upon use, patients also reported having issues with properly using their spirometer device. Changes in their position or using their spirometer device issues.

I have a hard time because I have to hold my phone, and I got to blow into it.... if I'm holding the phone down, now I'm having my neck down, it restricts some of my air. [Participant 2]

Relatedly, patients reported frustration from using their home spirometer.

... it deflates me emotionally because it's always low.... I'm not hitting the 3. I'm hitting like, 1.9. So medically, I'm okay. But to me, it keeps on failing. [Participant 2]

Feedback with the care team about data and concerns

Patients also wanted to know how their data were being used by the care team. They preferred to communicate with a member of their care team as opposed to entering their FEV₁ or symptom measurements into the chat. “The chat doesn't really offer that opportunity to talk to someone in real life” [PR1].

The automated aspect of the care chat also resulted in a perceived technological barrier between patients and their care team. Patients were keenly aware of the automation and felt it did not provide enough incentive to stay engaged.

Lung transplant is a partnership and a lot of hands and a lot of faith and a lot of luck. And a lot of things came together to make a successful surgery.... You're part of me now, you know. So yeah, any interaction is great. I'd rather talk to you. It's [the Care Chat] not very personal, very impersonal for the drastic surgery that we've had. [Participant 4]

Communication Breakdowns

The transition for patients from using the automated care chat to reaching out to their care team via phone call or secure portal message led to communication breakdowns. Phone calls would occur back and forth between patients and different clinic staff before they were able to reach the right individual and receive relevant answers.

Sometimes if you have a question, I guess you have to call your nurse coordinator. And at times you have to wait for them to call you back.... So, I'm at work, so sometimes I miss their phone calls. Sometimes we play phone tag... [PR1]

Patients had challenges with devices

Patients reported a diverse set of difficulties associated with the use of their spirometer and mobile devices at home. Challenges encompassed all facets of the experience, such as difficulties with installing the spirometer app on their mobile device and connecting their spirometers through Bluetooth.

My machine will not connect to my phone or anybody else's phone. I tried unloading it, reloading it, even took it to AT&T because I'm not very tech savvy. [Participant 3]

Low technological literacy was a barrier to engagement, and further product development is needed to make it easier to troubleshoot difficulties. One patient reported their preference to carry around an oximeter instead of using their spirometer due to a lack of confidence in their spirometer from experiencing device issues.

Upon use, patients also reported having issues with properly using their spirometer device. Changes in their position or using their spirometer device issues.

I have a hard time because I have to hold my phone, and I got to blow into it.... if I'm holding the phone down, now I'm having my neck down, it restricts some of my air. [Participant 2]

Relatedly, patients reported frustration from using their home spirometer.

... it deflates me emotionally because it's always low.... I'm not hitting the 3. I'm hitting like, 1.9. So medically, I'm okay. But to me, it keeps on failing. [Participant 2]

Feedback with the care team about data and concerns

Patients also wanted to know how their data were being used by the care team. They preferred to communicate with a member of their care team as opposed to entering their FEV₁ or symptom measurements into the chat. “The chat doesn't really offer that opportunity to talk to someone in real life” [PR1].

The automated aspect of the care chat also resulted in a perceived technological barrier between patients and their care team. Patients were keenly aware of the automation and felt it did not provide enough incentive to stay engaged.

Lung transplant is a partnership and a lot of hands and a lot of faith and a lot of luck. And a lot of things came together to make a successful surgery.... You're part of me now, you know. So yeah, any interaction is great. I'd rather talk to you. It's [the Care Chat] not very personal, very impersonal for the drastic surgery that we've had. [Participant 4]

To patients, the automated chat responses did not provide enough contextual feedback after a FEV₁ measurement was properly submitted to bring patients peace of mind. This resulted in concerns and a lack of confidence in at-home FEV₁ measurements.

I don't know what I'm doing in the sense that I'm not getting any feedback from the other end. So,.... thinking like, well, what am I doing ? [Participant 2]

Patients also wanted to know how their data were being used by the care team to assess their condition and wanted some acknowledgment that someone had reviewed their latest FEV₁ measurement. The lack of feedback from their care team about their inputted values caused patients to want to disengage. There was a strong desire for additional patient-provider interaction around the FEV₁ submission so that patients knew that their data were not going to be ignored.
I don’t even know if the team ever looks at what I submit via the care chat, because I don’t ever get any feedback from them. So, it’s kind of like, I think they see it, but I really don’t know. And in that sense, it’s like, well, do they really care? Maybe they do. Maybe they don’t? This instantly makes me go, oh, well, if Dr. Hays is looking at these numbers, then I definitely want to provide them. [Participant 5]

Patients Understand the Purpose of the Care Chat Experience

Patients strongly understood that their use of the care chat allowed providers to monitor their progress posttransplant. This resulted in patients being responsive to chat prompts. Furthermore, patients understood the clinical purpose of specifically collecting their FEV₁ measurement.

I think the purpose is to monitor our lung function from a distance and for the doctors to be able to get more frequent measurements of our FEV₁ without having us come in to do spirometry, or even go into our local hospital. It’s kind of a way to keep tabs on us from home in a way that’s safe for us and easier for them. [Participant 5]

At its core, home spirometry helps capture a more comprehensive view of a patient’s well-being through consistent remote clinical data measurements and symptom reporting. This more comprehensive view not only allows their care team to be better informed but also respond faster to concerning changes in condition.

I think it’s pretty great. I found it really helpful. And it’s been nice for me to have like a reason to do my FEV₁ and have a record of that. I like having more of a stand-up baseline that has more frequent measurements because I’ve always in clinic other than when I’ve been sick, it’s been really stable, but it’s nice to know at home like, oh yeah, this is really kind of where [my condition] lives. [Participant 5]

Discussion

Principal Findings

We sought to understand engagement in a cohort of 601 patients who underwent lung transplant engaged in a real-world home spirometry mHealth intervention. Engagement and continued use of mHealth tools are critical to effective remote care. However, consistent, quality engagement in mHealth is a difficult challenge and often requires the tailoring of interventions to specific subpopulations [17,18]. We describe several novel predictors of engagement in this intervention and report major themes resulting from qualitative interviews. During our interviews, we found that patients who underwent lung transplant and who had engaged in mHealth RPM spirometry have a strong desire for improved connection to their care team and that when they feel disconnected, are unsure if their data are being received or reviewed, have trouble escalating their concerns to their care team, or experience device difficulties, they are more likely to stop using mobile health (mHealth) tools and home monitoring devices. As the first study (to our knowledge) to focus on in-depth lung transplant patient experiences with drop-off and RPM, this study brings to light patient perspectives that future interventions can learn from for program design specifics.

Multivariate modeling found that having a recent lung transplant (within 1 year of a patient’s enrollment date) was consistently found to have a positive association with engagement with the chatbot. One possibility of this is that patients with longer times since transplants might be more confident managing their care without the use of an mHealth app. However, as the potential for chronic allograft dysfunction remains high even 1-year posttransplant, it remains important to engage and reengage patients to reduce the risk of missed allograft dysfunction. Future outreach efforts toward long-term transplant patients to show them the benefits of participation will be required. Additionally, the factors serving as proxies for socioeconomic status (urban or rural status, ADI percentile, and health insurance type) were not found to be associated with engagement in the mHealth program while controlled for other factors during multivariate modeling. This may be due to the fact the spirometers and the chat were provided to patients at no charge. Other factors such as race or ethnicity and age were also not consistently found to be predictive under our 3 definitions of engagement. Finally, our multivariate analysis centered on examining binary engagement with the chat found that having a non-English primary language was associated with lower engagement. Previous literature has shown that patients with a non-English primary language are significantly less likely to engage in mHealth and telemedicine [18]. Our finding has led us to plan to translate the program into Spanish, which is our second largest language cohort (6% of patients), as a next step toward equity.

Comparison With Previous Studies

Our study is not only the first to evaluate predictors of engagement of a mHealth RPM intervention in lung transplant recipients but also one of the first to examine RPM in a practice-wide, real-world setting, as opposed to a randomized controlled trial. Overall, the literature on factors influencing engagement in mHealth is mixed. Previous studies analyzing RPM programs have found age, income, and shorter time in program to be predictive of engagement [19]. Other studies have reported that drops in engagement are not associated with race or ethnicity, disease status, or geography [20]. Our analysis adds to the literature on RPM engagement: reporting that attrition over time remains a large complicated and multifaceted barrier and that trends in engagement are likely program and condition specific rather than being strongly influenced by socioeconomic factors. There is a need to continue to understand how social and clinical factors affect engagement in mHealth interventions to better engage vulnerable populations and not exacerbate existing disparities in care access. In addition, future work can be focused on comparative analyses between differences in engagement rates and outcome measurements (eg, emergency department visits, hospital readmissions).

In addition, interviews also revealed that some patients find the automated chatbot impersonal and crave additional interactions with their care teams. Program design can take this into account by creating defined instances and scenarios where providers
should reach out in case of stress or confusion. Patients desire human contact during their posttransplant care; however, it remains a challenge to balance the need to remotely monitor a large cohort of patients in an automated fashion without overwhelming clinical teams with notifications and also providing a meaningful experience that patients value. Future developments can focus on determining chat design and flow improvements that will offer patients more individualization and education, fine-tuning automated triage mechanisms to identify the right time to have patient-provider interactions, and streamlining patient-provider communication channels to keep patients informed and engaged. Content creation can also be initiated with patient reassurance in mind, such as the development of educational videos featuring their providers or trusted sources. Finally, many patients struggled with proper device setup, spirometer technological difficulties, and confidently using the device to obtain consistent results, leading to confusion and a lack of trust in program quality. These results are consistent with findings showing that low technological literacy and confusion remain major barriers toward even more widespread adoption of RPM and mHealth [10,21,22]. Spirometry device satisfaction levels have been found to greatly influence patient engagement levels, as well as constant contact and follow-up by physicians to increase patient satisfaction [23,24]. In response, program enhancements were built to alleviate the cognitive burden on patients by streamlining onboarding and reducing the number of unnecessary chats. Future developments can aim to help patients build confidence and trust with home spirometry, by adding staff for more detailed onboarding, focusing on education for patients who have technological issues or concerns during the onboarding period, and chat design incorporating more person-to-person real-time feedback by providers.

Limitations

Our study has several limitations. Our multivariate analysis did not account for health and digital literacy levels as these were not systematically assessed as part of routine clinical care and both have been shown to play major roles in a patient’s abilities to access health care. While our analysis factored in disease diagnosis type, there are likely additional clinical factors that may affect patient engagement, as transplant recipients with worse pulmonary function may be more apt to consistently engage to monitor their health. Future work can examine the efficacy of home spirometry monitoring in detecting adverse events when controlling for patients’ clinical conditions. Next, our study is a single-institution study of patients who have received lung transplant, and therefore the overall generalizability is reduced due to the specific nature of the patient’s condition, progression, and experiences with the program specifics such as Bluetooth spirometry. Interviewed patients were recruited only from those who initially dropped out after 1 month, and therefore the themes noted may only be representative of a dropped off in the engagement cohort. More work is required to elucidate the perspectives and reasons for continual engagement from patients whose engagements were consistent, steadily decreased over time, or dropped off at a later point in the program. In the future, we plan to interview patients who had excellent adherence to understand what factors are promoting adherence in these cases. Furthermore, the small sample size of our interview cohort increases the possibility of bias during the process of thematic analysis, particularly because we only interviewed 1 non-English speaker (using an interpreter, with the patient’s proxy present). Analyses from numerous implemented mHealth programs have shown that non-English speakers and patients of minority race or ethnicity are significantly less likely to engage in telemedicine [13,25,26]. Patient motivations for drop-off are nuanced and are likely strongly associated with sociocultural factors that can only be uncovered in larger, more diverse studies.

Conclusions

An mHealth intervention consisting of home spirometry paired with an automated care chat results in high engagement rates in patients who have received a lung transplant, particularly in those with more recent transplants. Interviews conducted on patients who have dropped off in engagement revealed program challenges and areas where mHealth care delivery can be improved to reduce engagement attrition, including addressing technological barriers and improving patient confidence.

Acknowledgments

WB was supported by the Agency for Healthcare Research and Quality (K12HS026383) and the National Center for Advancing Translational Sciences (KL2TR001870) during the writing of this research.

Authors’ Contributions

Conceptualization, project administration, data curation, formal analysis, investigation, writing—original draft, review, and editing were done by AWL. Project administration, data curation, formal analysis, investigation, writing—review, and editing were done by WB. Data curation, methodology, and project administration were done by NEM. Methodology and project administration were performed by ARM. Conceptualization and funding acquisition was done by OB. Funding acquisition and project administration were performed by EM. Data curation and formal analysis were performed by CS. Funding acquisition, resources, conceptualization, methodology, investigation, writing—review and editing, and supervision were done by SRH. Funding acquisition, conceptualization, formal analysis, methodology, investigation, writing—review and editing, and supervision were done by AYO.

Conflicts of Interest

None declared.
Multimedia Appendix 1
Predictors of the degree of engagement. Results from a multivariate logistic regression model.

[DOCX File, 16 KB - mhealth_v12i1e51236_app1.docx]

References


Abbreviations

ADI: Area Deprivation Index
FEV1: forced expiratory volume in the first second
mHealth: mobile health
OR: odds ratio
RPM: remote patient monitoring
UCSF: University of California, San Francisco

Edited by L Buis; submitted 26.07.23; peer-reviewed by J Edwards, C Peterson; comments to author 08.09.23; revised version received 27.10.23; accepted 12.12.23; published 20.03.24.

Please cite as:
Patient Engagement With and Perspectives on a Mobile Health Home Spirometry Intervention: Mixed Methods Study
JMIR Mhealth Uhealth 2024;12:e51236
URL: https://mhealth.jmir.org/2024/1/e51236
doi:10.2196/51236
PMID:38506896

©Andrew W Liu, William Brown, III, Ndubuisi E Madu, Ali R Maiorano, Olivia Bigazzi, Eli Medina, Christopher Sorric, Steven R Hays, Anobel Y Odisho. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 20.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Original Paper

Mining User Reviews From Hypertension Management Mobile Health Apps to Explore Factors Influencing User Satisfaction and Their Asymmetry: Comparative Study

Yunfan He¹, MSc; Wei Zhu²,³, PhD; Tong Wang⁴,⁵,⁶, PhD; Han Chen⁷, MD, PhD; Junyi Xin⁸, MS; Yongcheng Liu⁹, MS; Jianbo Lei¹⁰,¹¹,¹², MD, PhD; Jun Liang∥,¹³,¹⁴,¹⁵, MS

¹Center for Health Policy Studies, School of Public Health, Zhejiang University, Hangzhou, China
²Department of Cardiology, Second Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, China
³Cardiovascular Key Laboratory of Zhejiang Province, Hangzhou, China
⁴School of Health and Life Sciences, University of Health and Rehabilitation Sciences, Qingdao, China
⁵School of Basic Medical Sciences, Shandong University, Jinan, China
⁶Qingdao Hospital, University of Health and Rehabilitation Sciences (Qingdao Municipal Hospital), Qingdao, China
⁷Department of Cardiology, Second Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, China
⁸School of Information Engineering, Hangzhou Medical College, Hangzhou, China
⁹Netease Group, Hangzhou, China
¹⁰Clinical Research Center, Affiliated Hospital of Southwest Medical University, Luzhou, China
¹¹The First Affiliated Hospital, Hainan Medical University, Haikou, China
¹²Center for Medical Informatics, Health Science Center, Peking University, Beijing, China
¹³Department of AI and IT, Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China
¹⁴Key Laboratory of Cancer Prevention and Intervention., China National Ministry of Education, School of Medicine, Zhejiang University, Hangzhou, China
¹⁵School of Medical Technology and Information Engineering, Zhejiang Chinese Medical University, Hangzhou, China

Corresponding Author:
Jun Liang, MS
Department of AI and IT
Second Affiliated Hospital, School of Medicine
Zhejiang University
88 Jiefang Road, Shangcheng District
Hangzhou, 310000
China
Phone: 86 571 87783942
Fax: 86 571 87783526
Email: junl@zju.edu.cn

Abstract

Background: Hypertension significantly impacts the well-being and health of individuals globally. Hypertension management apps (HMAs) have been shown to assist patients in controlling blood pressure (BP), with their efficacy validated in clinical trials. However, the utilization of HMAs continues to be suboptimal. Presently, there is a dearth of real-world research based on big data and exploratory mining that compares Chinese and American HMAs.

Objective: This study aims to systematically gather HMAs and their user reviews from both China and the United States. Subsequently, using data mining techniques, the study aims to compare the user experience, satisfaction levels, influencing factors, and asymmetry between Chinese and American users of HMAs. In addition, the study seeks to assess the disparities in satisfaction and its determinants while delving into the asymmetry of these factors.

Methods: The study sourced HMAs and user reviews from 10 prominent Chinese and American app stores globally. Using the latent Dirichlet allocation (LDA) topic model, the research identified various topics within user reviews. Subsequently, the Tobit model was used to investigate the impact and distinctions of each topic on user satisfaction. The Wald test was applied to analyze differences in effects across various factors.

https://mhealth.jmir.org/2024/1/e55199

JMIR Mhealth Uhealth 2024 | vol. 12 | e55199 | p.436
(page number not for citation purposes)
Results: We examined a total of 261 HMAs along with their associated user reviews, amounting to 116,686 reviews in total. In terms of quantity and overall satisfaction levels, Chinese HMAs (n=91) and corresponding reviews (n=16,561) were notably fewer compared with their American counterparts (n=220 HMAs and n=100,125 reviews). The overall satisfaction rate among HMA users was 75.22% (87,773/116,686), with Chinese HMAs demonstrating a higher satisfaction rate (13,866/16,561, 83.73%) compared with that for American HMAs (73,907/100,125, 73.81%). Chinese users primarily focus on reliability (2165/16,561, 13.07%) and measurement accuracy (2091/16,561, 12.63%) when considering HMAs, whereas American users prioritize BP tracking (17,285/100,125, 17.26%) and data synchronization (12,837/100,125, 12.82%). Seven factors (easy to use: P<.001; measurement accuracy: P<.001; compatibility: P<.001; cost: P<.001; heart rate detection function: P=.02; blood pressure tracking function: P=.001; and interface design: P=.01) significantly influenced the positive deviation (PD) of Chinese HMA user satisfaction, while 8 factors (easy to use: P<.001; reliability: P<.001; measurement accuracy: P<.001; compatibility: P<.001; cost: P<.001; interface design: P<.001; real-time: P<.001; and data privacy: P=.001) affected the negative deviation (ND). Notably, BP tracking had the greatest effect on PD (β=3.703, P<.001), while cost had the most significant impact on ND (β=3.354, P<.001). All 12 factors (easy to use: P<.001; blood pressure tracking function: P<.001; data synchronization: P<.001; blood pressure management effect: P<.001; heart rate detection function: P<.001; data sharing: P<.001; reliability: P<.001; compatibility: P<.001; interface design: P<.001; advertisement distribution: P<.001; measurement accuracy: P<.001; and cost: P<.001) significantly influenced the PD and ND of American HMA user satisfaction. Notably, BP tracking had the greatest effect on PD (β=2.662, P<.001), while data synchronization had the most significant impact on ND (β=3.012, P<.001). In addition, the influencing factors of PD and ND in user satisfaction of HMA in China and the United States are different.

Conclusions: User satisfaction factors varied significantly between different countries, showing considerable asymmetry. For Chinese HMA users, ease of use and interface design emerged as motivational factors, while factors such as cost, measurement accuracy, and compatibility primarily contributed to user dissatisfaction. For American HMA users, motivational factors were ease of use, BP tracking, BP management effect, interface design, measurement accuracy, and cost. Moreover, users expect features such as data sharing, synchronization, software reliability, compatibility, heart rate detection, and nonintrusive advertisement distribution. Tailored experience plans should be devised for different user groups in various countries to address these diverse preferences and requirements.

(JMIR Mhealth Uhealth 2024;12:e55199) doi:10.2196/55199

KEYWORDS
hypertension management; mobile health; topic modeling; satisfaction; 2-factor model; comparative study

Introduction
The global prevalence of hypertension is on the rise. Hypertension management apps (HMAs) serve as convenient tools for effectively managing blood pressure (BP). These apps enhance users’ awareness of self-management, dietary and exercise habits, and medication adherence through features such as BP tracking, dietary guidance, exercise monitoring, educational resources, and medication reminders. The ultimate aim is to achieve effective BP control. Their effectiveness has been demonstrated in experimental settings. Globally, the number of patients with hypertension surged from 648 million in 1990 to 1.278 billion in 2019, marking a prevalence of 33% [1]. HMAs represent digital health tools with the potential for effectively controlling BP [2-9]. Their usability [10,11] and effectiveness [12] have been demonstrated in randomized controlled trials. In a systematic review and meta-analysis of 18 randomized controlled trials, Han et al [13] found that HMAs could significantly reduce BP and improve BP control rates. Moreover, within the framework of the current value-based medical policy, a hospital’s treatment efficacy relies not only on its in-hospital care but also on postdischarge patient attention and management. As a collaborative tool for out-of-hospital medical services, HMAs can assist patients in sustaining the effects of in-hospital treatments, thereby reducing the rate of hospital readmissions. McManus et al [14] demonstrated that compared with traditional nursing interventions, apps yield superior BP control effects within a year and incur lower incremental costs. However, in practical settings, many HMAs have been developed without adhering to evidence-based medicine [15] and lacked clinical certification before marketing [16], potentially posing adverse effects on hypertension management [17,18].

Although HMAs show significant BP management effects in controlled environments, their real-world outcomes are often unsatisfactory. Both user engagement and retention rates are low, with 62% of mobile health (mHealth) apps having fewer than 1000 monthly active users [19] and only 6.6% of patients with hypertension continuing to use HMAs [20]. These situations directly correlate with user satisfaction [21-24]. Therefore, enhancing user satisfaction can significantly improve the effectiveness of the app [25]. According to the Food and Drug Administration (FDA), in real-world scenarios, health information technology (HIT) is integrated into a complex sociotechnical system, and its actual impact is influenced by 4 primary factors aside from the product itself: people (whether they are involved or not), technologies (including HIT hardware and software), processes (the workflow of health care delivery), and organization (the procedure for HIT installation and configuration). Among these factors, external environmental elements (such as the policies and cultures of different countries) play a significant role [26]. The actual impacts of HMAs can vary significantly among user groups in different countries. Therefore, it is essential to explore and compare the satisfaction levels and influencing factors of user groups across various situations.
countries. This comparative analysis can enable targeted efforts to enhance the practical effectiveness of HMAs.

Traditional studies on user satisfaction of HMAs primarily consist of qualitative analyses [27,28]. However, these studies often lack breadth and depth, characterized by limitations such as small sample sizes, susceptibility to adverse observer effects and recall bias, and limited generalizability of conclusions. Furthermore, these studies often overlook the impact of external environmental factors and do not compare or analyze the satisfaction levels of user groups across different countries. For instance, Li et al [29] conducted semistructured interviews with 13 English-speaking patients with hypertension to explore their surface-level satisfaction with HMAs and analyze factors contributing to dissatisfaction. Kang and Park [30] developed an English HMA grounded on clinical guidelines for hypertension management. They used the modified Morisky Scale to assess perceived usefulness and satisfaction with this app among 38 patients diagnosed with hypertension. Although qualitative analyses allow for a deep exploration of individual attitudes, they have limitations such as a restricted number of apps that can be analyzed [31,32], a small sample size confined to a specific region [33], high research costs, and potential biases in conclusions [34].

Research on user satisfaction with HMAs lacks both quantitative studies driven by large volumes of user-generated content and in-depth exploration of the factors influencing satisfaction, particularly in terms of asymmetry. In addition, there is a notable absence of comparative analyses of user groups across different countries. Plante et al [35] manually annotated and summarized English user reviews of an HMA (Instant BP) from an online app store, discovering that users expressed greater satisfaction with apps yielding lower measurement results. Similarly, Wang et al [36] and Nuo et al [37] conducted quantitative analyses of user reviews for weight and sleep management apps to investigate user satisfaction and influencing factors. However, none of these studies conducted an in-depth evaluation of variations in satisfaction and influencing factors among user groups across different countries. The renowned Herzberg 2-factor theory [38] in management suggests that the factors influencing user satisfaction exhibit asymmetry and can be categorized into motivational factors that enhance satisfaction and hygienic factors that mitigate dissatisfaction. Using large-sample data and incorporating the asymmetry of factors impacting user satisfaction, while also considering variations among user groups in different countries, can effectively mitigate errors in constructing explanatory models. This approach enhances the predictability and generalizability of the model [39].

Therefore, this study adopts the 2-factor model and uses unsupervised clustering algorithms to quantitatively analyze user reviews of HMAs from major Chinese and American app stores globally. By considering the macro usage environment, the study aims to extract and compare the primary opinions of Chinese and American users, assess differences in satisfaction and its influencing factors, and explore the asymmetry within these factors.

**Methods**

**Informed Consent and Study Approval Statement**

All data used in this study were sourced from publicly accessible internet mobile app stores, encompassing app information and user reviews. Hence, this study does not entail any medical ethics concerns.

**Data Collection**

**HMA Identification**

In this comparative study, to conduct a comprehensive review of the primary HMAs, we consulted previous studies [36,37] and identified 10 widely used Chinese and American app stores from the 2 major mobile phone platforms (iOS [Apple Inc.] and Android [Google LLC]). These include 8 Chinese platforms (China Apple, Huawei, Xiaomi, OPPO, VIVO, Baidu, 360, and Application Treasure) and 2 American platforms (US Apple App Store and US Google Play Store). App stores in different countries can only access user data from their respective regions. Therefore, in April 2023, we conducted searches across the aforementioned 10 Chinese and American app stores using keywords such as “hypertension,” “high blood pressure,” “blood pressure management,” and “blood pressure recording.” We retrieved a total of 5016 apps, out of which 3591 remained after deduplication (Table S1 in Multimedia Appendix 1). Following guidelines from previous studies [40], we formulated detailed inclusion and exclusion criteria (Table S2 in Multimedia Appendix 1). Following guidelines from previous studies [40], we formulated detailed inclusion and exclusion criteria (Table S2 in Multimedia Appendix 1), which were independently screened by 2 researchers (YPH and J Liang). Both researchers underwent standardized training before the screening process, resulting in high consistency in their screening results ($\kappa=0.84$). Any discrepancies between the researchers were resolved through arbitration by another cardiovascular clinical expert (WZ). The specific screening process, following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines [41], is depicted in Figure 1.
User Review Collection

All 10 Chinese and American app platforms offer user ratings and review features, enabling users to provide quantitative ratings and qualitative descriptions of their satisfaction with apps. These ratings typically range from 1 (very poor) to 5 (very good). We used Python scripts (Python Foundation) and the Qimai mobile app data analysis platform [42] to extract all quantitative ratings and qualitative user reviews of the included HMAs. As of April 23, 2023, we collected a total of 295,927 quantitative ratings and qualitative user reviews, comprising 250,193 reviews from American users and 45,734 reviews from Chinese users. The proportion of Chinese reviews and downloads (16,561/24,204,832, 0.07%) is similar to that of American reviews and downloads (100,125/148,869,181, 0.07%).

User Review Preprocessing

The presence of false and meaningless user reviews in the app store data significantly impacted the topic mining of user reviews and interfered with the assessment of user satisfaction with HMAs. Hence, we conducted preprocessing on the user reviews with the following steps:

- Removing data containing only ratings without accompanying user reviews.
- Excluding user reviews posted on bot accounts, using the tweetbotornot package in R [43]. This resulted in the exclusion of 53,431 American and 16,940 Chinese reviews.
- Eliminating duplicate reviews, blank values, non-Chinese or English reviews, garbled characters, and reviews deemed meaningless. This process led to the exclusion of 80,829 American and 9697 Chinese reviews.
- Following this, we used the sentiment knowledge–enhanced pretraining algorithm [44] to calculate the emotional polarity of each user review, categorized into negative, neutral, and positive sentiments. Furthermore, contradictory data showing inconsistencies between user ratings and reviews were eliminated [45], resulting in the exclusion of 9243 American and 1362 Chinese reviews. Specifically, we removed data with user ratings of 1 or 2 points that lacked negative emotional polarities in user reviews, data with user ratings of 3 points that lacked neutral emotional polarities in user reviews, and data with user ratings of 4 or 5 points that missed positive emotional polarities in user reviews (Figure 2). Finally, we uniformly labeled different data with the same concept (Table S3 in Multimedia Appendix 1). Following the initial data preprocessing steps, a total of 124,425 user reviews were included in the LDA model for topic modeling, comprising 106,690 American reviews and 17,735 Chinese reviews.
Data Analysis

Overview

We used natural language processing technology and the LDA topic model to extract the main topics from both Chinese and American user reviews. Subsequently, following the 2-factor model, we constructed a Tobit model to analyze the correlation between different topics and user satisfaction. Finally, we used the Wald test to analyze the differences in the impact of each topic on user satisfaction.

LDA Topic Modeling

LDA is a widely used probability-based topic modeling algorithm [46]. It is known for its ease of operation, high efficiency, and positive impact on topic clustering and prediction accuracy [47]. As LDA is primarily a language model rather than a classification prediction model, perplexity serves as a common and effective indicator for evaluating the quality of the language model, rather than sensitivity. To efficiently and accurately extract the primary opinions and topics from Chinese and American user reviews, we used LDA, which is a 3-level hierarchical Bayesian model. LDA calculates the distribution probabilities of words and topics, enabling the clustering of latent semantic structures within user reviews to summarize the main topics.

For the LDA model input, we used the word segmentation set and specified the number of Chinese and American review topics, along with manually induced topics. Consequently, we executed the LDA topic modeling process as follows: Initially, we used the Jieba [48] and NLTK packages [49] in Python to segment the Chinese and American reviews, respectively. We then used stop-word lists compiled from Baidu, Harbin Institute of Technology, Sichuan University Machine Intelligence Laboratory, and standard Chinese and English stop words [50]. These stop-word lists were applied to delete stop words, including numbers, punctuation marks, emoticons, and blank values, from both Chinese and English word segmentations. Subsequently, any blank reviews were removed after the stop words had been eliminated. In addition, we conducted morphological restoration of English word segmentation. Subsequently, based on the perplexity curve and actual clustering effect [51], we determined that 11 topics were optimal for Chinese reviews and 12 topics for American reviews (Figures S1–S4 in Multimedia Appendix 1). Furthermore, we used the Gensim package [52] in Python to construct the LDA topic models for both Chinese and American reviews. Two researchers (YFH and J Liang) independently summarized and named the topics for each keyword set in the Chinese and American review topic models, respectively. Any discrepancies were resolved through arbitration by a third expert (WZ). Finally, the topics for each user review were determined based on their distribution probabilities across all topics generated by the LDA model [53].

Statistical Analysis

To delve deeper into the factors influencing user satisfaction with HMAs, in line with the 2-factor theory [38], we introduced 2 variables: positive deviation (PD) and negative deviation (ND) [54]. These variables are defined by the disparity between the user’s individual rating of HMAs and the overall rating displayed in the app store. They serve to indicate the discrepancy between the user’s personal rating and the average rating. PD and ND are mutually exclusive if a user’s deviation is positive and the ND value is 0, and vice versa. To simplify the calculation, we applied absolute value processing to ND. Consequently, the higher the ND value, the greater the degree of dissatisfaction. Both PD and ND values ranged between 0 and 4, calculated based on the difference between the user rating and the average rating displayed in the app store.
Results

Chinese and American HMA User Review Topics

We initially retrieved 5016 related apps from both Chinese and American app stores, yielding a total of 295,927 original reviews in both languages. Following screening and data preprocessing, we identified 261 HMAs with user reviews. Among these, 41 (15.7%) were exclusively available in Chinese, 170 (65.1%) were exclusively available in American, and 50 (19.2%) were available in both app stores. Ultimately, 116,686 user reviews were included in the analysis, with 100,125 (85.81%) in English and 16,561 (14.19%) in Chinese. Among these reviews, 87,773 (including 73,907 American reviews and 13,866 Chinese reviews) were rated 4 stars and above, resulting in an overall satisfaction rate of 75.22%; 73,907/100,125 (73.81%) for American apps and 13,866/16,561 (83.73%) for Chinese apps.

The results of LDA modeling revealed differences between Chinese and American reviews. In Chinese reviews, the number of reviews related to software reliability (2165/16,561, 13.07%) and measurement accuracy (2091/16,561, 12.63%) substantially exceeded other topics. Conversely, specific hypertension management functions such as BP tracking (17,285/100,125, 17.26%) and data synchronization (12,837/100,125, 12.82%) substantially outnumbered other topics in American reviews.

The LDA modeling revealed 11 topics in the Chinese reviews, with 5 (ease of use: P<.001 in both PD and ND; interface design: P=.01 in PD and P<.001 in ND; measurement accuracy: P<.001 in both PD and ND; compatibility: P<.001 in both PD and ND; and cost: P<.001 in both PD and ND) of them being significant (including 2 motivational factors and 3 hygienic factors). By contrast, the American reviews yielded 12 topics, all of which were significant (easy to use: P<.001 in both PD and ND; blood pressure tracking function: P<.001 in both PD and ND; data synchronization: P<.001 in both PD and ND; blood pressure management effect: P<.001 in both PD and ND; heart rate detection function: P<.001 in both PD and ND; data sharing: P<.001 in both PD and ND; reliability: P<.001 in both PD and ND; compatibility: P<.001 in both PD and ND; interface design: P<.001 in both PD and ND; advertisement distribution: P<.001 in both PD and ND; measurement accuracy: P<.001 in both PD and ND; cost: P<.001 in both PD and ND), comprising 6 motivational factors and 6 hygienic factors (Tables 1 and 2; Tables S4–S7 in Multimedia Appendix 1).

Table 1. Topics and keywords of Chinese reviews formulated by latent Dirichlet allocation topic modeling (N=16,561).

<table>
<thead>
<tr>
<th>Topics</th>
<th>Keywords</th>
<th>Reviews, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic 1: Easy to use</td>
<td>simple, convenient, easy to use, practical, rapid, recommended, easy to operate</td>
<td>6863 (41.44)</td>
</tr>
<tr>
<td>Topic 2: Reliability</td>
<td>healthy, quality, comprehensive, supportive, stable, professional, normal</td>
<td>2165 (13.07)</td>
</tr>
<tr>
<td>Topic 3: Measurement accuracy</td>
<td>measurement, accurate, collect, accuracy, test, data, inaccurate</td>
<td>2091 (12.63)</td>
</tr>
<tr>
<td>Topic 4: Attitude (positive)</td>
<td>very good, like, recommend, every day, awesome, useful, look forward to</td>
<td>1850 (11.17)</td>
</tr>
<tr>
<td>Topic 5: Compatibility</td>
<td>version, download, try, blood pressure monitor, Apple, connection, platform</td>
<td>878 (5.30)</td>
</tr>
<tr>
<td>Topic 6: Cost</td>
<td>fee, subscription, free, payment, upgrade, refund, paid</td>
<td>653 (3.94)</td>
</tr>
<tr>
<td>Topic 7: Heart rate detection function</td>
<td>heart rate, detection, value, body, monitor, watch, indicator</td>
<td>597 (3.60)</td>
</tr>
<tr>
<td>Topic 8: Blood pressure tracking function</td>
<td>blood pressure, function, record, tool, data, share, form</td>
<td>502 (3.03)</td>
</tr>
<tr>
<td>Topic 9: Interface design</td>
<td>updated, special, interface, clear, design, components, good-looking</td>
<td>471 (2.84)</td>
</tr>
<tr>
<td>Topic 10: Real-time</td>
<td>trial, view, status, anytime, anywhere, patient, daily</td>
<td>299 (1.81)</td>
</tr>
<tr>
<td>Topic 11: Data privacy</td>
<td>account, personal, information, security, management, privacy, licensing</td>
<td>192 (1.16)</td>
</tr>
</tbody>
</table>
Factors Affecting Chinese and American HMA User Satisfaction

We computed the variance inflation factor of the PD and ND models for both Chinese and American reviews to assess multicollinearity among the independent variables. The regression variance inflation factors of all independent variables in the 4 models were found to be <5, indicating the absence of multicollinearity-related issues [22] (Tables S8-S11 in Multimedia Appendix 1). Furthermore, in Chinese reviews, the topic “attitude (positive)” pertains solely to users’ positive attitudes and does not encompass opinions regarding the functions and utility of apps. Therefore, it was not included in the Tobit regression model.

In Table 3, model 1 displays the PD model results for Chinese reviews. With the exception of reliability ($P=56$, real-time ($P=21$), and data privacy ($P=52$), the other 7 topics (easy to use: $P<.001$; measurement accuracy: $P<.001$; compatibility: $P<.001$; cost: $P<.001$; heart rate detection function: $P=.02$; blood pressure tracking function: $P<.001$; and interface design: $P=.01$) discussed by users significantly influenced the PD of user satisfaction. Among them, the factors with the most substantial positive and negative effects were the BP tracking function ($\beta=354$, $P<.001$) and cost ($\beta=-232$, $P<.001$), respectively. Model 2 presents the ND model results for Chinese reviews. Except for the heart rate detection ($P=.64$) and BP tracking ($P=.14$) functions, the other 8 topics (easy to use: $P<.001$; reliability: $P<.001$; measurement accuracy: $P<.001$; compatibility: $P<.001$; cost: $P<.001$; interface design: $P<.001$; real-time: $P<.001$; and data privacy: $P=.001$) significantly impacted user satisfaction NDs. The factors with the most substantial positive and negative effects were cost ($\beta=3.703$, $P<.001$) and interface design ($\beta=-1.619$, $P<.001$), respectively. In Table 4, model 3 presents the PD model results for American reviews. All 12 topics (easy to use: $P<.001$; blood pressure tracking function: $P<.001$; data synchronization: $P<.001$; blood pressure management effect: $P<.001$; heart rate detection function: $P<.001$; data sharing: $P<.001$; reliability: $P<.001$; compatibility: $P<.001$; interface design: $P<.001$; advertisement distribution: $P<.001$; measurement accuracy: $P<.001$; and cost: $P<.001$) included in the model significantly influenced the user satisfaction PD. Among them, the factors with the most significant positive and negative effects were the BP tracking function ($\beta=312$, $P<.001$) and data synchronization ($\beta=-593$, $P<.001$), respectively. Model 4 displays the ND model results for American reviews. All 12 topics (easy to use: $P<.001$; blood pressure tracking function: $P<.001$; data synchronization: $P<.001$; blood pressure management effect: $P<.001$; heart rate detection function: $P<.001$; data sharing: $P<.001$; reliability: $P<.001$; compatibility: $P<.001$; interface design: $P<.001$; advertisement distribution: $P<.001$; measurement accuracy: $P<.001$; and cost: $P<.001$) significantly influenced the user satisfaction ND. Data synchronization ($\beta=2.662$, $P<.001$) had the greatest positive effect, while the BP management effect ($\beta=-2.035$, $P<.001$) had the most substantial negative effect.
Table 3. Determinant factors for rating deviations (Chinese reviews).

<table>
<thead>
<tr>
<th>Influencing factor</th>
<th>Model 1</th>
<th></th>
<th>Model 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β (95% CI)</td>
<td>SE</td>
<td>P value</td>
<td>SE</td>
</tr>
<tr>
<td>Topic 1: Easy to use</td>
<td>0.080 (0.043 to 0.117)</td>
<td>0.019</td>
<td>&lt;.001</td>
<td>-1.575 (-1.792 to -1.358)</td>
</tr>
<tr>
<td>Topic 2: Reliability</td>
<td>0.013 (-0.030 to 0.055)</td>
<td>0.022</td>
<td>.56</td>
<td>-0.692 (-0.937 to -0.448)</td>
</tr>
<tr>
<td>Topic 3: Measurement accuracy</td>
<td>-0.181 (-0.224 to -0.138)</td>
<td>0.022</td>
<td>&lt;.001</td>
<td>2.789 (2.564 to 3.014)</td>
</tr>
<tr>
<td>Topic 5: Compatibility</td>
<td>-0.219 (-0.275 to -0.163)</td>
<td>0.029</td>
<td>&lt;.001</td>
<td>3.170 (2.906 to 3.434)</td>
</tr>
<tr>
<td>Topic 6: Cost</td>
<td>-0.232 (-0.293 to -0.171)</td>
<td>0.031</td>
<td>&lt;.001</td>
<td>3.703 (3.424 to 3.981)</td>
</tr>
<tr>
<td>Topic 7: Heart rate detection function</td>
<td>-0.073 (-0.132 to -0.014)</td>
<td>0.030</td>
<td>.02</td>
<td>0.077 (-0.250 to 0.403)</td>
</tr>
<tr>
<td>Topic 8: Blood pressure tracking function</td>
<td>0.354 (0.298 to 0.411)</td>
<td>0.029</td>
<td>&lt;.001</td>
<td>-0.239 (-0.556 to 0.078)</td>
</tr>
<tr>
<td>Topic 9: Interface design</td>
<td>0.082 (0.019 to 0.144)</td>
<td>0.032</td>
<td>.01</td>
<td>-1.619 (-1.934 to -1.304)</td>
</tr>
<tr>
<td>Topic 10: Real-time</td>
<td>-0.047 (-0.121 to 0.026)</td>
<td>0.038</td>
<td>.21</td>
<td>1.459 (1.091 to 1.826)</td>
</tr>
<tr>
<td>Topic 11: Data privacy</td>
<td>-0.028 (-0.113 to 0.057)</td>
<td>0.043</td>
<td>.52</td>
<td>0.761 (0.315 to 1.208)</td>
</tr>
</tbody>
</table>

aIn Chinese reviews, the topic labeled “topic attitude (positive)” pertains solely to users’ positive attitudes and does not encompass opinions regarding the functions and utility of apps. Therefore, it was not included in the Tobit regression model.
bPositive deviations: The maximum likelihood estimate of model 1 was –16521.383.
cNegative deviations: The maximum likelihood estimate of model 2 was –12843.104.

Table 4. Determinant factors for rating deviations (American reviews).

<table>
<thead>
<tr>
<th>Influencing factor</th>
<th>Model 1</th>
<th></th>
<th>Model 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β (95% CI)</td>
<td>SE</td>
<td>P value</td>
<td>SE</td>
</tr>
<tr>
<td>Topic 1: Easy to use</td>
<td>0.068 (0.058 to 0.078)</td>
<td>0.005</td>
<td>&lt;.001</td>
<td>-0.946 (-0.995 to -0.898)</td>
</tr>
<tr>
<td>Topic 2: Blood pressure tracking function</td>
<td>0.312 (0.301 to 0.323)</td>
<td>0.006</td>
<td>&lt;.001</td>
<td>-1.139 (-1.195 to -1.083)</td>
</tr>
<tr>
<td>Topic 3: Data synchronization</td>
<td>-0.593 (-0.606 to -0.581)</td>
<td>0.006</td>
<td>&lt;.001</td>
<td>2.662 (2.612 to 2.711)</td>
</tr>
<tr>
<td>Topic 4: Blood pressure management effect</td>
<td>0.247 (0.235 to 0.260)</td>
<td>0.006</td>
<td>&lt;.001</td>
<td>-2.035 (-2.100 to -1.970)</td>
</tr>
<tr>
<td>Topic 5: Heart rate detection function</td>
<td>-0.201 (-0.214 to -0.188)</td>
<td>0.007</td>
<td>&lt;.001</td>
<td>0.807 (0.750 to 0.863)</td>
</tr>
<tr>
<td>Topic 6: Data sharing</td>
<td>-0.150 (-0.165 to -0.135)</td>
<td>0.008</td>
<td>&lt;.001</td>
<td>0.707 (0.642 to 0.771)</td>
</tr>
<tr>
<td>Topic 7: Reliability</td>
<td>-0.269 (-0.283 to -0.255)</td>
<td>0.007</td>
<td>&lt;.001</td>
<td>1.356 (1.295 to 1.417)</td>
</tr>
<tr>
<td>Topic 8: Compatibility</td>
<td>-0.233 (-0.252 to -0.214)</td>
<td>0.010</td>
<td>&lt;.001</td>
<td>1.467 (1.394 to 1.540)</td>
</tr>
<tr>
<td>Topic 9: Interface design</td>
<td>0.155 (0.138 to 0.171)</td>
<td>0.008</td>
<td>&lt;.001</td>
<td>-1.235 (-1.324 to -1.146)</td>
</tr>
<tr>
<td>Topic 10: Advertisement distribution</td>
<td>-0.577 (-0.599 to -0.555)</td>
<td>0.011</td>
<td>&lt;.001</td>
<td>2.644 (2.565 to 2.723)</td>
</tr>
<tr>
<td>Topic 11: Measurement accuracy</td>
<td>0.039 (0.019 to 0.058)</td>
<td>0.010</td>
<td>&lt;.001</td>
<td>-0.597 (-0.694 to -0.499)</td>
</tr>
<tr>
<td>Topic 12: Cost</td>
<td>0.217 (0.192 to 0.243)</td>
<td>0.013</td>
<td>&lt;.001</td>
<td>-0.664 (-0.920 to -0.408)</td>
</tr>
</tbody>
</table>

aPositive deviations: The maximum likelihood estimate of model 3 was –81779.181.
bNegative deviations: The maximum likelihood estimate of model 4 was –102855.05.

Motivational and Hygienic Factor Asymmetry

The effect asymmetry results are presented in Tables S12 and S13 in Multimedia Appendix 1. In models 1 and 2, with the exception of the heart rate detection and BP tracking functions, the remaining 8 factors exhibited significant differences in asymmetric effect (easy to use: P<.001; reliability: P<.001; measurement accuracy: P<.001; compatibility: P<.001; cost: P<.001; interface design: P<.001; real-time: P<.001; and data privacy: P=.001). In models 3 and 4, all 12 influencing factors demonstrated significant differences (easy to use: P<.001; blood pressure tracking function: P<.001; data synchronization: P<.001; blood pressure management effect: P<.001; heart rate detection function: P<.001; data sharing: P<.001; reliability: P<.001; compatibility: P<.001; interface design: P<.001; advertisement distribution: P<.001; measurement accuracy: P<.001; and cost: P<.001). Consequently, all factors exhibited a significant asymmetric effect on user satisfaction with Chinese (easy to use: P<.001; reliability: P<.001; measurement accuracy: P<.001; compatibility: P<.001; cost: P<.001; interface design:
Regarding the influencing factors of user satisfaction with Chinese HMAs, both ease of use ($P<0.001$ for both) and interface design ($P=0.01$ and $P<0.001$, respectively) had significant positive or negative effects on the PD or ND models, respectively. In addition, the effects on the 2 models were significantly different ($P<0.001$). Similarly, measurement accuracy ($P<0.001$ for both), compatibility ($P<0.001$ for both), and cost ($P<0.001$ for both) all exhibited significant positive or negative impacts on the PD or ND model. Moreover, there was a significant difference in the effect on the 2 models ($P<0.001$). Data synchronization ($P<0.001$ for both), heart rate detection function ($P<0.001$ for both), data sharing ($P<0.001$ for both), reliability ($P<0.001$ for both), compatibility ($P<0.001$ for both), and advertisement distribution ($P<0.001$ for both) each had a significant negative or positive impact on the PD or ND model. Furthermore, the impact on the 2 models was significantly different ($P<0.001$).

### Discussion

#### Principal Findings

To the best of our knowledge, this study represents the first quantitative analysis of user satisfaction, influencing factors, and asymmetry of HMA-related factors based on user reviews. By encompassing a large number of apps and user samples, our research achieved high credibility at a low cost, rendering the findings highly generalizable. Furthermore, leveraging the sociotechnical model widely used in HIT evaluation, we conducted a comparative analysis of Chinese and American HMAs, elucidating differences in user satisfaction and influencing factors between Chinese and American user groups. Consequently, we offered targeted improvement suggestions based on our findings. Although the numbers of Chinese HMAs and reviews were lower than those of American apps, user satisfaction was higher with Chinese HMAs. Furthermore, the main factors influencing user satisfaction and dissatisfaction with Chinese HMAs were the BP tracking function and cost, respectively. Conversely, the main factors affecting user satisfaction and dissatisfaction with American HMAs were the BP tracking function and data synchronization, respectively. Regarding the asymmetry of influencing factors, all factors exhibited significantly different effects on user satisfaction and dissatisfaction. Moreover, there were notable disparities in the motivational and hygienic factors between Chinese and American HMA users.

#### Differences in the Use of Chinese and American HMAs

Globally, users are distributed unevenly, with considerably lower usage of Chinese HMAs compared with American HMAs. As HMAs represent a typical form of digital health software, their usage status can be effectively analyzed using sociotechnical models [25], which are widely used in HIT. External environmental factors, such as medical policies, diagnosis and treatment methods, and payment methods in different countries, serve as major determinants of HIT. These factors may explain the significant disparities in the use of Chinese and American HMA. Regarding medical policies and treatment methods, Chinese patients tend to rely more on hospital-based doctors, while American patients often prefer active health management guided by family doctors. In China, a 3-level accreditation system for general hospitals has been implemented, and diagnosis and treatment modalities primarily revolve around hospital visits. This has resulted in the accumulation of many patients in a small number of tertiary hospitals with high-quality medical resources for offline diagnosis and treatment [57]. Consequently, personal active health management is rarely implemented. However, hierarchical diagnosis and treatment systems and family doctor consultation models have been primarily implemented in American countries [58]. In the American context, family doctors, who constitute around 80% of doctors, are responsible for 80%–90% of primary diagnosis and treatment services [59]. These services primarily entail disease prevention through active health management. Contrastingly, China relies predominantly on traditional hospitals and drug payments. As a result, Chinese hospitals primarily prioritize in-hospital drug efficacy and often overlook out-of-hospital patient management. In addition, limited software copyright protection in China restricts profits for HMA developers from software downloads. Consequently, the substantial economic costs have impeded the development of HMAs in China. However, American countries primarily implement medical value-based payment policies. Medical insurance payers prioritize patient rehospitalization rates, imposing fines on medical institutions for frequent patient rehospitalizations. This has prompted American medical institutions to use digital health apps to manage patients after discharge and monitor their health status in real-time, aiming to reduce rehospitalization rates. Furthermore, in the United States, HMAs can be prescribed to patients by family doctors or specialists through digital therapy prescriptions. Moreover, the software is granted patent rights, and app developers can cover app development–related costs through software downloads or paid functions, thereby promoting the continuous upgrading and iteration of HMAs.

#### Satisfaction and Focus of Chinese and American MHA Users

The overall user satisfaction with HMAs was generally poor, with Chinese HMAs exhibiting higher user satisfaction compared with American HMAs. The functions, software design, technical stability, and cost of HMAs were common areas of concern for both Chinese and American users, although
their specific focuses differed. The disparity in user satisfaction between Chinese and American HMA users was consistent with the topic clustering results of user reviews, with only Chinese HMA user reviews containing topics reflecting positive user attitudes. Furthermore, the overall user satisfaction obtained in this study, using big data generated by real-world users, was lower than that reported by previous small-sample surveys [60] (87,773/116,686, 75.22%, vs 93.5%). Nevertheless, HMA user engagement and intention to continue use were low [19,20], indicating that our study results are more reliable than those of previous small-sample questionnaire surveys.

Regarding qualitative user descriptions, both Chinese and American users expressed concerns about hypertension management functions, software design, technical stability, and costs. Specifically, most users were particularly concerned about hypertension management functions such as BP tracking and heart rate detection, including BP and heart rate measurements, BP recording, and the visual display of BP change trends. Software design, encompassing usability and interface design, was also a primary concern among users. Usability emerged as the topic of greatest concern among both Chinese (6863/16,561, 41.44%) and American (34,443/100,125, 34.40%) users. Attributes such as simple operating procedures, convenient usage environments, excellent interface design, and clear information display were highlighted as factors contributing to improved user satisfaction. Discussions on technical stability primarily revolved around software reliability, including issues such as software crashes and the inability to open software, as well as system compatibility problems such as software version mismatches and difficulties connecting via Bluetooth. In addition, concerns were raised about measurement accuracy. Finally, users also expressed concerns about software costs, including whether the software is free, its affordability, and the possibility of refunds.

However, there was a significant difference in the focus of HMAs between Chinese and American user groups. Chinese users primarily paid attention to technological stability, such as software reliability and measurement accuracy. By contrast, American users were more concerned about specific hypertension management functions. Moreover, Chinese and American users exhibited different concerns regarding software functions and design. Chinese users showed more interest in the real-time software monitoring function, while American users focused more on data synchronization, data sharing, and BP management effects. Furthermore, Chinese users mentioned software design and personalization, whereas American users were more inclined to allow advertisements that did not disrupt normal app use.

### Motivation and Hygiene Factors of Chinese and American HMAs

Easy-to-use features and interface design, including simple software operation and convenience, were common motivational factors for both Chinese and American HMA users. Compatibility issues, such as mismatched software versions and the inability to connect via Bluetooth, were common hygienic factors. However, other motivational and hygienic factors differed considerably between the 2 user groups. Tobit model analysis and the Wald test revealed influencing factors with significantly different effects on user satisfaction PD and ND, indicating asymmetric impacts of related factors on Chinese and American HMA user satisfaction. This suggests the presence of both motivational and hygienic factors in influencing user satisfaction.

Simple operating procedures and convenient HMA usage effectively improve user satisfaction. However, other motivational factors had significantly different effects on user satisfaction. In addition to the motivational factors for Chinese users, American users prioritized factors such as the BP tracking function, BP management effect, measurement accuracy, and cost-effectiveness. The presence of these influencing factors in user reviews increased the degree of user satisfaction, while their absence had the opposite effect. Hence, these factors were identified as the main contributors to user satisfaction, although user dissatisfaction was less associated with them. Therefore, according to the 2-factor theory, the aforementioned influencing factors were considered motivational factors for HMA user satisfaction. Enhancing these factors can lead to increased user satisfaction and, consequently, strengthen users’ intention to continue using the HMA.

Addressing issues such as software version mismatches and the inability to connect via Bluetooth can effectively mitigate user dissatisfaction. However, other hygienic factors had varying effects on user satisfaction. In addition to measurement accuracy and cost, hygienic factors for Chinese users encompassed aspects such as data sharing (ie, data uploadable to family doctors). By contrast, hygienic factors for American users included reliability, data synchronization, advertisements that did not affect normal use, and heart rate detection functions. The presence of the aforementioned influencing factors in user reviews increased the degree of user dissatisfaction, and vice versa. Hence, while users feel dissatisfied when these factors do not meet their expectations, their satisfaction is not significantly affected. Therefore, according to the 2-factor theory, the aforementioned influencing factors were considered motivational factors for HMA user satisfaction. Improving these factors can help mitigate user dissatisfaction, thereby enhancing user participation.

### Suggestions for Improving User Engagement and Continued Use of HMA

Given the considerable differences in motivational and hygienic factors for HMA use between Chinese and Americans, software developers should tailor improvements to hygienic factors based on the specific needs of each user group. This approach can effectively reduce user dissatisfaction and increase user participation. Furthermore, efforts should be directed toward enhancing motivational factors to improve user satisfaction and foster continued use intention among both Chinese and American users. Considering that hygiene factors have a direct impact on the utilization of HMA, it is advisable for developers to prioritize improving these factors to enhance user participation. Once hygiene factors are satisfactorily addressed, developers can then focus on enhancing motivational factors to increase users’ willingness to continue using the HMA. By improving corresponding motivational and hygienic factors tailored to different user groups, developers can facilitate the
adoption and utilization of HMA, thereby aiding patients in effectively controlling BP. For Chinese HMAs, developers should prioritize enhancing the accuracy of indicator measurements by improving the sensitivity of measurement sensors and optimizing software measurement algorithms. This targeted improvement can effectively mitigate user abandonment of Chinese HMAs. In addition, HMA developers should promptly address technical issues related to software compatibility, such as mismatches between phone systems and software versions, as well as Bluetooth connectivity issues. Moreover, given that Chinese users are more price-sensitive, high software charges have resulted in the loss of a significant number of users. Therefore, developers should consider implementing pricing strategies that align with the preferences and financial capabilities of Chinese users to mitigate user attrition. Developers should consider reducing the number of charging items, shifting away from the traditional model of charging for software downloads, and exploring new business models to optimize profits. For instance, HMAs could be bundled with BP monitors or other relevant health devices to provide added value to users. Furthermore, government intervention is crucial to support the growth of digital medical companies and provide funding for the development of mHealth solutions, especially those focused on chronic disease management. Policy support can encourage entrepreneurship in the digital health sector and foster innovation in HMA development. Finally, HMA developers can enhance user satisfaction by focusing on improving app usability and interface design. This could involve streamlining the operation process, providing clear operation guides, and enhancing the overall user-friendliness of the software. By prioritizing these aspects, developers can create a more intuitive and enjoyable user experience, ultimately increasing user satisfaction and engagement with the HMA.

For American HMAs, developers should promptly address software compatibility issues to enhance HMA usage. This includes resolving mismatches between wearable devices and software, as well as fixing Bluetooth connection failures. In addition, HMA developers should consider incorporating health data sharing and export functions to fulfill the fundamental requirements of American users. Furthermore, developers should optimize HMA reliability and promptly address technical issues, such as software crashes and start-up failures. In addition, providing data synchronization functionality for wearable device monitoring software is essential to ensure real-time BP tracking. Furthermore, developers should consider developing additional heart rate detection functionality to meet users’ needs for tracking heart rate indicators. In addition, HMA developers should reconsider the placement and frequency of advertisements to avoid disrupting routine HMA use. Finally, in addition to enhancing the satisfaction of American HMA users, developers can improve the effectiveness of BP management within the software, enhance indicator measurement accuracy, optimize interface design, and establish a reasonable charging model. For instance, they could refine BP measurement algorithms and validate measurement results against those obtained from a BP monitor. They can incorporate additional BP management features such as exercise and diet management, BP warnings, and visual displays of BP trends. Enhancing the interface aesthetics and integrating HMAs into medical insurance schemes are also crucial measures to consider.

**Advantages Compared With Previous Research**

This study marks a significant advancement by exploring user satisfaction and its influencing factors on HMAs based on real-world user reviews. Unlike previous research that primarily relied on on-site surveys and qualitative analysis, often gathering subjective comments through interviews or questionnaires, this study leverages real user feedback obtained from online reviews. By using advanced computational methods such as natural language processing and topic modeling, it provides a more comprehensive and data-driven analysis of user satisfaction factors. For instance, Breil et al [61] discovered, through questionnaire surveys, that both patients and doctors generally accept HMAs, with expected performance being a crucial determining factor. Similarly, Kang et al [30] developed an HMA and assessed user satisfaction using a scale. However, our study adopts a big data-driven approach to quantitatively explore the influencing factors of user satisfaction in HMAs. Using the LDA model, we mined user perspectives from user-generated content and established a connection between user perspectives and satisfaction through the Tobit model, rendering the results more objective and reliable. For practical applications, previous studies have primarily focused on user groups in specific regions, overlooking the exploration of user satisfaction asymmetry. For instance, Hui et al [62] evaluated the functional availability and user satisfaction of HMAs in the United Kingdom, highlighting the need for further improvement. Meanwhile, Melin et al [63] developed an app user satisfaction evaluation scale and assessed app user satisfaction using linear weighting methods. This study compared user satisfaction and influencing factors of HMAs in China and the United States, applying the 2-factor theory to analyze user satisfaction asymmetry. This provided deeper insights into the attributes of influencing factors and offered more accurate improvement suggestions.

**Limitations**

This study had several limitations. First, due to variations in medical policies, diagnoses, and treatment models between China and the United States, Chinese individuals use HMAs less frequently, resulting in a proportionately lower number of Chinese reviews and downloads (16,561/24,204,832, 0.07%) compared with American ones (100,125/148,869,181, 0.07%). Therefore, the number of collected Chinese HMA user reviews was significantly lower than that for the American apps. However, for the representativeness of user reviews, we did not conduct special sampling but obtained all user reviews. These user reviews were deemed sufficient to reflect user satisfaction and the influencing factors. Second, considering that some users used the HMA but did not leave reviews, there may have been bias in population selection, and we were unable to explore the satisfaction of such users. However, given that 116,686 reviews were included, and the process of user reviews is random, with users of various opinions possibly not leaving reviews, we believe our results are representative and provide useful data for discovering factors and attributes associated with HMA user satisfaction. Third, through user reviews, this study identified
the factors influencing user satisfaction and explored their asymmetry. However, an in-depth analysis of the impact paths could be further explored based on these influencing factors. Finally, through data screening and aggregation, we observed that the number of Chinese HMA user reviews was substantially lower than that of American HMA user reviews, and the factors affecting user satisfaction in China and the United States were considerably different. Although we provided a preliminary analysis of the reasons for these differences through a sociotechnical model, a more in-depth analysis is required in the future. In addition, as users are more concerned about the therapeutic effect of HMAs, it would be meaningful to pair clinical efficacy data with satisfaction in future research to further explore the relationship between user satisfaction and specific clinical efficacy.

Conclusions
Our study reveals that only 87,773/116,686 (75.22%) users are satisfied with HMA use. We also found that the factors influencing Chinese and American HMA user satisfaction exhibit asymmetry. Furthermore, because of differences in user groups and macro usage environments, the motivational and hygienic factors for users in China and the United States are significantly different. Thus, to enhance user participation, developers of HMAs should devise personalized and comprehensive strategies that address issues related to hygienic factors as a priority. Furthermore, efforts should be made to enhance motivational factors to encourage sustained HMA usage.

Acknowledgments
This work was supported by the National Natural Science Foundation of China (grant 81871455), the Zhejiang Provincial Natural Science Foundation of China (grant LY22H180001), the Municipal Natural Science Foundation of Beijing (grant 7222306), the National TCM Innovation Team and Talent Support Projects (grant ZYYCXTD-C-202210), the Key Research and Development Program of Zhejiang Province (grant 2022C03111), the Hainan Province Science and Technology special fund (grant ZDYF2022SHFZ292), the Hainan Province Clinical Medical Center (grant QWYH2022341), and The 13th Five Year Education Reform Project in Zhejiang Province (grant JG20190551).

Data Availability
We used crawler technology and Qimai app data official website to obtain all hypertension management apps and user reviews in 8 Chinese app stores (China Apple App Store, Huawei App Store, Xiaomi App Store, OPPO App Store, VIVO App Store, Baidu App store, 360 App store, and Application Treasure App Store) and 2 American app stores (US Apple App Store and US Google Play Store). The source code for our analysis is publicly available on GitHub [64].

Authors' Contributions
J Liang conceptualized and designed the study, performed data screening, and made significant revisions to the manuscript. YFH drafted the manuscript; summarized the literature review; and performed the data acquisition, screening, analysis, and interpretation. J Lei supervised the data filtering and analysis process. J Liang and WZ supervised the review method and data interpretation and provided valuable suggestions for improvement. HC, JX, TW, and YCL critically revised the manuscript for important intellectual content.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Search terms and results; inclusion and exclusion criteria; annotation and examples of reviews (in both Chinese and English). We also present the results of multicollinearity analysis.

References
3. Alnooh G, Alessa T, Hawley M, de Witte L. The use of dietary approaches to stop hypertension (DASH) mobile apps for supporting a healthy diet and controlling hypertension in adults: systematic review. JMIR Cardiol 2022 Nov 02;6(2):e35876. [FREE Full text] [doi: 10.2196/35876] [Medline: 36322108]


https://mhealth.jmir.org/2024/1/e55199


39. He et al. JMIR MHEALTH AND UHEALTH 2024 | vol. 12 | e55199 | p.449


49. Natural language toolkit. NLTK Project. URL: https://www.nltk.org/ [accessed 2023-09-03]

50. The most comprehensive NLP Chinese and English stop list for the entire site. CSDN. URL: https://blog.csdn.net/qq_52181283/article/details/125013666 [accessed 2023-12-18]


52. Rehurek R, Sojka P. Gensim Python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University. 2011. URL: https://code.activestate.com/pypi/gensim/ [accessed 2023-12-10]


56. Stata is statistical software for data science. Stata. URL: https://www.stata.com [accessed 2023-09-03]


Abbreviations

BP: blood pressure
FDA: Food and Drug Administration
HIT: health information technology
HMA: hypertension management app
LDA: latent Dirichlet allocation
mHealth: mobile health
ND: negative deviation
PD: positive deviation
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

https://mhealth.jmir.org/2024/1/e55199

JMIR Mhealth Uhealth 2024 | vol. 12 | e55199 | p.450

(page number not for citation purposes)
©Yunfan He, Wei Zhu, Tong Wang, Han Chen, Junyi Xin, Yongcheng Liu, Jianbo Lei, Jun Liang. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 28.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
The Real-World Impact of App-Based Mindfulness on Headspace Members With Moderate and Severe Perceived Stress: Observational Study

Christine Callahan, PhD; Justin Kimber, PhD; Emily Hu, PhD; Leah Tanner, MSPH; Sarah Kunkle, MPH
Headspace, Santa Monica, CA, United States

Corresponding Author:
Christine Callahan, PhD

Abstract

Background: Perceived stress in the United States has drastically increased since the COVID-19 pandemic and is associated with negative mental health outcomes such as depression and anxiety. Digital mental health (DMH) interventions are efficacious tools to address negative mental health outcomes and have helped reduce the severity of psychological symptoms, such as anxiety, depression, and perceived stress, compared to waitlist controls. Although DMH tools have been studied in controlled settings, less is known about the real-world evidence of such interventions.

Objective: This study aimed to (1) characterize patterns in baseline perceived stress and changes in perceived stress among Headspace members with moderate and severe baseline perceived stress and (2) examine associations between engagement with Headspace content and changes in perceived stress (ie, evaluate whether there is a dose-response relationship).

Methods: We evaluated real-world perceived stress and engagement data at 2 time points among Headspace app members with baseline moderate and severe perceived stress. Perceived stress was measured using the Perceived Stress Scale (PSS-10) and engagement using active days and active minutes engaged with Headspace as well as the number of user sessions. Descriptive statistics were computed for all variables. Correlations between baseline and follow-up scores, percent change in PSS-10 scores, days between PSS-10 use, active days, active days per week, active minutes, active minutes per day, sessions, and sessions per week were evaluated. We used t tests to investigate differences in the abovementioned parameters between (1) participants who did and those who did not see improvements in PSS-10 scores (yes vs no improvement) and (2) participants who saw ≥30% improvement versus those who saw a <30% improvement in PSS-10 scores.

Results: Overall, 21,088 Headspace members were included in these analyses. On average, members saw a 23.52% decrease in PSS-10 scores from baseline to follow-up. On average, members had 2.42 (SD 1.76) active days per week and 25.89 (SD 33.40) active minutes per day, and completed 7.11 (SD 8.34) sessions per week. t tests suggest that members who saw improvements in PSS-10 scores from baseline to follow-up had significantly higher baseline PSS-10 scores (Cohen d=0.56), more active days per week (Cohen d=0.33), and more sessions per week (Cohen d=0.27) than those who did not see improvements in PSS-10 scores (all P<.001). Additional t tests suggest that members with ≥30% improvement in PSS-10 scores had significantly higher baseline PSS-10 scores (Cohen d=0.35), more active days per week (Cohen d=0.36), and more sessions per week (Cohen d=0.31) than those with a >30% improvement (all P<.001).

Conclusions: Real-world use of Headspace is associated with decreased perceived stress. Furthermore, data suggest that more engagement, specifically weekly active days and sessions, is associated with a greater likelihood of stress reduction.

(JMIR Mhealth Uhealth 2024;12:e52968) doi:10.2196/52968

KEYWORDS
digital mental health; meditation; real word evidence; mental health; app; application; stress; observational study; depression; anxiety; tool; tools; psychological; psychological symptom; engagement; stress reduction; PSS-10

Introduction

Stress in the United States has increased, with significant impacts from the COVID-19 pandemic. While the pandemic itself caused a drastic increase in stress from 2019 to 2020 [1], stress levels still remain high and continue to impact the majority of Americans [2]. Higher perceived stress is associated with poorer mental health outcomes such as depression and anxiety [3]. Further meta-analytic research indicates small to medium effects of the relationship between perceived stress and mental health outcomes, suggesting that peoples’ appraisal of stressful situations in their lives might be a predictor of mental health outcomes [4,5]. Higher perceived stress also has economic costs, and studies have estimated that perceived stress, primarily work-related perceived stress, accounts for US $221 million to US $187 billion in both direct (eg, medical) and indirect costs.
(eg, absenteeism, burnout, and decreased productivity) [6]. This is important in the larger context of mental health costs and outcomes and how people manage stressors daily in their lives. Digital mental health (DMH) interventions may provide an accessible, scalable way to mitigate perceived stress.

DMH interventions are scalable and accessible; incorporate evidence-based practices (ie, mindfulness meditation and cognitive behavioral therapy techniques); are efficacious for a range of mental health concerns including anxiety, depression, and posttraumatic stress disorder [3]; and have shown to be effective in a wide range of populations with all levels of mental health concerns (ie, mild, moderate, and severe), including college students, employees, graduate trainees, and rural communities [7-9]. Studies suggest that digital mindfulness-based interventions significantly reduce perceived stress [10-12]. Specifically, a randomized controlled trial evaluated an app-based mindfulness tool among individuals with moderate and severe baseline perceived stress and suggested a 30.12% decrease in perceived stress from baseline to 8 weeks (intervention completion), with these reductions retained at 12-week follow-up (a 31.24% decrease) [12]. In addition to randomized controlled trials, current meta-analytic and systematic review data suggest medium effect sizes for DMH interventions for perceived stress from baseline to post intervention [4,5]. Beyond clinical outcomes, DMH interventions improve access to mental health care and provide individuals with effective, cost-effective care readily available via a mobile app or website.

While DMH interventions in clinical trials have been shown to be effective, less is known about their use in real-world settings. Real-world evidence builds upon clinical trials to improve our understanding of an intervention’s efficacy in a person’s daily functioning, providing data on the effectiveness and accessibility of DMH products as well as the external validity of interventions. As DMH interventions are delivered via a mobile device or computer to use within one’s own environments, real-world evidence highlights intervention effectiveness outside of controlled settings. As such, there is an increasing need to understand the real-world effectiveness of widely used commercial apps and DMH interventions that help improve mental health outcomes for the overall population.

Headspace is a popular and evidence-based DMH platform with over 100 million downloads and supported by >50 published peer-reviewed studies. The Headspace app offers a range of services, most notably to this study, mindfulness and meditation-based content that teaches coping strategies to manage daily stressors. Previous clinical trials on the efficacy of the Headspace app show evidence of improved mindfulness, focus, stress, sleep quality, burnout, resilience, anxiety, depression, and quality of life. Given the reach and scale of Headspace’s membership, there is an opportunity to better understand real-world outcomes beyond clinical trials. As such, this study aims to use real-world data to evaluate perceived stress among Headspace members. To accomplish this goal, this study’s aims are to (1) characterize patterns in baseline perceived stress and changes in perceived stress among Headspace members with moderate and severe baseline perceived stress and (2) examine associations between engagement with Headspace content and changes in perceived stress (ie, to evaluate if there is a dose-response relationship).

Methods

Study Design and Participants

This real-world observational study examined perceived stress among Headspace members. This study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) framework for reporting observational studies [13]. Individuals were included in this study if they enrolled on Headspace between March 2020 and January 2023, completed a baseline and follow-up perceived stress questionnaire (the Perceived Stress Scale [PSS-10]), completed the baseline PSS-10 within 90 days of enrollment, completed the follow-up PSS-10 at >7 and <60 days from the baseline, and reported moderate or severe perceived stress levels at baseline (Figure 1).
The Headspace App

Study participants had access to all Headspace app offerings and used the app at their own discretion. The Headspace app includes evidence-based, expert-led content including guided mindfulness, meditation with animated guided videos, progressive muscle relaxation, psychoeducation, and gratitude exercises. Mindfulness content ranges from content for general self-care to that specific to mental health disorders such as anxiety, stress, and sleep disorders. When participants open the Headspace app, they are first directed to the “Today” tab, which contains personalized, daily content recommendations to encourage health habit formation throughout the day. Content on the “Today” tab includes a breathing exercise, educational video, and 3 meditations (1 each for the morning, afternoon, and evening). All other content can be found in the “Explore” tab, which includes a search bar; content organized into large categories of meditate, sleep, move, and music at the top; and as they scroll further, more specific categories and courses such as Beginning Meditation, Mindfulness at Work, Mindful Eating, etc. Finally, all participants have a personalized “Profile” tab, which tracks activity history and overall app statistics (minutes meditated, sessions completed, and days in a row of content engagement). Study participants had access to all Headspace app offerings and navigated the app to choose their own content (Figure 2).
**Ethical Considerations**

This study represents a secondary analysis of preexisting deidentified data [14]. The study team does not have access to the participants’ identifying information and will not contact the participants. Therefore, in accordance with the US Department of Health and Human Services’ regulations (45 CRF 46.104), this study is deemed exempt from institutional review board oversight and informed consent. Participants are notified through the Headspace terms of service that their deidentified data may be used for research.

**Measures**

Perceived stress was measured using the PSS-10 [15]. Participants were prompted to complete the PSS-10 through the Headspace app and answered 10 questions about their perceived feelings and thoughts regarding stress with a 5-point Likert scale (0=never to 4=very often). Scores range from 0 to 40, with higher scores indicating higher perceived stress. The scale has high test-retest reliability (Cronbach $\alpha=.85$). PSS-10 scores were categorized to indicate low (<14), moderate (14-26), and high (27-40) levels of perceived stress [15]. Those with low levels of perceived stress at baseline were excluded from this study. Study participants completed a baseline and follow-up PSS-10. Total scores for each time point were reported, along with the raw change in PSS-10 score and percent change in PSS-10 score. Participants completed the PSS-10 within the Headspace app, and final scores were extracted directly from the app. Scores were calculated and extracted only for those
who completed all questions on the PSS-10 (ie, incomplete data were excluded from this study).

Engagement was assessed using active days and minutes engaged with Headspace as well as the number of sessions started. Additionally, ratios for active days per week, active minutes per day, and sessions per week were calculated to further define engagement. Engagement data are tracked for all members within the Headspace app and were extracted directly from the app.

**Statistical Analysis**

Means, SDs, medians, and 95% CIs were computed for baseline and follow-up PSS-10 scores, raw change in PSS-10 score, percent change in PSS-10 score, days between PSS-10 assessments, active days, active days per week, active minutes, active minutes per day, sessions, and sessions per week. Frequencies and percentages were computed for baseline and follow-up PSS-10 scores (moderate vs high).

Mean percent change in PSS-10 scores was segmented out by active days per week, active minutes per day, and sessions per week to delineate the relationship between app engagement and percent change in PSS-10 score. Correlations between baseline and follow-up PSS-10 scores, percent change in PSS-10 score, days between PSS-10 assessments, active days, active days per week, active minutes, active minutes per day, sessions, and sessions per week were evaluated.

We used *t*-tests to evaluate the differences in baseline PSS-10 scores, follow-up PSS-10 scores, percent change in PSS-10 score, active days, active days per week, active minutes, active minutes per day, sessions, and sessions per week between those who saw and those who did not see improvements in PSS-10 scores (yes vs no improvement) and (2) those who saw a ≥30% improvement versus those who saw a <30% improvement. A 30% improvement threshold was chosen to coincide with a clinically meaningful change in perceived stress [16]. *P* values of <.05 were considered significant, and effect sizes were reported using the Cohen *d* to determine the magnitude of significance.

**Results**

**Sample and Perceived Stress Descriptives**

Overall, 344,544 Headspace members completed 2 PSS-10s between March 2020 and January 2023. Of them, 21,088 met this study’s inclusion criteria and were included as study participants (Figure 1).

At baseline, 15,127 (71.73%) participants reported moderate levels of perceived stress and 5961 (28.27%) reported severe levels of perceived stress with a mean baseline PSS-10 score of 23.14 (SD 5.69). On average, members completed their follow-up PSS-10 33.25 (SD 11.17) days after their first PSS-10 with a mean follow-up PSS-10 score of 20.41 (SD 6.47). At follow-up, 2878 (13.56%) participants reported mild to moderate levels of perceived stress, 14,376 (68.17%) reported moderate levels, and 3834 (18.18%) reported high levels. On average, members saw a 23.52% decrease in PSS-10 scores from baseline to follow-up, and 13,692 (64.93%) participants saw a decrease in their PSS-10 score (Table 1).

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Mean (SD)</th>
<th>Median</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PSS-10</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>23.14 (5.69)</td>
<td>23</td>
<td>23.06 to 23.21</td>
</tr>
<tr>
<td>Follow-up</td>
<td>20.41 (6.47)</td>
<td>20</td>
<td>20.33 to 20.50</td>
</tr>
<tr>
<td>Raw change in the PSS-10 score</td>
<td>−2.72 (6.06)</td>
<td>−3</td>
<td>−2.80 to −2.64</td>
</tr>
<tr>
<td>Percent change in the PSS-10</td>
<td>−23.52 (62.60)</td>
<td>−13.04</td>
<td>−24.37 to −22.68</td>
</tr>
<tr>
<td>score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Days between PSS-10 assessments</td>
<td>33.25 (11.17)</td>
<td>31</td>
<td>33.10 to 33.40</td>
</tr>
<tr>
<td><strong>Engagement metrics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active days</td>
<td>18.04 (11.42)</td>
<td>17</td>
<td>17.89 to 18.20</td>
</tr>
<tr>
<td>Active days per week</td>
<td>2.42 (1.76)</td>
<td>2.01</td>
<td>2.39 to 2.44</td>
</tr>
<tr>
<td>Active minutes</td>
<td>547.4 (1004.81)</td>
<td>281.95</td>
<td>533.83 to 560.96</td>
</tr>
<tr>
<td>Active minutes per day</td>
<td>25.89 (33.40)</td>
<td>16.68</td>
<td>25.44 to 26.34</td>
</tr>
<tr>
<td>Sessions</td>
<td>49.73 (48.14)</td>
<td>35</td>
<td>49.08 to 50.38</td>
</tr>
<tr>
<td>Sessions per week</td>
<td>7.11 (8.34)</td>
<td>4.33</td>
<td>7.00 to 7.23</td>
</tr>
</tbody>
</table>

*aPSS-10: Perceived Stress Scale.*
Engagement

On average, members engaged with Headspace content for 18.08 (SD 11.42) days and 547.40 (SD 1004.81) minutes and started 49.73 (SD 48.14) sessions. On average, members had 2.42 (SD 1.76) active days per week, 25.89 (SD 33.40) active minutes per day, and completed 7.11 (SD 8.34) sessions per week (Table 1). The percent change in PSS-10 score by active days per week, active minutes per day, and sessions per week are presented in Table 2, with data suggesting peak changes in PSS-10 scores at 7 active days per week, 11 - 15 active minutes per day, and 19 - 20 sessions per week.

Table . Mean percent change in PSS-10 scores by active days per week, active minutes per day, and sessions per week.

<table>
<thead>
<tr>
<th>Active days per week</th>
<th>Mean percent change in PSS-10 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-21.11</td>
</tr>
<tr>
<td>2</td>
<td>-21.55</td>
</tr>
<tr>
<td>3</td>
<td>-23.48</td>
</tr>
<tr>
<td>4</td>
<td>-28.94</td>
</tr>
<tr>
<td>5</td>
<td>-23.69</td>
</tr>
<tr>
<td>6</td>
<td>-32.61</td>
</tr>
<tr>
<td>7</td>
<td>-43.24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Active minutes per day</th>
<th>Mean percent change in PSS-10 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-5</td>
<td>-19.86</td>
</tr>
<tr>
<td>6-10</td>
<td>-21.33</td>
</tr>
<tr>
<td>11-15</td>
<td>-24.41</td>
</tr>
<tr>
<td>16-20</td>
<td>-24.26</td>
</tr>
<tr>
<td>21-25</td>
<td>-23.84</td>
</tr>
<tr>
<td>26-30</td>
<td>-25.65</td>
</tr>
<tr>
<td>&gt;30</td>
<td>-21.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sessions per week</th>
<th>Mean percent change in PSS-10 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>-17.18</td>
</tr>
<tr>
<td>2-4</td>
<td>-21.75</td>
</tr>
<tr>
<td>5-6</td>
<td>-25.88</td>
</tr>
<tr>
<td>7-8</td>
<td>-23.25</td>
</tr>
<tr>
<td>9-10</td>
<td>-27.56</td>
</tr>
<tr>
<td>11-12</td>
<td>-27.89</td>
</tr>
<tr>
<td>13-14</td>
<td>-29.64</td>
</tr>
<tr>
<td>15-16</td>
<td>-32.21</td>
</tr>
<tr>
<td>17-18</td>
<td>-35.94</td>
</tr>
<tr>
<td>19-20</td>
<td>-41.71</td>
</tr>
<tr>
<td>&gt;20</td>
<td>-25.39</td>
</tr>
</tbody>
</table>

aPSS-10: Perceived Stress Scale.

Engagement and Perceived Stress

Correlations between engagement and perceived stress are reported in Table 3. The results of t tests investigating the association between engagement and perceived stress are reported in Table 4. Participants who demonstrated an improvement in perceived stress had significantly higher baseline PSS-10 scores ($t_{16,575}=-40.08; P<.001; Cohen d=0.56$) and significantly more active days ($t_{14,722}=-9.00; P<.001; Cohen d=0.13$), active days per week ($t_{17,152}=-24.18; P<.001; Cohen d=0.33$), sessions ($t_{16,515}=-11.18; P<.001; Cohen d=0.16$), and sessions per week ($t_{19,025}=-20.53; P<.001; Cohen d=0.27$) than those who did not demonstrate an improvement.

Participants with a ≥30% improvement in perceived stress had significantly higher baseline PSS-10 scores ($t_{12,794}=-23.52; P<.001; Cohen d=0.35$), more active days per week ($t_{11,401}=-23.10; P<.001; Cohen d=0.36$), and more sessions per week.
week ($t_{10.150} = -18.87; P < .001; \text{Cohen } d = 0.31$) than those with a <30% improvement. Although effect sizes suggest smaller relationships, $t$ tests also suggest that participants with a $\geq 30\%$ improvement in PSS-10 scores had significantly more active days ($t_{12.984} = -12.02; P < .001; \text{Cohen } d = 0.17$), more active minutes ($t_{13.569} = -2.61; P < .001; \text{Cohen } d = 0.04$), and more sessions ($t_{11.539} = -12.68; P < .001; \text{Cohen } d = 0.19$) than those with a <30% improvement in PSS-10 score.

### Table. Pearson correlations between PSS-10 scores and engagement outcomes.

<table>
<thead>
<tr>
<th></th>
<th>Baseline PSS-10 score</th>
<th>Follow-up PSS-10 score</th>
<th>Percent change in PSS-10 score</th>
<th>Days between PSS-10 assessments</th>
<th>Active days</th>
<th>Active days per week</th>
<th>Active minutes</th>
<th>Active minutes per day</th>
<th>Sessions</th>
<th>Sessions per week</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline PSS-10 score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r$</td>
<td>N/A&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.51</td>
<td>-0.14</td>
<td>-0.03</td>
<td>-0.12</td>
<td>0.05</td>
<td>0.01</td>
<td>0.04</td>
<td>-0.03</td>
<td>0.08</td>
</tr>
<tr>
<td>$P$ value</td>
<td>N/A</td>
<td>&lt;.001&lt;sup&gt;c&lt;/sup&gt;</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>.66</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Follow-up PSS-10 score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r$</td>
<td>0.51</td>
<td>N/A</td>
<td>0.51</td>
<td>0.03</td>
<td>-0.17</td>
<td>-0.13</td>
<td>-0.01</td>
<td>0.03</td>
<td>-0.11</td>
<td>-0.09</td>
</tr>
<tr>
<td>$P$ value</td>
<td>&lt;.001</td>
<td>N/A</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Percent change in PSS-10 score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r$</td>
<td>-0.14</td>
<td>0.51</td>
<td>N/A</td>
<td>0.03</td>
<td>-0.05</td>
<td>-0.12</td>
<td>-0.01</td>
<td>0.03</td>
<td>-0.07</td>
<td>-0.10</td>
</tr>
<tr>
<td>$P$ value</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>N/A</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Days between PSS-10 assessments</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r$</td>
<td>-0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>N/A</td>
<td>0.33</td>
<td>-0.11</td>
<td>0.12</td>
<td>-0.01</td>
<td>0.16</td>
<td>-0.13</td>
</tr>
<tr>
<td>$P$ value</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>N/A</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>.12</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Active days</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r$</td>
<td>-0.12</td>
<td>-0.17</td>
<td>-0.05</td>
<td>0.33</td>
<td>N/A</td>
<td>0.63</td>
<td>0.46</td>
<td>0.21</td>
<td>0.73</td>
<td>0.43</td>
</tr>
<tr>
<td>$P$ value</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>N/A</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Active days per week</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r$</td>
<td>0.05</td>
<td>-0.13</td>
<td>-0.12</td>
<td>-0.11</td>
<td>0.63</td>
<td>N/A</td>
<td>0.31</td>
<td>0.19</td>
<td>0.58</td>
<td>0.78</td>
</tr>
<tr>
<td>$P$ value</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>N/A</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Active minutes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r$</td>
<td>0.01</td>
<td>-0.01</td>
<td>-0.01</td>
<td>0.12</td>
<td>0.46</td>
<td>0.31</td>
<td>N/A</td>
<td>0.87</td>
<td>0.52</td>
<td>0.36</td>
</tr>
<tr>
<td>$P$ value</td>
<td>.66</td>
<td>.05</td>
<td>.18</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>N/A</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Active minutes per day</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r$</td>
<td>0.04</td>
<td>0.03</td>
<td>-0.01</td>
<td>-0.01</td>
<td>0.21</td>
<td>0.19</td>
<td>0.87</td>
<td>N/A</td>
<td>0.35</td>
<td>0.30</td>
</tr>
<tr>
<td>$P$ value</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>.97</td>
<td>.12</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>N/A</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Sessions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r$</td>
<td>-0.03</td>
<td>-0.11</td>
<td>-0.07</td>
<td>0.16</td>
<td>0.73</td>
<td>0.58</td>
<td>0.52</td>
<td>0.35</td>
<td>N/A</td>
<td>0.78</td>
</tr>
<tr>
<td>$P$ value</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>N/A</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Sessions per week</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r$</td>
<td>0.08</td>
<td>-0.09</td>
<td>-0.10</td>
<td>-0.13</td>
<td>0.43</td>
<td>0.78</td>
<td>0.36</td>
<td>0.30</td>
<td>0.78</td>
<td>N/A</td>
</tr>
<tr>
<td>$P$ value</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>N/A</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

<sup>a</sup>PSS-10: Perceived Stress Scale.

<sup>b</sup>N/A: not applicable.

<sup>c</sup>Italicized values are significant at $P<.05$.
Table. Differences in PSS-10 scores and engagement metrics between participants who saw an improvement in their PSS-10 score and those who did not and between participants who saw a $\geq 30\%$ improvement in the PSS-10 score and those who saw a $<30\%$ improvement in their PSS-10 score.

<table>
<thead>
<tr>
<th></th>
<th>PSS-10 score improvement, mean (SD)</th>
<th>No PSS-10 score improvement, mean (SD)</th>
<th>95% CI</th>
<th>$t$ test (df)</th>
<th>$P$ value</th>
<th>Cohen $d$</th>
<th>$\geq 30%$ improvement in PSS-10 score, mean (SD)</th>
<th>$&lt;30%$ improvement in PSS-10 score, mean (SD)</th>
<th>95% CI</th>
<th>$t$ test (df)</th>
<th>$P$ value</th>
<th>Cohen $d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline PSS-10 score</td>
<td>24.22 (5.69)</td>
<td>21.14 (5.12)</td>
<td>$-3.23$ to $-2.93$</td>
<td>$-40.08$ (16,576)</td>
<td>$&lt;.001$ b</td>
<td>0.56</td>
<td>24.49 (5.59)</td>
<td>22.52 (5.63)</td>
<td>$-2.12$ to $-1.80$</td>
<td>(14,364)</td>
<td>$&lt;.001$</td>
<td>0.35</td>
</tr>
<tr>
<td>Follow-up PSS-10 score</td>
<td>18.16 (5.73)</td>
<td>24.60 (5.62)</td>
<td>$6.28$ to $6.60$</td>
<td>78.87 (15,418)</td>
<td>$&lt;.001$</td>
<td>1.13</td>
<td>15.13 (4.52)</td>
<td>22.81 (4.52)</td>
<td>7.54 to 7.83</td>
<td>(14,980)</td>
<td>$&lt;.001$</td>
<td>1.42</td>
</tr>
<tr>
<td>Percent change in PSS-10 score</td>
<td>$-43.42$ (69.50)</td>
<td>13.32 (11.84)</td>
<td>$55.55$ to $57.94$</td>
<td>93.07 (15,121)</td>
<td>$&lt;.001$</td>
<td>1.01</td>
<td>$-74.19$ (90.34)</td>
<td>$-50.53$ (17.34)</td>
<td>71.45 to 75.85</td>
<td>(6682)</td>
<td>$&lt;.001$</td>
<td>1.40</td>
</tr>
<tr>
<td>Total active days</td>
<td>18.57 (11.28)</td>
<td>17.07 (11.62)</td>
<td>$-1.82$ to $-1.17$</td>
<td>$-9.00$ (14,774)</td>
<td>$&lt;.001$</td>
<td>0.13</td>
<td>19.43 (11.21)</td>
<td>17.41 (11.46)</td>
<td>$-2.35$ to $-1.69$</td>
<td>(14,236)</td>
<td>$&lt;.001$</td>
<td>0.17</td>
</tr>
<tr>
<td>Active days per week</td>
<td>2.62 (1.82)</td>
<td>2.04 (1.57)</td>
<td>$-0.63$ to $-0.53$</td>
<td>$-24.18$ (17,152)</td>
<td>$&lt;.001$</td>
<td>0.33</td>
<td>2.85 (1.88)</td>
<td>2.22 (1.66)</td>
<td>$-0.68$ to $-0.57$</td>
<td>(13,437)</td>
<td>$&lt;.001$</td>
<td>0.36</td>
</tr>
<tr>
<td>Total active minutes</td>
<td>557.59 (969.71)</td>
<td>528.52 (1066.57)</td>
<td>$-58.31$ to $-0.17$</td>
<td>$-1.95$ (13,969)</td>
<td>.05</td>
<td>0.03</td>
<td>573.54 (956.53)</td>
<td>535.53 (1025.86)</td>
<td>$-66.52$ to $-9.51$</td>
<td>(14,700)</td>
<td>.01</td>
<td>0.04</td>
</tr>
<tr>
<td>Active minutes per day</td>
<td>26.10 (32.33)</td>
<td>25.50 (35.29)</td>
<td>$-1.56$ to $0.38$</td>
<td>$-1.20$ (14,062)</td>
<td>.23</td>
<td>0.02</td>
<td>26.25 (31.60)</td>
<td>25.73 (34.18)</td>
<td>$-1.46$ to $0.43$</td>
<td>(14,540)</td>
<td>.28</td>
<td>0.01</td>
</tr>
<tr>
<td>Total sessions</td>
<td>52.36 (49.64)</td>
<td>44.85 (44.85)</td>
<td>$-8.83$ to $-6.20$</td>
<td>$-11.18$ (16,515)</td>
<td>$&lt;.001$</td>
<td>0.16</td>
<td>56.21 (51.62)</td>
<td>46.79 (46.18)</td>
<td>$-10.87$ to $-7.96$</td>
<td>(13,549)</td>
<td>$&lt;.001$</td>
<td>0.19</td>
</tr>
<tr>
<td>Sessions per week</td>
<td>7.90 (9.00)</td>
<td>5.65 (6.72)</td>
<td>$-2.47$ to $-2.04$</td>
<td>$-20.53$ (19,025)</td>
<td>$&lt;.001$</td>
<td>0.27</td>
<td>8.88 (9.82)</td>
<td>6.31 (7.44)</td>
<td>$-2.83$ to $-2.30$</td>
<td>(12,439)</td>
<td>$&lt;.001$</td>
<td>0.31</td>
</tr>
</tbody>
</table>

aPSS-10: Perceived Stress Scale.
bItalicized values are significant at $P<.05$.

Discussion

This study focused on using real-world data from Headspace members to evaluate changes in perceived stress and its association with app engagement. Our findings suggest that participants experienced a significant reduction in perceived stress scores and those who used Headspace more frequently experienced greater reductions in perceived stress. This study builds on prior clinical trials and provides real-world evidence supporting the use of Headspace to improve perceived stress.

On average, participants experienced a 23.52% reduction in perceived stress scores from baseline to the follow-up assessment (ie, approximately a 30-day period). Furthermore, 64.93% of participants saw a reduction in their PSS-10 scores from baseline to follow-up, suggesting that Headspace improved perceived stress for a majority of individuals. Shifts in perceived stress levels also provide evidence supporting the use of Headspace, as 10% of members who reported high levels of perceived stress at baseline no longer met those criteria at follow-up (ie, reported moderate or mild levels of perceived stress at follow-up). These shifts in individual perceived stress are important for self-management behaviors (eg, mindfulness or self-regulation) that are associated with improved patient well-being and mental health outcomes [17]. Thus, it is likely that shifts in perceived stress might be an important predictor of patients’ mental health outcomes. These findings are consistent with previous clinical research examining the impact of DMH interventions to improve perceived stress [10,18,19] and provide additional real-world evidence to support those claims.

We also identified relationships between app engagement and changes in perceived stress. Specifically, correlations suggest that active days per week and active sessions per week were associated with a higher percent change in PSS-10 scores. App-based mindfulness interventions of longer duration, such as those from 4 to 12 weeks, have shown improvements in not only mindfulness and perceived stress but also depression, anxiety, and overall well-being [11,20,21]. Few app-based mindfulness interventions are less than 4 weeks in duration, and provide additional real-world evidence to support those claims.

On average, participants experienced a 23.52% reduction in perceived stress scores from baseline to the follow-up assessment (ie, approximately a 30-day period). Furthermore, 64.93% of participants saw a reduction in their PSS-10 scores from baseline to follow-up, suggesting that Headspace improved perceived stress for a majority of individuals. Shifts in perceived stress levels also provide evidence supporting the use of Headspace, as 10% of members who reported high levels of perceived stress at baseline no longer met those criteria at follow-up (ie, reported moderate or mild levels of perceived stress at follow-up). These shifts in individual perceived stress are important for self-management behaviors (eg, mindfulness or self-regulation) that are associated with improved patient well-being and mental health outcomes [17]. Thus, it is likely that shifts in perceived stress might be an important predictor of patients’ mental health outcomes. These findings are consistent with previous clinical research examining the impact of DMH interventions to improve perceived stress [10,18,19] and provide additional real-world evidence to support those claims.

We also identified relationships between app engagement and changes in perceived stress. Specifically, correlations suggest that active days per week and active sessions per week were associated with a higher percent change in PSS-10 scores. App-based mindfulness interventions of longer duration, such as those from 4 to 12 weeks, have shown improvements in not only mindfulness and perceived stress but also depression, anxiety, and overall well-being [11,20,21]. Few app-based mindfulness interventions are less than 4 weeks in duration, supporting the use of longer interventions. While this study revealed that a higher number of active days (total and per week) is associated with larger improvements in perceived stress, our
data did not reveal a relationship between active minutes (total and per week) and percent change in perceived stress. These data support the findings of a previous study comparing 10 and 30 minutes of daily mindfulness, which reported no difference in mindfulness or psychological distress between the dosage groups [22]. This study’s findings, in addition to those of previous studies investigating mindfulness dosage, may suggest that active days are more important than total time spent practicing mindfulness or engaging with the app. Future research is necessary to further understand these relationships to more accurately suggest mindfulness dosage for clinical practice and research.

Identifying changes in perceived stress by active days per week, active minutes per day, and sessions per week provide more nuanced data informing the duration and dosage of mindfulness interventions. The largest percent change in perceived stress occurred with 7 active days per week; however, data show a sizable increase at 4 active days per week with a leveling off in those active for more days. We also note peak percent change in PSS-10 score at 11 - 15 active minutes per day and 19 - 20 sessions per week. These data indicate that consistent Headspace use results in more prominent improvements in perceived stress. Previous research has reported that the psychological status of approximately 25% of patients improved after 1 psychotherapy session, with steadier improvements occurring over 8 weeks [23]. This study’s findings indicate a similar pattern that emphasizes a shift in perceived stress from baseline to the second administration of the PSS-10 during the first 3 weeks of engaging with the Headspace platform.

Higher engagement, specifically more active days and sessions, was significantly associated with a higher likelihood of improving stress (both overall and ≥30%), suggesting that frequent engagement with the Headspace app might be related to improved outcomes. This is an important finding given the breadth of research on increased patient engagement with psychotherapy and improved mental health outcomes [24]. In particular, the effect sizes for active days per week and sessions per week of −0.33 and 0.27 in the group that saw improvements in perceived stress and the group that did not see an improvement and 0.36 and 0.31 in the group that saw a ≥30% improvement in perceived stress and the group that saw a <30% improvement are notable. These results build on previously reported correlations, suggesting that higher engagement days and a higher number of sessions are associated with a higher percent change in PSS-10 scores and provide further support for the intervention dosage, suggesting multiple bouts of mindfulness each day [22].

Previous research highlights positive mental health outcomes for people who establish consistent health behaviors. In light of higher costs and access to mental health care, these preliminary findings suggest that consistently engaging with Headspace may decrease perceived stress, thus supporting the use of DMH in real-world settings to accessibly improve outcomes.

Strengths, Limitations, and Future Research

This study has several limitations and strengths. As this study was the first real-world study investigating the Headspace app, our data are largely focused on descriptive and group mean differences. While this information is important to establish overall benchmarks for this study, we are limited in the types of questions (eg, prospective) we can ask with this study’s format and recognize that the current statistical approach does not allow for causal inferences. Additionally, these data did not include the demographic information of Headspace members; therefore, we were unable to investigate the impact of demographies on the study outcomes. Future research should examine longitudinal data and changes in perceived stress, while accounting for the demographic characteristics of Headspace members and engagement factors. As a real-world evidence study, we were unable to identify specific programming used by members; we focused only on overall engagement outcomes. Future studies should include more in-depth engagement outcomes to better understand how certain programming is used and impacts perceived stress. Finally, data for this study were collected during the COVID-19 pandemic (March 2020 and January 2023). As noted, stress greatly increased among individuals in the United States during this time, which may have impacted PSS-10 scores (ie, increased scores). However, members were only included if they completed their 2 PSS-10s between 7 and 60 days, and the average number of days between PSS-10 assessments was approximately 1 month, suggesting that if members completed their baseline PSS-10 at the beginning of the pandemic, when stress levels were higher nationwide, they would have completed their follow-up at a similar time period within the pandemic.

A primary strength of the study is the methodology incorporated to understand changes in perceived stress and engagement on the platform. The use of a real-world evidence methodology often helps researchers and clinicians to observe the feasibility and generalizability of interventions in daily functioning. Given the current sample size, which is over 20,000 participants, it is evident that the Headspace platform provides benefits to people who frequently engage with the platform. Furthermore, this study’s large sample size also offers strong evidence for the generalizability of our findings in the real world. The large sample size coupled with the current methodology also allowed us to establish reliable findings for our current platform; therefore, establishing a foundational understanding of how much change is possible on the platform while understanding overall engagement trends.

Conclusions

Our findings suggest that members using Headspace experienced significant reductions in perceived stress in a real-world setting. Furthermore, data suggest that members who engaged with the platform more regularly were more likely to experience improvements in perceived stress. This study is the first to provide real-world evidence of the DMH Headspace platform aimed to reduce participants’ perceived stress. Our results have implications for clinical practice, which include incorporating mediation and mental health psychoeducation as an adjunct to psychotherapy or as a preventive intervention to reduce stress.
Authors' Contributions

CC contributed to the study design, data analysis, manuscript writing, and manuscript review. JK contributed to the study design, manuscript writing, and manuscript review. EH contributed to the study design, data analysis, manuscript writing, and manuscript review. LT contributed to the data analysis and manuscript review. SK contributed to the study design, data analysis, manuscript writing, and manuscript review.

Conflicts of Interest

CC, LT, and SK are currently employed by Headspace. JK and EH were employed by Headspace at the time of data collection.

References


Abbreviations

DMH: digital mental health
PSS-10: Perceived Stress Scale
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

© Christine Callahan, Justin Kimber, Emily Hu, Leah Tanner, Sarah Kunkle. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 4.3.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Weight Loss Using an mHealth App Among Individuals With Obesity in Different Economic Regions of China: Cohort Study

Xinru Huang1,*, MSc; Yefei Shi2,*, PhD; Hongyun Yao1,*, MSc; Mingjie Li1, MSc; Zhijun Lei2, PhD; Jiayun Shi2, MSc; Bo Li2, PhD; Weiwei Zhang1, PhD; Weixia Jian1, MD, PhD

1 these authors contributed equally

Corresponding Author:
Weixia Jian, MD, PhD

Abstract

Background: With the increasing prevalence of obesity, weight loss has become a critical issue in China. Self-managed weight loss through a mobile health (mHealth) app may be a prospective method. However, its practicability in different economic regions of China is unknown.

Objective: This study aims to evaluate the effectiveness of self-managed weight loss through an mHealth app among individuals with obesity in different economic regions of China and to demonstrate the feasibility of online self-management for weight loss.

Methods: A total of 165,635 Chinese adults who signed up for the mHealth app were included to analyze the body composition characteristics of individuals from different economic regions by \( \chi^2 \) analyses. Furthermore, 2 types of participants with obesity using mHealth monitoring, including 74,611 participants with a BMI \( \geq 24.0 \text{ kg/m}^2 \) and 22,903 participants with a normal BMI but an excessive percentage of body fat (PBF), were followed for 6 months to explore the weight loss and fat loss effects in different economic regions of China and to find independent predictors associated with weight loss success by 2-tailed Student \( t \) test and multivariable logistic regression analysis.

Results: There were 32,129 users from low-income regions and 133,506 users from high-income regions. The proportion of users with obesity in low-income regions was higher than in high-income regions, both based on BMI (15,378/32,129, 47.9% vs 59,233/133,506, 44.4%; \( P < .001 \)) and PBF classification (19,146/32,129, 59.6% vs 72,033/133,506, 54%; \( P < .001 \)). Follow-up analyses showed that the weight loss effect among participants with overweight or obesity in low-income regions was greater than in high-income regions (mean –4.93, SD 6.41 vs mean –4.71, SD 6.14 kg; \( P < .001 \)), while there was no significant difference in fat loss (mean –2.06%, SD 3.14% vs mean –2.04%, SD 3.19%; \( P = .54 \)). In the population with normal-weight obesity, the weight loss (mean –2.42, SD 4.07 vs mean –2.32, SD 4.21 kg; \( P = .004 \)) and fat loss effects (mean –1.43%, SD 2.73% vs mean –1.27%, SD 2.63%; \( P < .001 \)) were stronger in high-income regions than in low-income regions. In addition, multivariable logistic regression analyses showed that age, baseline PBF, skeletal muscle rate, and measurement frequency were related to weight loss, whereas gender and baseline body metabolic rate only showed a correlation with weight loss in the population in high-income regions.

Conclusions: This study found a high proportion of mHealth app users with obesity in low-income regions. Individuals with overweight and obesity in different economic regions of China experienced significant weight loss and fat loss using an mHealth app. Moreover, individuals in high-income regions paid more attention to body fat and had better fat reduction effects. Therefore, promoting self-monitoring of weight and PBF through an mHealth app could be an important intervention that could be implemented across all regions of China.

(JMIR Mhealth Uhealth 2024;12:e48675) doi:10.2196/48675

KEYWORDS

weight loss; obesity; normal-weight obesity; economic regions; mHealth app; mobile health; China; mHealth

Introduction

Obesity is a global health crisis that has reached pandemic proportions in many countries, including China [1,2]. The World Health Organization estimated that over 2.1 billion people worldwide are overweight or obese, and this number is expected to continue rising in the coming years [3,4]. Obesity is a significant risk factor for several chronic health conditions, including heart disease, stroke, diabetes, and certain types of cancer, and has been linked to decreased quality of life and premature death [5-7]. In China, the prevalence of overweight and obesity has substantially increased in recent years. This has
been attributed to several factors, including increased urbanization, changes in dietary habits, and decreased physical activity levels [8,9]. The rising burden of obesity in China has led to growing concerns about the health and economic consequences of this trend and has prompted the need for effective weight loss.

Standard behavioral treatment for obesity included dietary and physical activity counseling and self-monitoring of body weight, activity, and diet [10,11]. Consistent self-weighing over time promoted the awareness of behaviors, environments, or situations that might lead to desired or undesired changes in weight. Researchers have established a correlation between self-weighing and successful weight loss, with studies showing that self-weighing significantly improved weight loss outcomes during the first 6 months of a weight loss intervention [12,13].

A new online weight management system, the smart body fat scale, calculates the percentage of body fat (PBF), records and synchronizes the data to mobile health (mHealth) apps, and offers more advantages than a traditional scale [14,15]. In addition to measuring PBF, smart body fat scales usually offer additional features such as the ability to track weight and body fat over time and the ability to measure other health-related metrics such as muscle mass [16]. The use of mHealth with smart body fat scales has become increasingly popular in recent years, as individuals seek to monitor their health and fitness more closely and make more informed decisions about their diets and physical activity patterns.

Currently, the increase in obesity rates is decelerating in high-income areas of China. In contrast, obesity is showing a significant increase in low-income areas, which indicates a requirement for targeted health policies to prevent a further increase in obesity among the general population [17]. Despite the growing popularity of smart body fat scales and the cost-effective potential of digital platforms for reaching a large number of individuals, there was limited research on their impact on weight loss across different economic regions. Here, we conducted a cohort study that analyzed the data of obesity-related anthropometric indices recorded through an mHealth app, which connected to the body fat scale, in individuals with overweight and obesity from low- and high-income regions of China. Our study aimed to investigate the weight loss effects of mHealth connected to smart body fat scales in different economic regions and to test our hypothesis that using mHealth in different economic regions could achieve significant weight loss and fat loss. This would also provide important insights into the potential for self-managed mHealth methods to promote healthy weight loss and improve health outcomes in this population.

Methods

Participants

This study analyzed the data of 165,635 adults aged 18 to 79 years who signed up for the Qingniu Health app between January 2020 and July 2022 and lived in different economic regions of China. Users with a baseline BMI outside the 95% range were excluded. The baseline data were used to determine the general and body composition characteristics of mHealth users from different economic regions. Furthermore, we followed 74,611 participants with a BMI $\geq 24.0$ kg/m$^2$ and 22,903 participants with a normal BMI but an excessive PBF. All participants were followed for 6 months to assess weight and fat loss (Figure 1).
Ethical Considerations
This study was approved by the Ethics Committee of Xinhua Hospitals (approval XHEC-D-2022-195). Electronic consent was provided by all participants. The data were deidentified, and our study did not involve compensation.

Data Collection
The Qingniu Health app was used for the data collection. Participants were required to provide their baseline information, including gender, age, height, and city of residence, at the time of registering in the app. Body composition data, including weight, PBF, skeletal muscle rate (SMR), and basal metabolic rate (BMR), were measured using the same body fat scales (CS10C; Yolanda Technology Co., Ltd). These scales used multifrequency bioelectrical impedance analysis, as previously noted in another study [14].

Subgroups
For the analyses, participants were divided into groups based on gender, age, BMI, PBF, measurement frequency, and economic class of the residential city. Participants younger than 40 years, 40-59 years of age, and 60 years and older were categorized as adults, middle-aged adults, and older adults, respectively. The BMI classification followed standards set by the Ministry of Health’s Disease Control Department for Chinese People. Those with a BMI less than 18.5 kg/m² were considered underweight, those with a BMI between 18.5 and 23.9 kg/m² were considered normal, those with a BMI between

Figure 1. Study flowchart. PBF: percentage of body fat.
24.0 and 27.9 kg/m$^2$ were considered overweight, and those with a BMI of 28.0 kg/m$^2$ or more were considered obese [9]. Male adults with a PBF $\geq$25.0% and female adults with a PBF $\geq$30.0% were considered obese [18], whereas the others were considered nonobese. Individuals with normal weight but had a high PBF were considered to have “normal-weight obesity.” Participants were also evenly divided into groups based on their frequency of measurements: low, medium, and high. First-tier, new first-tier, and second-tier cities were classified as high-income regions, whereas third-tier, fourth-tier, and fifth-tier cities were classified as low-income regions.

Follow-Up Outcomes
The results of the follow-up were measured for 6 months (180±30 d) after the initial registration. A previous study showed that a 5% weight loss reduced the incidence of obesity-related diseases [19]. Therefore, participants who lost more than 5% of their initial body weight were considered to have achieved effective weight loss.

Statistical Analyses
The continuous variables were reported as means and SDs, and the categorical variables were reported as counts and percentages. The frequency of measurements was characterized by medians and IQRs. Differences among groups were analyzed using the unpaired, 2-tailed Student $t$ test, and the differences in the constituent ratios were evaluated with the $\chi^2$ test. Multivariable logistic regression analysis was used to identify the independent factors that influenced the results. The statistical analyses were performed using SPSS 26.0 (IBM Corp). A $P$ value $<$0.05 was considered statistically significant.

Results
General Characteristics of Participants
A total of 165,635 Chinese adults registered in the Qingniu Health app were analyzed, consisting of 32,129 (19.4%) citizens from low-income regions and 133,506 (80.6%) citizens from high-income regions (Table 1). Users were mainly distributed in eastern coastal cities and inland provincial capitals (Figure 2). The majority of users were women (n=133,377, 80.5%) and those aged 18 to 40 years (n=126,796, 76.6%), and the proportion of women in low-income regions was greater than that in high-income regions (27,119/32,129, 84.4% vs 106,258/133,506, 79.6%; $P$<.001). Body composition analyses revealed that the proportion of individuals with overweight and obesity, as classified by BMI, and the proportion of individuals with obesity with a PBF above the upper limit were higher in low-income regions than in high-income regions ($P$<.001). Based on the frequency of measurements during the 6 months, all users were equally divided into 3 groups: low (median 2.29, IQR 1.76 times), medium (median 7.26, IQR 3.95 times), and high (median 22.88, IQR 14.70 times).
<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Low-income regions (n=32,129)</th>
<th>High-income regions (n=133,506)</th>
<th>P value&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total participants (n=165,635), n (%)</td>
<td>32,129 (19.4)</td>
<td>133,506 (80.6)</td>
<td>N/A&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Women, n (%)</td>
<td>27,119 (84.4)</td>
<td>106,258 (79.6)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Age group (years), n (%)</td>
<td></td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>18-40</td>
<td>24,382 (75.9)</td>
<td>102,414 (76.7)</td>
<td></td>
</tr>
<tr>
<td>40-60</td>
<td>7419 (23.1)</td>
<td>29,438 (22)</td>
<td></td>
</tr>
<tr>
<td>60-80</td>
<td>328 (1)</td>
<td>1654 (1.2)</td>
<td></td>
</tr>
<tr>
<td>BMI (kg/m&lt;sup&gt;2&lt;/sup&gt;), n (%)</td>
<td></td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Underweight</td>
<td>696 (2.2)</td>
<td>3662 (2.7)</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>16,055 (50)</td>
<td>70,611 (52.9)</td>
<td></td>
</tr>
<tr>
<td>Overweight</td>
<td>10,625 (33.1)</td>
<td>41,389 (31)</td>
<td></td>
</tr>
<tr>
<td>Obesity</td>
<td>4753 (14.8)</td>
<td>17,844 (13.4)</td>
<td></td>
</tr>
<tr>
<td>Percentage of body fat, n (%)</td>
<td></td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Normal</td>
<td>12,983 (40.4)</td>
<td>61,473 (46)</td>
<td></td>
</tr>
<tr>
<td>Obesity</td>
<td>19,146 (59.6)</td>
<td>72,033 (54)</td>
<td></td>
</tr>
<tr>
<td>Baseline skeletal muscle rate (%), mean (SD)</td>
<td>41.33 (3.93)</td>
<td>42.06 (4.22)</td>
<td>.002</td>
</tr>
<tr>
<td>Baseline basal metabolic rate (kcal), mean (SD)</td>
<td>1339.99 (160.81)</td>
<td>1350.04 (174.93)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Frequency of measurements, n (%)</td>
<td></td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Low</td>
<td>10,638 (33.1)</td>
<td>44,608 (33.4)</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>10,488 (32.6)</td>
<td>44,768 (33.5)</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>11,003 (34.2)</td>
<td>44,130 (33.1)</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> All P values were compared between different economic regions.

<sup>b</sup> N/A: not applicable.
Figure 2. The number of users in mainland China. The pink areas have more than 2000 registrants, the purple areas have 1000-2000 registrants, and the blue areas have fewer than 1000 registrants.

Baseline Characteristics of the Participants With Obesity

In our study, 2 types of obesity were defined based on BMI and PBF: overweight and obese with a BMI $\geq 24.0$ kg/m$^2$ and normal-weight obese with a normal BMI but an excessive PBF. A total of 74,611 participants with overweight and obesity and 22,903 participants with normal-weight obesity were included in the analysis of baseline characteristics. The proportion of users with overweight and obesity in low-income regions was close to half (15,378/32,129, 47.9%), rising to 62.3% (20,021/32,129) when users with normal-weight obesity were included. High-income regions (77,493/133,506, 58%) had a slightly lower proportion compared to low-income regions. There was no significant difference in the distribution of measurement frequency among individuals with overweight and obesity between different economic areas ($P=.08$). The percentage of middle-aged adults was higher (3958/15,378, 25.7% vs 14,748/59,233, 24.9%; $P=.002$) and the mean baseline BMR was lower in low-income areas compared to high-income areas (Table 2).
Baseline characteristics of participants with obesity grouped by regional economic classification.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Overweight and obesity</th>
<th>Normal-weight obesity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low-income regions</td>
<td>High-income regions</td>
</tr>
<tr>
<td></td>
<td>(n=15,378)</td>
<td>(n=59,233)</td>
</tr>
<tr>
<td>Total participants (low-income</td>
<td>15,378 (47.9)</td>
<td>59,233 (44.4)</td>
</tr>
<tr>
<td>regions: n=32,129; high-income</td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>regions: n=133,506), n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline basal metabolic rate</td>
<td>1420.76 (165.97)</td>
<td>1446.57 (181.40)</td>
</tr>
<tr>
<td>(kcal), mean (SD)</td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Age group (years), n (%)</td>
<td>.002</td>
<td>.001</td>
</tr>
<tr>
<td>18-40</td>
<td>11,226 (73)</td>
<td>43,543 (73.5)</td>
</tr>
<tr>
<td>40-60</td>
<td>3958 (25.7)</td>
<td>14,748 (24.9)</td>
</tr>
<tr>
<td>60-80</td>
<td>194 (1.3)</td>
<td>942 (1.6)</td>
</tr>
<tr>
<td>BMI (kg/m²), n (%)</td>
<td>.06</td>
<td>N/A</td>
</tr>
<tr>
<td>Overweight</td>
<td>10,625 (69.1)</td>
<td>41,389 (69.9)</td>
</tr>
<tr>
<td>Obesity</td>
<td>4753 (30.9)</td>
<td>17,844 (30.1)</td>
</tr>
<tr>
<td>Percentage of body fat, n (%)</td>
<td>&lt;.001</td>
<td>N/A</td>
</tr>
<tr>
<td>Normal</td>
<td>875 (5.7)</td>
<td>5462 (9.2)</td>
</tr>
<tr>
<td>Obesity</td>
<td>14,503 (94.3)</td>
<td>53,771 (90.8)</td>
</tr>
<tr>
<td>Frequency of measurements, n (%)</td>
<td>.08</td>
<td>.06</td>
</tr>
<tr>
<td>Low</td>
<td>4308 (28)</td>
<td>16,779 (28.3)</td>
</tr>
<tr>
<td>Medium</td>
<td>4850 (31.5)</td>
<td>19,073 (32.2)</td>
</tr>
<tr>
<td>High</td>
<td>6220 (40.4)</td>
<td>23,381 (39.5)</td>
</tr>
</tbody>
</table>

All P values were compared between different economic regions.

Moreover, the proportion of individuals with normal-weight obesity was also slightly higher in low-income regions than in high-income regions (4643/32,129, 14.5% vs 18,260/133,506, 13.7%; P<.001), and there was no significant difference in the mean baseline BMR of individuals with normal-weight obesity (mean 1234.80, SD 77.04 vs mean 1236.48, SD 77.93 kcal; P=.19; Table 2). Compared with the percentage of middle-aged adults with overweight and obesity (defined by BMI: 18,748/74,611, 25.1%), the percentage of middle-aged adults with normal-weight obesity was higher (9754/22,903, 42.6%).

Weight Loss and Fat Loss of the Participants With Obesity

We followed the aforementioned individuals with obesity for 6 months and found that participants with overweight and obesity who used the self-monitoring mHealth app experienced significant weight loss and fat loss in both regions (low-income regions: mean –4.93, SD 6.41 kg; high-income regions: mean –4.71, SD 6.14 kg; P<.001). Men with overweight and obesity in low-income regions lost weight markedly more than those in high-income regions (P=.01; Table 3). When grouped by age, baseline BMI, and measurement frequency, weight loss was greater among users in low-income regions than in high-income regions. Additionally, individuals with normal-weight obesity showed less weight loss than individuals with overweight and obesity. In contrast, in the population with normal-weight obesity, individuals lost more weight in high-income regions than those in low-income regions.
### Table 4. Weight loss grouped by regional economic classification.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Weight loss in the overweight and obesity group (kg), mean (SD)</th>
<th>Weight loss in the normal-weight obesity group (kg), mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low-income regions</td>
<td>High-income regions</td>
</tr>
<tr>
<td>All participants</td>
<td>–4.93 (6.41)</td>
<td>–4.71 (6.14)</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>–4.40 (6.81)</td>
<td>–4.08 (6.51)</td>
</tr>
<tr>
<td>Women</td>
<td>–5.08 (6.27)</td>
<td>–4.99 (5.94)</td>
</tr>
<tr>
<td><strong>Age group (years)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-40</td>
<td>–5.08 (6.68)</td>
<td>–4.82 (6.37)</td>
</tr>
<tr>
<td>40-60</td>
<td>–4.56 (5.59)</td>
<td>–4.39 (5.39)</td>
</tr>
<tr>
<td>60-80</td>
<td>–3.68 (4.86)</td>
<td>–4.33 (5.75)</td>
</tr>
<tr>
<td><strong>BMI (kg/m^2)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overweight</td>
<td>–4.07 (5.48)</td>
<td>–3.92 (5.16)</td>
</tr>
<tr>
<td>Obesity</td>
<td>–6.85 (7.77)</td>
<td>–6.53 (7.64)</td>
</tr>
<tr>
<td><strong>Percentage of body fat (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obesity</td>
<td>–5.08 (6.40)</td>
<td>–4.94 (6.19)</td>
</tr>
<tr>
<td><strong>Frequency of measurements</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>–3.24 (7.44)</td>
<td>–2.92 (6.74)</td>
</tr>
<tr>
<td>Medium</td>
<td>–3.96 (5.72)</td>
<td>–3.84 (5.53)</td>
</tr>
<tr>
<td>High</td>
<td>–6.86 (5.58)</td>
<td>–6.70 (5.57)</td>
</tr>
</tbody>
</table>

^a All P values were compared between different economic regions.

^b N/A: not applicable.

Considering that the goal of participants with normal-weight obesity was to reduce body fat, we further explored the differences in the efficiency of fat loss in different economic regions. We found that the fat loss effect on the population with normal-weight obesity was greater in high-income regions. However, grouped by gender, age, and measurement frequency, we conducted unpaired 2-tailed Student t tests for each group and found that there were no differences in the fat loss effect on people with overweight and obesity in different economic regions (Table 4).
Table. Fat loss of participants with obesity grouped by regional economic classification.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Fat loss in the overweight and obesity group (PBF(^a), %), mean (SD)</th>
<th>Fat loss in the normal-weight obesity group (PBF, %), mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low-income regions</td>
<td>High-income regions</td>
</tr>
<tr>
<td>All participants</td>
<td>(-2.06 (3.14))</td>
<td>(-2.04 (3.19))</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>(-2.00 (3.57))</td>
<td>(-2.03 (3.76))</td>
</tr>
<tr>
<td>Women</td>
<td>(-2.07 (3.01))</td>
<td>(-2.04 (2.91))</td>
</tr>
<tr>
<td><strong>Age group (years)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-40</td>
<td>(-2.13 (3.26))</td>
<td>(-2.11 (3.30))</td>
</tr>
<tr>
<td>40-60</td>
<td>(-1.87 (2.82))</td>
<td>(-1.85 (2.86))</td>
</tr>
<tr>
<td>60-80</td>
<td>(-1.31 (2.12))</td>
<td>(-1.70 (2.83))</td>
</tr>
<tr>
<td><strong>BMI (kg/m(^2))</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overweight</td>
<td>(-1.85 (2.92))</td>
<td>(-1.82 (2.95))</td>
</tr>
<tr>
<td>Obesity</td>
<td>(-2.52 (3.55))</td>
<td>(-2.55 (3.65))</td>
</tr>
<tr>
<td><strong>PBF (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obesity</td>
<td>(-2.13 (3.09))</td>
<td>(-2.15 (3.13))</td>
</tr>
<tr>
<td><strong>Frequency of measurements</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>(-1.42 (3.80))</td>
<td>(-1.33 (3.60))</td>
</tr>
<tr>
<td>Medium</td>
<td>(-1.64 (2.92))</td>
<td>(-1.68 (2.94))</td>
</tr>
<tr>
<td>High</td>
<td>(-2.81 (2.61))</td>
<td>(-2.84 (2.90))</td>
</tr>
</tbody>
</table>

\(^a\)PBF: percentage of body fat.
\(^b\)All \(P\) values were compared between different economic regions.
\(^c\)N/A: not applicable.

**Independent Factors Linked With Successful Weight Loss in Different Economic Regions**

We conducted multivariable logistic regression analyses to examine the relationship between various baseline factors and successful weight loss in individuals with overweight and obesity. In the population from low-income regions, younger participants with higher baseline PBF, baseline SMR, and measurement frequency were more likely to succeed in weight loss (Table 5). In addition, women in high-income regions were more likely to achieve successful weight loss. The results indicated that the frequency of measurement was the most critical independent factor in both low-income (odds ratio 5.036, 95% CI 4.618-5.491; \(P<.001\)) and high-income (odds ratio 5.271, 95% CI 5.042-5.511; \(P<.001\)) regions, particularly a high measurement frequency. Similarly, we further estimated the dependent factors associated with successful fat loss in the populations with overweight and obesity. In different economic regions, younger men with a higher baseline PBF and measurement frequency and lower BMR were more likely to lose body fat (Table 5).
Discussion

Principal Findings

After 6 months of follow-up, the results showed that significant weight and fat loss were found in participants with overweight and obesity using the mHealth app in different economic regions. Furthermore, individuals in low-income regions lost more weight than individuals in high-income regions in the population with overweight and obesity, and there was no difference between individuals in fat loss. Providing further health education and online weight loss monitoring in low-income regions was beneficial to the population. Interestingly, in the population with normal-weight obesity, individuals in high-income regions lost more fat than individuals in low-income regions.

Comparison With Prior Work

Obesity has become a major public health challenge worldwide, with strong links to metabolic disorders such as cardiovascular diseases [5] and diabetes [20]. With the transformation of China’s economic and social structure, dietary patterns and nutritional status have undergone significant changes, and the problem of obesity has become increasingly serious [21].

Addressing obesity can help reduce the incidence of chronic diseases [22], and the economic benefits of long-term nonsurgical weight loss in individuals with obesity have been well established [23]. In the previous study, we found that a large number of Chinese individuals with overweight and obesity were able to achieve weight loss goals through an mHealth app during a long-term follow-up, leading us to conclude that mHealth with body fat scales might be a promising method for weight loss and fitness [24]. However, given the vast territory and large population of China (Figure 2), we aimed to investigate whether self-management through mHealth could achieve consistent results across different economic regions, further demonstrating the feasibility of online self-management for weight loss.

The mHealth app with the body fat scale is a widely used self-weighing tool due to its affordability and convenience, different from the traditional paper record [25]. We analyzed the user data of the Qingniu Health app linked to the body fat scale. We divided all the included users into 2 groups: high-income regions and low-income regions according to the economic classification of their living regions. Results showed that the number of users in high-income regions was more than quadruple that in low-income regions (n=133,506 vs n=32,129).

Table. Factors linked with successful weight loss and fat loss in overweight and obese group.

| Characteristics | Weight loss | | | Fat loss | | |
|----------------|------------|----------------|----------------|----------------|----------------|
|                | Low-income regions | OR (95% CI) | P value | High-income regions | OR (95% CI) | P value |
|                | Low-income regions | OR (95% CI) | P value | High-income regions | OR (95% CI) | P value |
|                | OR (95% CI) | P value | OR (95% CI) | P value | OR (95% CI) | P value |
| Gender         | N/A \(^b\) | N/A | 0.748 (0.680-0.823) | <.001 | 2.568 (2.139-3.081) | <.001 |
|                | OR (95% CI) | P value | OR (95% CI) | P value | OR (95% CI) | P value |
| Age            | 0.981 (0.977-0.985) | <.001 | 0.983 (0.981-0.985) | <.001 | 0.971 (0.967-0.975) | <.001 |
|                | OR (95% CI) | P value | OR (95% CI) | P value | OR (95% CI) | P value |
| Baseline PBF\(^c\) | 1.109 (1.087-1.132) | <.001 | 1.079 (1.071-1.088) | <.001 | 1.047 (1.034-1.060) | <.001 |
| Baseline SMR\(^d\) | 1.052 (1.029-1.075) | <.001 | 1.036 (1.025-1.048) | <.001 | N/A | N/A |
| Baseline BMR\(^e\) | N/A | N/A | 1.001 (1.000-1.001) | <.001 | 0.999 (0.999-1.000) | <.001 |
| Frequency of measurements classification | | | | | | |
| Medium-low     | 1.611 (1.479-1.755) | <.001 | 1.771 (1.695-1.851) | <.001 | 1.557 (1.427-1.700) | <.001 |
|                | OR (95% CI) | P value | OR (95% CI) | P value | OR (95% CI) | P value |
| High-low       | 5.036 (4.618-5.491) | <.001 | 5.271 (5.042-5.511) | <.001 | 4.606 (4.224-5.023) | <.001 |
|                | OR (95% CI) | P value | OR (95% CI) | P value | OR (95% CI) | P value |

\(^a\)OR: odds ratio.
\(^b\)N/A: not applicable.
\(^c\)PBF: percentage of body fat.
\(^d\)SMR: skeletal muscle rate.
\(^e\)BMR: basal metabolic rate.
and the vast majority of users were women in both groups, which was similar to other research [26]. This revealed that people in high-income regions paid more attention to obesity prevention and monitoring, particularly women. Furthermore, the age distribution of users showed a majority of young and middle-aged people. When grouped by baseline BMI and PBF, both the proportion of people with an excessive BMI and the proportion of people with an excessive PBF in low-income regions were significantly higher than those in high-income regions, as noted in a comment published in Nature [27]. In all high-income countries, overweight and obesity levels were already higher in rural areas than in urban areas, and the same phenomenon might be occurring in China [27,28]. These results suggest that self-management and lifestyle interventions are required to prevent further development of metabolic-related diseases in people with excess body weight and body fat in low-income areas.

Previous studies have established a close association between socioeconomic status and the risk of obesity [29-31]. Despite this, the relationship between weight loss and economic regions among Chinese adults remains largely unknown. Our study of the user data of the Qingniu Health app linked to the body fat scale found that nearly half of the users met the criteria for a diagnosis of obesity based on their BMI. The average baseline BMR for users in low-income regions was lower, and the individual BMR was correlated with multiple factors such as age, gender, body composition, and BMI [32]. There were several users with overweight and obesity with a normal baseline PBF and some people with a normal BMI but excessive PBF, which we called the population with normal-weight obesity. Normal-weight obesity was closely associated with metabolic and cardiovascular diseases [33]. Our findings showed that the proportion of the population with normal-weight obesity in low-income regions was higher than in high-income regions. Moreover, the increase in the proportion of middle-aged adults with normal-weight obesity was associated with an increase in biological age, a lack of physical activity, and other factors. This phenomenon highlighted the need for increased focus on body fat control in this age group [34]. As expected, we found no statistical difference in the composition of measurement frequency between participants in different economic regions, showing that users were able to maintain a certain frequency of use with the app, which suggested that promoting the use of mHealth in our country was a feasible and effective approach to monitoring obesity.

In our previous study [24], we found that a younger age was the most important contributing factor to fat loss success. In this study, the age composition of participants with normal-weight obesity in high-income regions was not different from that in low-income regions but showed a more obvious trend of weight loss and fat loss, reflecting that participants from high-income regions had focused on body fat in addition to body weight. To better understand the factors affecting successful weight loss in different economic regions, we found that age, baseline PBF, SMR, and measurement frequency were significant predictors for weight loss success, whereas gender and baseline BMR only showed a correlation with weight loss in the population in high-income regions, which suggests women in high-income regions paid more attention to self-management than those in low-income regions.

**Limitations**

Our study had a large sample size and a prolonged follow-up period, which accurately reflected the real-world situation of online self-management in China. Nevertheless, there were certain limitations in our study. We must acknowledge that our study was a nonrandomized cohort study without controls. Nevertheless, we observed a significant weight loss outcome in individuals with overweight and obesity from the 2 different economic regions in this large-scale follow-up study. Furthermore, individuals who used the smart body fat scale with a higher frequency exhibited better weight loss results compared to those with a lower use frequency. This suggests that the use of an mHealth app connected to a smart body fat scale has a certain impact on effective weight reduction, regardless of economic region. In the future, we can consider adding control groups using alternative weight loss methods to further clarify the role of using an mHealth app connected to a smart body fat scale in self-initiated weight and fat loss. Additionally, we did not consider the 2 important influencing factors of participants' diet and exercise volume. We also did not consider the overall health status of the participants; factors such as hydration level might impact the accuracy of body fat measurement using bioelectrical impedance analysis.

**Conclusions**

A high proportion of individuals with obesity from low-income regions was found in our study, and individuals with overweight and obesity who used body fat scales in different economic regions of China experienced significant weight loss and fat loss. Individuals from high-income regions paid more attention to body fat and had better fat loss than those from low-income regions, and in the middle-aged population, the issue of normal-weight obesity required more attention. Therefore, promoting self-monitoring of weight and fat through the use of body fat scales connected to an mHealth app could be an important intervention measure for the population with overweight and obesity across all regions of China.

**Acknowledgments**

This study was supported by the Chinese National Natural Science Foundation (grant 82270867) and the Natural Science Foundation of Shanghai (grant 20ZR1435300). We also acknowledge all the users involved in the study and the staff of the Qingniu Health app, Yolanda Technology Co., Ltd.
Data Availability
Data are available from the corresponding author with the permission of the Qingniu Health app, Yolanda Technology Co., Ltd.

Authors' Contributions
XH, YS, and HY contributed equally. XH, YS, HY, and WJ designed and wrote the manuscript. ML and ZL analyzed the data. JS, BL, and WZ reviewed the analyses and manuscript. All authors approved the final version.

Conflicts of Interest
None declared.

References


Abbreviations

| BMR: | basal metabolic rate |
| mHealth: | mobile health |
| PBF: | percentage of body fat |
| SMR: | skeletal muscle rate |

© Xinru Huang, Yefei Shi, Hongyun Yao, Mingjie Li, Zhijun Lei, Jiayun Shi, Bo Li, Weiwei Zhang, Weixia Jian. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 16.1.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Original Paper

Use and Engagement With Low-Intensity Cognitive Behavioral Therapy Techniques Used Within an App to Support Worry Management: Content Analysis of Log Data

Paul Farrand¹²*, PhD; Patrick J Raue³*, PhD; Earlise Ward⁴*, PhD; Dean Repper⁵*, MSc; Patricia Areán³*, PhD

¹Clinical Education, Development and Research, Faculty of Health and Life Sciences, University of Exeter, Exeter, United Kingdom
²Department of Psychology, Faculty of Health and Life Sciences, University of Exeter, Exeter, United Kingdom
³AIMS CENTER, Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, United States
⁴School of Medicine and Public Health, Carbone Comprehensive Cancer Center, University of Wisconsin-Madison, Madison, WI, United States
⁵Trent PTS, Improving Access to Psychological Therapies, Derby, United Kingdom

*all authors contributed equally

Corresponding Author:
Paul Farrand, PhD
Clinical Education, Development and Research
Faculty of Health and Life Sciences
University of Exeter
Sir Henry Wellcome Building for Mood Disorders Reserarch
Perry Road
Exeter, EX4 4QG
United Kingdom
Phone: 44 01392725793
Email: p.a.farrand@exeter.ac.uk

Abstract

Background: Low-intensity cognitive behavioral therapy (LICBT) has been implemented by the Improving Access to Psychological Therapies services across England to manage excessive worry associated with generalized anxiety disorder and support emotional well-being. However, barriers to access limit scalability. A solution has been to incorporate LICBT techniques derived from an evidence-based protocol within the Iona Mind Well-being app for Worry management (IMWW) with support provided through an algorithmically driven conversational agent.

Objective: This study aims to examine engagement with a mobile phone app to support worry management with specific attention directed toward interaction with specific LICBT techniques and examine the potential to reduce symptoms of anxiety.

Methods: Log data were examined with respect to a sample of “engaged” users who had completed at least 1 lesson related to the Worry Time and Problem Solving in-app modules that represented the “minimum dose.” Paired sample 2-tailed t tests were undertaken to examine the potential for IMWW to reduce worry and anxiety, with multivariate linear regressions examining the extent to which completion of each of the techniques led to reductions in worry and anxiety.

Results: There was good engagement with the range of specific LICBT techniques included within IMWW. The vast majority of engaged users were able to interact with the cognitive behavioral therapy model and successfully record types of worry. When working through Problem Solving, the conversational agent was successfully used to support the user with lower levels of engagement. Several users engaged with Worry Time outside of the app. Forgetting to use the app was the most common reason for lack of engagement, with features of the app such as completion of routine outcome measures and weekly reflections having lower levels of engagement. Despite difficulties in the collection of end point data, there was a significant reduction in severity for both anxiety (t₅₃=5.5; P<.001; 95% CI 2.4-5.2) and low mood (t₅₃=2.3; P=.03; 95% CI 0.2-3.3). A statistically significant linear model was also fitted to the Generalized Anxiety Disorder–7 (F₂,5₁=6.73; P<.001), while the model predicting changes in the Patient Health Questionnaire–8 did not reach significance (F₂,5₁=2.33; P=.11). This indicates that the reduction in these measures was affected by in-app engagement with Worry Time and Problem Solving.

Conclusions: Engaged users were able to successfully interact with the LICBT-specific techniques informed by an evidence-based protocol although there were lower completion rates of routine outcome measures and weekly reflections. Successful interaction
with the specific techniques potentially contributes to promising data, indicating that IMWW may be effective in the management of excessive worry. A relationship between dose and improvement justifies the use of log data to inform future developments. However, attention needs to be directed toward enhancing interaction with wider features of the app given that larger improvements were associated with greater engagement.

(JMIR Mhealth Uhealth 2024;12:e47321) doi:10.2196/47321

**KEYWORDS**

cognitive behavioral therapy; low-intensity; mCBT; app; log data; worry management; CBT; management; application; therapy; implementation; treatment; symptoms; anxiety; worry; engagement

**Introduction**

Excessive worry represents a core characteristic associated with generalized anxiety disorder (GAD) [1], characterized as 6 or more months of chronic worry about several different events and little belief worry can be controlled, and is associated with general somatic symptoms of anxiety [2]. It is highly pervasive in high-income countries, with a lifetime prevalence rate of 7.8% in the United States [3] and a median age of onset between 24 and 50 years, and is more common in women and people who are unemployed [4]. Excessive worry is deleterious to the individual, and if it manifests into GAD, it places a significant burden on society and employers with an average of 6.3 days per month of work absenteeism [5]. Furthermore, treatment is associated with increased service use [6], thereby placing a significant burden on primary care settings within both the United Kingdom [7] and United States [8].

Despite excessive worry impacting on the individual quality of life with progression to GAD representing a major public health problem [9], treatment availability remains limited. The treatment gap for GAD has been reported to be as high as 67% in the United Kingdom and 70% in the United States [10]. In an attempt to address the treatment gap [11], innovations in the delivery of evidence-based psychological therapy have been witnessed, for example, by broadening the workforce delivering cognitive behavioral therapy (CBT) for late-life GAD with no differences in effectiveness demonstrated when delivered by bachelor-level lay providers compared with PhD-level experienced therapists [12]. Further attempts to address the treatment gap have been addressed within the Improving Access to Psychological Therapies (IAPT) program, implementing low-intensity cognitive behavioral therapy (LICBT) self-help for the management of mild to moderate common mental health problems alongside therapist-delivered high-intensity CBT for moderate to severe presentations [13].

While improved access has been facilitated through the IAPT program, predictions indicate that access will only be increased to 25% of the community prevalence of depression and anxiety disorders by 2023-2024 [14]. Furthermore, between 2021 and 2022, only 37% of patients completed a course of therapy [15]. Difficulties in reducing the treatment gap are further dependent on a number of individual-level barriers such as stigma, desire to handle the problem independently, and limited willingness to disclose problems [11].

Greater implementation of digital health technologies such as smartphone apps [16] has potential to address barriers at the level of the individual. Furthermore, with high use of smartphones, for example, 81% of adults within the United States [17], apps offer the promise of delivering mental well-being interventions at scale and derive cost savings with respect to delivery and increased productivity within the workplace [18]. The expansion of apps to manage challenges with well-being has predominantly been based on CBT [19]. CBT is particularly well suited to inform mental well-being apps with emphasis placed on active engagement with specific techniques rather than exclusive reliance on a supportive relationship with a behavioral health coach or therapist [13]. This is especially salient with regard to an approach based on “collaborative empiricism,” whereby after engagement with specific CBT techniques, people are encouraged to explore outcomes for themselves [20]. In doing so, a better understanding of the way their mental health difficulty is affecting them can be derived through an appreciation of the cognitive behavioral model.

Despite CBT providing a compatible evidence-based approach for mental well-being apps with the potential to deliver at scale to close the treatment gap, implementation and uptake remain limited [21]. This is partly dependent on poor engagement with apps relying on factors such as poor usability, techniques inconsistent with user expectations, and poor health literacy [22]. Exploring ways to enhance engagement is of significance given that greater engagement has been reported to predict improvements in mental well-being [23]. In particular, focusing on obstacles and difficulties encountered in completing LICBT techniques is of significance given their effectiveness [24], while not dismissing common factors to establish a good “therapeutic relationship” generating a sense of genuineness, warmth, and collaborative working [25]. Focusing on both common and specific LICBT techniques used within the Iona Mind Well-being app for Worry management (IMWW) is therefore of importance given that the combination of both is crucial for bringing about therapeutic change [26].

This paper reports log data [27] to examine engagement with IMWW based on LICBT to help in the management of excessive worry. It has been proposed that rather than looking at overall engagement regarding areas such as number of sessions engaged with or session duration, it is better to focus attention on specific user interactions [28]. The focus of this paper is therefore directed toward appreciating engagement and interaction with specific LICBT techniques and wider features of IMWW to examine potential enhancements alongside wider usability. Furthermore, the relationship between engagement and outcomes...
will be explored to inform ongoing development to maximize effectiveness.

**Methods**

**Design**

Areas representing the focus of log data analysis have been informed by recommendations identified as useful when seeking to address the effectiveness of eHealth technology alongside behavioral and theoretical models [27]. Analysis was only undertaken on data collected regarding “engaged users” and their log data associated with engagement with the LICBT techniques. To be considered an engaged user, the user was required to have completed at least 1 lesson related to the Worry Time and Problem Solving in-app modules within any time period. These criteria represented the “minimum dose” [29] necessary for the user to be considered to have engaged enough to be able to understand the in-app CBT techniques and apply them outside of the app to manage excessive worry. This approach to represent “engaged users” has previously been adopted. For example, with respect to a feasibility trial examining internet-administered LICBT for parents of children treated for cancer [30].

A \( \chi^2 \) test of homogeneity was undertaken to compare demographic data provided by the engaged users and those who failed to engage with specific LICBT techniques to receive a minimum dose.

**Sample**

Over 6 months (July 19, 2022, to February 19, 2023), 956 adults 18 years and older with a GAD-7 score of 6 and above downloaded and opened IMWW and completed the sign-up process. Of these, 803 (84%) adults did not engage sufficiently with the app to be considered an “engaged user,” resulting in 153 (16%) users engaging with the app sufficiently to be considered to have received a minimum dose (Table 1).

Approximate data regarding the continent the user was accessing the app from were automatically collected by the app from the time zone set on the user’s phone and therefore collected on all 956 adults. \( \chi^2 \) tests of homogeneity indicate that differences between engaged users and those who had downloaded the app but failed to receive a minimum dose were not significant at the 5% level across any of the demographic variables. In addition to the collection of demographic data, 41 of 153 (27%) engaged users responded to a question regarding receipt of other therapy, of whom 34 (83%) users indicated that they were not.

<table>
<thead>
<tr>
<th>Table 1. Demographic questionnaire responses completed (N=956).</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variable</strong></td>
</tr>
<tr>
<td><strong>Gender, n (%)</strong></td>
</tr>
<tr>
<td>Women</td>
</tr>
<tr>
<td>Men</td>
</tr>
<tr>
<td>Other</td>
</tr>
<tr>
<td><strong>Age range (years; n=53), n (%)</strong></td>
</tr>
<tr>
<td>18-24</td>
</tr>
<tr>
<td>25-34</td>
</tr>
<tr>
<td>35-44</td>
</tr>
<tr>
<td>45-54</td>
</tr>
<tr>
<td>55-64</td>
</tr>
<tr>
<td>≥65</td>
</tr>
<tr>
<td><strong>Continent, n (%)</strong></td>
</tr>
<tr>
<td>Americas</td>
</tr>
<tr>
<td>Europe</td>
</tr>
<tr>
<td>Asia</td>
</tr>
<tr>
<td>Africa</td>
</tr>
<tr>
<td>Australasia</td>
</tr>
<tr>
<td>Unknown</td>
</tr>
</tbody>
</table>

\( ^a \)A total of 60 engaged users and 59 users who did not receive the minimum dose responded.

**Iona Mind Well-Being App for Worry Management**

The IMWW is, in part, informed by the techniques described in the LICBT written self-help intervention for managing excessive worry [31], based on the CBT protocol for the management of GAD [32]. LICBT is recommended for the management of GAD [33] and is one of the most commonly adopted written self-interventions used within the IAPT program [34]. The focus of IMWW is explained during onboarding where the user is required to explicitly acknowledge its purpose as a
well-being tool. Users wishing to continue engagement acknowledge that they understand conditions related to use and consent to have their data processed.

**Collection of Demographic Data**

Demographic data were not used to inform the delivery or functionality of the app. Therefore, a screen requesting demographic data, or a question regarding receipt of other therapy, was only presented once the user had engaged with IMWW on 2 occasions at least 6 hours apart following enablement, and no other higher priority messages were pending. If higher priority messages were pending, the request to provide demographic data was repeatedly postponed to the following day until the user had supplied all data, completed specific questions, or declined the request to open the screen (Table 1). Due to the optionality and logic surrounding whether demographic data were requested from users who downloaded the app, such data were therefore not requested from all users and were not prioritized over other more useful app functionality. Collecting demographic data was not prioritized given that answering such questions on an app can increase the risk of disengagement [35].

**Supporting Interaction**

Interaction is supported by an algorithmically driven chatbot simulating a “conversation” between the conversational agent (CA) and the user. Users interact by entering raw text or selecting a predetermined response. The type of response depends on the type of content being delivered and varies between selecting a button from a list of options or entering free text when a personalized response is requested. User feedback is collected at the level of a user message supplied within the “conversation” through selecting a “thumbs up” or “thumbs down” icon.

Key principles associated with user-centered systems design were adopted to inform the development of IMWW [36]. The app was developed to manage excessive worry and support emotional wellness with an “SOS” button prominently displayed for users finding themselves in significant emotional distress. If selected, signposting information to a comprehensive list of local and international crisis helplines is presented alongside mindfulness practices to assist with mood stabilization. Before engaging, the user is further reminded that IMWW is not intended to deliver treatment but rather is a tool to support well-being and is not designed for anyone who has been diagnosed with a psychiatric disorder. Consequently, it is stressed that the app is not to be used outside of the context of a well-being self-help aid.

**Progress Through IMWW**

Engagement begins with users landing on the Today home screen and progressing through 6 educational modules through which they learn about and interact with the LICBT techniques (Multimedia Appendix 1).

Educational modules are chronologically ordered and unlocked as engagement is initiated. The order in which they are unlocked is dependent on user choice, reflecting whether they wish to initially address practical or hypothetical worries. However, the user is able to move forward and backward between these specific factors to address the different types of worries where preferred. After onboarding, the user is given the opportunity to complete the GAD-7 and Patient Health Questionnaire–8 (PHQ-8) weekly during their weekly review, with scores presented on a progress screen. However, following the completion of these measures during onboarding, subsequent completion is voluntary.

**Home Screen**

Informed by the CBT protocol for the management of GAD, which can also be used in the context of improving emotional well-being by supporting the management of worries [37], the Today (Home) screen supports the user to record their worries. This screen also presents a timer that counts down to the user’s scheduled worry time and offers tips to complete Worry Time (Multimedia Appendix 1). Should the user not have completed the lesson, a placeholder is displayed inviting them to learn more about Worry Time. The screen has been designed to make it as easy as possible for users to record worries, plan Worry Time, and access the CBT content. A navigation bar at the bottom of the screen links to the screens related to the LICBT techniques through which users can engage depending on preference.

**CBT Techniques**

To promote engagement, the IMWW is informed by collaborative empiricism where the user actively engages with techniques associated with CBT [38]. Collaborative empiricism has been identified as core to the therapeutic relationship [39], supporting “learning by doing” fundamental to CBT [40]. Accordingly, LICBT techniques are presented as skills to be mastered through regular practice. The user is initially introduced to the CBT model followed by supporting them to record and categorize their worries. The user is then provided with the choice of Problem Solving or Worry Time to address practical or hypothetical worries, respectively.

**CBT Model**

An interactive CBT Five Areas model (introduced in October 2022) is presented [41], and the user is encouraged to interact and identify a current situation in the “here and now.” In response to this current situation, the user is also encouraged to interact with boxes reflecting “Thoughts” that go through their head, “Behaviors” engaged in, “Physical Feelings,” and to recognize “Emotions” (Multimedia Appendix 1). Additionally, the model helps them appreciate ways in which the specific LICBT techniques presented to address practical and hypothetical worries may be helpful. The model serves as psychoeducation, enabling the user to recognize the interaction between each of these areas, understand the nature of their worry, and appreciate factors that maintain their worry behavior and the impact of physical symptoms associated with anxiety.

**Worry Diary**

The primary function of the Worry Diary is to enable the user to actively add new worries as they arise throughout the day and as a record of worries for subsequent review (Multimedia Appendix 1). Prompts and predefined categories are used to
enable the user to differentiate between practical worries that have a solution and hypothetical worries that do not.

Problem Solving

Where practical worries are logged, the user is guided to list potential solutions, consider strengths and weaknesses for each solution, and select the most appropriate one. A time to try the solution out is then optionally scheduled by the user with a reminder given to complete it at the chosen time. After the chosen time has passed or 30 hours elapsed, on opening the app, the CA will ask the user to review how their solution went. Prompts ask the user if the problem was resolved and if not request further information regarding the challenges encountered. Advice is offered where problems have been encountered, putting the solution into action (eg, to break the problem down or work through and apply another solution).

Worry Time

Unless explicitly overridden by the user, all worries identified as hypothetical are displayed only during Worry Time at a time determined by the user and are blurred out at all other times. Users are reminded that their worry time is starting with a push notification to their phone. If the user opens the app during Worry Time, they are prompted to work through the time they have set aside to worry with the CA. If they choose to do so, the CA will list out all user worries and request that these be worried about for the specified period of time. Subsequently, the CA will review each worry with the user, asking them if it still remains an issue or if it now better represents a practical worry. Worry Time represents a form of cognitive exposure with users exposed to hypothetical worries written down during the day. This is proposed to overcome avoidance behavior and reduce intolerance of uncertainty when it is recognized that there are no solutions to the worry [37].

Maintaining Engagement

A chatbot informed by theoretically driven techniques is adopted to help establish a “therapeutic approach” to maintain and promote user engagement embedded within IMWW [38]. Such techniques help to establish an approach based on collaborative empiricism [39], whereby the user is encouraged to explore outcomes arising from engagement for themselves.

Conversational Agent

Support is omnipresent throughout the engagement and comes in the form of an algorithmically driven chatbot stimulating engagement between IMWW, the CA, and users. This helps them overcome difficulties encountered with the specific LICBT techniques and uses common factors to maintain engagement. Upon recognition that difficulties are experienced with any of the specific techniques, the CA is deployed to enable users to work through the specific techniques. If the user reports difficulties in trying out a solution they have planned, the CA will ask questions to determine the nature of the difficulty encountered and direct them to the appropriate parts of the app. For example, the user would be directed back to Problem Solving should they need to break the problem down, or Worry Diary if the worry appears to be hypothetical rather than practical (Table 1).

On other occasions, the CA provides the user with helpful tips and advice or the opportunity to ask FAQs to navigate difficulties experienced. For example, 2 days after learning about Worry Time, the CA will check back in with the user and ask how the exercise has been going. Depending on user response, advice will be given. For example, if the user forgets to engage with Worry Time, they are reminded to turn on their notifications and set an alarm on their phone to serve as a prompt. Consistent with the delivery of CBT, during engagement with the LICBT techniques, the CA brings the user back to the CBT model to reinforce their understanding of the intervention and maintain motivation for continued engagement.

Common Factors

The CA uses nontherapeutic common factor skills in the form of “therapeutic empathy” to instill a sense of hopefulness and encouragement to maximize engagement with the specific factors linked to symptom reduction [42]. Statements include those demonstrating an empathic stance highlighting a desire to help alongside empathic attunement where statements demonstrate an appreciation of the user’s emotional experience [42]. When recognizing that the user is experiencing difficulties in engaging or is not improving, the CA uses empathy to maintain engagement.

Behavior Change Techniques

Within the module on recording worries, Behavioral Contracting [43] encourages the user to sign an agreement to consistently engage with IMWW throughout the 6-week program with a separate Goal Setting lesson guiding the user to set approach, rather than avoidance, goals [44]. Behavior “Push” notifications serve to prompt or maintain behavior change while engaging with the app. Furthermore, constructs derived from self-determination theory [45] promote autonomy and intrinsic motivation that serve to facilitate collaborative empiricism [37].

Monitoring Progress

Throughout engagement, a progress screen presents a summary of the user’s app use and engagement with in-app lessons, previously entered goals, and scores regarding symptom severity associated with anxiety (GAD-7) [46] and low mood (PHQ-8) [47]; it also presents links to the settings page, which houses operational features such as typing speed (Multimedia Appendix 1). Given a potential association between providing feedback and improved outcomes, all data collected are repeatedly presented to the user throughout engagement [48].

Weekly Reflection

Consistent with face-to-face CBT [39], on a weekly basis, the CA prompts the user to reflect on their engagement with IMWW and the features found most helpful (Table 1). Using reflective learning within the app facilitates learning, with the CA encouraging engagement to promote self-discovery [49]. Furthermore, during the weekly reflection, the CA requests information on the LICBT techniques engaged with and highlights those found most helpful. In the event an identified technique was not engaged with, the CA also requested information as to the main reasons from a range of options provided.
Data Collection and Analysis

Log Data
Consistent with the aims of the study, analysis was undertaken on log data collected from engaged users to reflect their engagement with the specific LICBT techniques. Log data were collected by IMWW automatically logging the actions of each engaged user and requests to complete surveys throughout the use of the app. From these data, summary statistics for use in this paper were extracted. Progress of engaged users through IMWW was monitored and informed by data regarding the number of sessions completed, session duration, weekly reflections, and completion of LICBT techniques alongside summary statistics recorded. Engagement with IMWW was explored with respect to the number of users who reached the CBT model and interacted with it, text entered into each area, completion of the LICBT techniques, and general input and behavior during the engagement. Specific worry management techniques were examined with respect to the number of worries entered and the proportion classified as practical or hypothetical problems. With respect to Problem Solving, data analysis included the number of users who completed the lesson, the number of times the in-app tool was used to solve a practical problem, the number of practical problems entered, the number of users prompted to follow up on their problem-solving with the CA, and the number who engaged with it. Furthermore, analysis was undertaken on the number of engaged users who sought to manage hypothetical worries by learning about Worry Time, set a time for Worry Time, and started an in-app session alongside the number of hypothetical worries entered being recorded.

User responses from the Weekly Reflection conversation within IMWW were also collected and analyzed to gauge general engagement with the specific techniques. As a proxy for behavior change approaches adopted to maintain engagement within the app, the number of users who were delivered at least 1 push notification and the number of those who interacted were also examined. For engaged users completing more than 1 GAD-7 or PHQ-8 at assessment, the log of assessments and the number of times IMWW was used for more than 10 seconds, which is defined as a “session,” were analyzed.

Potential Effectiveness
To examine the potential effectiveness of IMWW for engaged users, separate paired samples 2-tailed t-tests were undertaken to examine the difference between outcome data collected regarding the severity of anxiety (GAD-7) and low mood (PHQ-8). This analysis was only undertaken for the 54 of 153 (35%) engaged users who completed the outcome measures during onboarding and at the end of the engagement.

Impact of Engagement on Potential Effectiveness
A multivariate linear regression model was used to investigate the impact that engagement with IMWW had on improvement in anxiety and low mood. In particular, the extent to which specific features were used to complete therapeutic exercises impacted on scores over time. Engagement with, and completion of, Worry Time and Problem Solving was expected to lead to improvements in the symptoms of anxiety, and hence a model to analyze this was specified. Because there are multiple discrete interventions being applied within IMWW and the dependent variable is not univariate, the multiple regression $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ was adopted. Within this model, $y$ is the change in GAD-7 or PHQ-8 from the initial score at onboarding to the final input during progress review, $x_1$ is the binary variable indicating whether the user completed Problem Solving and resolved their problem, $x_2$ is the binary variable indicating whether the user completed at least 1 instance of Worry Time in-app, and $\epsilon$ is the stochastic error term. Additional controls were added to the model to examine the extent to which the number of in-app sessions completed, and the number of worries, problems, and solutions recorded predicted improvement in GAD-7 and PHQ-8. All models met OLS model assumptions associated with multicollinearity, heteroskedasticity, and normality of residuals.

Ethical Considerations
Users were only able to download IMWW after agreeing to Iona Mind’s Terms of Service and Privacy policy, which required them to acknowledge that they understand conditions related to use and consent to have their anonymized data processed. Being based on anonymous, routinely collected log data from a nonclinical population, research ethics was not required for this study.

Results

General Engagement
Analysis of log data collected from the 153 engaged users indicated engagement with 1108 sessions (mean 7.2, SD 7.7) with an average session length of 6.2 (SD 6.2) minutes. The number of sessions and session length varied significantly across users with a median session length of 4.5 minutes and 6 being the median number of sessions (Multimedia Appendix 2).

Engagement With LICBT Techniques
Since inclusion (October 2022), 36 users started filling out the CBT model to reflect their current difficulties with anxiety, and of these, 31 (86%) users completed all areas in an average of 2.4 minutes (SD 1.7; median 1.7 minutes). The lesson on the Worry Time technique was successfully completed by almost all users (147/153, 96%). However, of these users, only 50 (33%) were observed to have performed Worry Time at their chosen time using the in-app tools. Problem Solving was engaged with by 114 of 153 (74.5%) users; however, only 89 of 153 (58.2%) users actually completed the lesson. This indicates that 25 of 114 (21.9%) users engaged with the in-app tools to problem-solve one of their practical worries without completing the lesson. This behavior is permitted within the IMWW user experience because the user is able to choose the specific features of the app they wish to engage with. The majority of app features start in an unlocked state to encourage exploration and self-discovery. Only 42 of 153 (27%) users completed a follow-up conversation to review their solutions and progress using the Problem Solving protocol on their worries. During the interaction, engaged users recorded a total of 720 worries (mean 4.7, SD 6.1), and a median of 3
worries were recorded for each user. Of the worries recorded, 399 (55%) were categorized by the user as practical, 306 (43%) as hypothetical, and only 15 (2%) worries were not categorized. With respect to practical worries, 244 (61%) worries were problem-solved using the in-app tools with at least 1 possible solution added.

**Weekly Reflection**

In response to the CA asking the user to reflect on their experience of engaging with IMWW, in-app Weekly Reflections were completed by 58 of 153 (38%) users who recorded 206 responses (mean 3.6) identifying LICBT techniques engaged with, alongside 48 responses identifying the technique found most helpful (Table 2).

During the Weekly Reflection, 27 of 58 (47%) individual users reflected on engaging with Worry Time, of whom 19 (70%) were observed to have used the in-app tooling to complete it at their chosen time. A total of 8 of 58 (14%) users therefore engaged with Worry Time without using the in-app tools. In addition to asking which features of IMWW the user had engaged with, the CA also asked which feature they found most helpful. The users were asked this question during the weekly review, and for each weekly review, they could give at most 1 response.

**Table 2.** Weekly reflection techniques engaged with and found most helpful.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Engaged with (n=206), n (%)</th>
<th>Most helpful (n=48), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journaling worries</td>
<td>50 (24)</td>
<td>16 (33)</td>
</tr>
<tr>
<td>Worry time</td>
<td>46 (22)</td>
<td>11 (23)</td>
</tr>
<tr>
<td>Problem-solving</td>
<td>35 (17)</td>
<td>10 (21)</td>
</tr>
<tr>
<td>Avoiding worry behaviors</td>
<td>28 (14)</td>
<td>4 (8)</td>
</tr>
<tr>
<td>Watching out for different worry types (Worry categorization)</td>
<td>17 (8)</td>
<td>4 (8)</td>
</tr>
<tr>
<td>CBT modelc</td>
<td>8 (4)</td>
<td>1 (2)</td>
</tr>
</tbody>
</table>

aUsers can respond multiple times.
bCBT: cognitive behavioral therapy.
cIntroduced in October 2022.

**Maintaining Engagement**

To maintain engagement with IMWW, 142 of 153 (93%) users were sent at least 1 push notification with 113 of 153 (74%) users responding. An average of 84 (SD 75) push notifications were sent to each engaged user throughout their engagement, although the quantity of push notifications per user varied substantially with use pattern and duration.

**Potential Effectiveness**

Separate paired sample 2-tailed t tests were conducted to examine the difference between the GAD-7 and PHQ-8 scores for 54 of 153 (35%) engaged users who completed the measures during onboarding and the final score provided. There was a significant reduction in both anxiety ($t_{53}=5.5; P<.001; 95\% \text{ CI 2.4-5.2}$) and low mood ($t_{53}=2.3; P=.03; 95\% \text{ CI 0.2-3.3}$), with severity dropping from moderate to mild in both instances (Figure 1).

Examination of individual-level data indicates that the vast majority of users (43/53, 81%) experienced a reduction in anxiety between baseline and final observation with the score of 2 (4%) users remaining unchanged. The majority of users (35/53, 66%) also saw a reduction in PHQ-8 with no difference arising for 4 (8%) users. Deterioration in GAD-7 was experienced by 9 (17%) users and rose to 15 (28%) users for low mood.
Impact of Engagement on Potential Effectiveness

The multivariate linear regression predicting changes in GAD-7 based on engagement reached statistical significance ($F_{2,51}=6.73; P=0.03$), but the model predicting changes in the PHQ-8 did not ($F_{2,51}=2.33; P=0.11$). Two-sided 2-tailed $t$ tests were performed on the slope estimates in the models. The model predicting changes in GAD-7 estimates that the marginal effect of a user completing in-app Worry Time ($\beta_2$) is a $-3.3$ change in GAD-7 and is significant ($P=0.02$). The constant $\beta_0$ and the marginal effect of the user successfully completing Problem Solving $\beta_1$ had respective values of $-1.6$ and $-3.0$. However, they failed to reach significance at $\alpha=0.05$ ($P_0=0.07$, $P_1=0.08$). The model had an $R^2$ of 0.21. Furthermore, an improvement in the GAD-7 and PHQ-8 was not predicted by the number of in-app sessions completed ($P=0.09$) or the number of worries ($P=0.36$), problems ($P=0.27$), and solutions ($P=0.16$) recorded.
Discussion

Principal Findings

While engaged users represented a minority of those who downloaded IMWW, a large number of these interacted with the LICBT techniques associated with the CBT protocol to manage excessive worry and support emotional well-being [31]. The vast majority of those who engaged completed all areas presented with the CBT model and were able to successfully record worries and categorize them as practical or hypothetical. The CA was commonly used to help engaged users overcome difficulties when engaging with practical worries. Worry Time was engaged with to a much lesser extent within the app; however, several users reported engaging with it outside of the app. Forgetting to engage with the LICBT techniques was identified as the most common reason for lack of engagement, while experiencing them as too difficult to comprehend was only reported by a small minority of engaged users. The change in the user GAD-7 score was predominantly explained by engagement with the LICBT techniques as opposed to the number of times they used IMWW.

Poor engagement with an app following download is not uncommon, with only 14% of people often using it the following day [50] and even lower rates typically associated with mental health apps [22]. Despite using common factors and behavior change techniques, however, only a minority of users who engaged with IMWW had enough engagement with the LICBT techniques to be considered engaged users. This is of some concern given that users failing to engage to a point where they have received an adequate dose to bring about recovery may serve as a barrier to seeking further support.

While engagement following download was poor, log data identified that engaged users had moderate to good levels of interaction and fidelity [51], with the CBT model alongside recording and categorizing worries. Fidelity and interaction with Problem Solving were also good, potentially arising from support provided through the CA. When engaging with Problem Solving, the CA was commonly used to support users to overcome difficulties in engaging with the LICBT techniques and to encourage continued engagement. There was less within-app engagement with Worry Time; however, some users reported engaging with it outside of the app. Engagement with the LICBT techniques included within IMWW may therefore have been greater than log data alone suggest. This supports the additional benefits of exploring out-of-app engagement with specific techniques to get a full appreciation of interaction [28]. Exploring ways to promote out-of-app engagement is of benefit given that engagement with techniques in face-to-face CBT between support sessions as “homework” is identified as important to improve clinical outcomes related to anxiety [52].

Although there were moderate levels of interaction with LICBT techniques used within IMWW, exploring additional ways to enhance interaction across all techniques and promote prolonged engagement would be highly beneficial. Enhancing engagement through approaches such as involving personalized support, guidance, and feedback regarding engagement has also been associated with improved effectiveness for mental well-being digital tools [48]. Furthermore, recommendations to enhance out-of-app homework compliance to deliver better outcomes have also been proposed [53]. These include ensuring that app content is congruent to the therapeutic approach adopted, learning is consolidated through engagement, and emphasis is placed on completion. Additionally, recommendations include ensuring that the app is tailored to specific populations and building connections with others has been identified as supporting engagement with homework [53]. Within IMWW the CA was used to enhance engagement through the use of common factor skills to encourage and motivate the user. However, greater focus needs to be directed toward maximizing the ability of the CA to enhance engagement within and outside of the app.

Maximizing engagement may be achieved by implementing mental well-being apps for use adjuvant to health professional support and integrated into clinical settings [54]. Benefits associated with providing support are recognized by the National Institute of Health and Care Excellence recommendations for supported LICBT for anxiety and depression [14]. This has resulted in Psychological Practitioner support for LICBT adopted by the IAPT program implemented across England [14]. Support enables the patient to engage with the interventions by using personalized common factor skills, monitor progress, and provide encouragement during weekly support sessions. However, it does not include a therapeutic role in the delivery of LICBT techniques within the clinical sessions [13].

However, nonprofessional forms of support have also been demonstrated to enhance engagement and improve outcomes with LICBT. For example, group support within community settings is provided by trained volunteers with varying backgrounds [55]. Furthermore, forms of support through technology such as web-based communities providing constructive peer support [55] and discussion forums [56] have been identified to enhance engagement with digital tools [54]. Potentially, therefore, using IMWW adjuvant to some form of minimal-contact support provided by a practitioner, volunteers within community organizations, or mediated through technology offers promise to result in enhanced effectiveness at reduced delivery costs.

With respect to outcomes, the average level of anxiety and low mood improved among users who engaged with IMWW to a degree they would be considered to have received a minimum dose of the LICBT techniques [29]. That anxiety and low mood are identified to share mechanisms has led to recommendations to combine techniques within a single app to reduce the commitment needed by users to maximize engagement [19]. However, when exploring recovery at the level of the individual user, the low mood of several more users deteriorated compared with anxiety. However, it would remain possible to develop a single app that included protocol-informed LICBT techniques to target low mood or anxiety once the main emotional difficulty being experienced was determined.

Strengths and Limitations

Providing a clear description of the LICBT techniques contained within IMWW informed by a theoretical basis represents a real strength of the paper. This has enabled the analysis of log data
to be interpreted with respect to interactions with the techniques. Clearer conclusions regarding the relationship between engagement and outcomes regarding the management of symptoms associated with anxiety were therefore able to be reached. This facilitates specific targeting of future development work on IMWW to ensure greater levels of engagement to derive improved outcomes.

There was a large difference between the number of people who downloaded IMWW and those who interacted with at least 1 lesson related to Worry Time and Problem Solving for them to be considered engaged users. While it is known that the background demographics of these 2 groups did not significantly differ, it is unclear as to why a large number of those who downloaded IMWW never went on to engage with one of these specific LICBT techniques. Unfortunately, reasons behind failing to engage with IMWW were not requested, and therefore the extent to which poor usability may have been a relevant factor is unknown. As the use of digital health technologies continues to increase [16], understanding the usability of apps is of increasing interest [57]. Future research exploring log data could therefore consider using a measure of usability, such as the mHealth Usability Questionnaire [58], alongside the collection of log data to gain a better understanding of the way in which an app is used alongside potential barriers to usability.

Finally, while data regarding outcomes can be considered promising with respect to IMWW as a tool to support worry management, this study does not enable definitive conclusions regarding effectiveness to be reached. As a consequence of the lack of clear end points when using log data, users can stop using the app at any time without completing outcome measures. This makes it difficult to reach conclusions regarding effectiveness. The use of multivariate regression with terms to represent proxy use of techniques was adopted to compensate for this. However, this cannot be considered to represent a substitute for the collection of clear and reliably collected end point data within a trial design comparing IMWW with an appropriate control [59]. Furthermore, reaching conclusions regarding effectiveness is further confounded given that only a minority of engaged users responded to a question regarding the current receipt of treatment.

**Conclusions**

While a large number of people downloaded IMWW, only a minority engaged with the app to be considered engaged users. Of these users, however, analysis of log data identified good interaction with the LICBT techniques associated with an evidence-based protocol to support worry management [31]. Although there were good levels of interaction, exploring additional ways to promote interaction with the LICBT techniques and other features of the app to result in prolonged engagement remains beneficial. This could involve adopting a “user-centric” design process whereby potential users are directly involved in ongoing development [22]. Considering log data as part of a user-centric design process may enhance engagement to a point where more users receive an appropriate “dose” to bring about improvement [29]. Log data can therefore be used to inform ongoing development to maximize engagement and protocol fidelity [51]. This is significant given the relationship between engagement and effectiveness. While effectiveness data associated with IMWW can only be seen as promising, capturing log data will serve to enhance ongoing intervention development. A high-quality randomized controlled trial would then enable definitive conclusions regarding effectiveness to be reached [54]. This would help address concerns that the current level of evidence derived from poor-quality trials does not enable recommendations regarding apps to enhance mental well-being to be reached [60].

**Authors' Contributions**

PF, PA, and PJR conceptualized and designed the project with EW and DR providing theoretical input informing the background of the paper. PF wrote the initial draft of the paper with all authors contributing to the development of the paper, interpretation of the analysis, and editing of the final manuscript. PF was not involved in any part of data analysis.

**Conflicts of Interest**

PF is currently on a paid sabbatical with Iona Mind from the University of Exeter. All other authors have confirmed they have no conflicts of interest to declare.

**Multimedia Appendix 1**
Progress through the Iona Mind Well-being app for Worry management.
[DOCX File, 951 KB - mhealth_v12i1e47321_app1.docx ]

**Multimedia Appendix 2**
Number of sessions of varying length.
[PNG File, 60 KB - mhealth_v12i1e47321_app2.png ]

**References**


17. Smartphone ownership is growing rapidly around the world, but not always equally. Pew Research Center. 2019. URL: http://tinyurl.com/3erjjbjt [accessed 2023-12-01]


Abbreviations

CA: conversational agent
CBT: cognitive behavioral therapy
GAD-7: Generalized Anxiety Disorder–7
IAPT: improving access to psychological therapies
IMWW: Iona Mind Well-being app for Worry management
LICBT: low-intensity cognitive behavioral therapy
PHQ-8: Patient Health Questionnaire–8

©Paul Farrand, Patrick J Raue, Earlise Ward, Dean Repper, Patricia Areán. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 10.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
A Behaviorally Informed Mobile App to Improve the Nutritional Quality of Grocery Shopping (SwapSHOP): Feasibility Randomized Controlled Trial

Carmen Piernas1,2, MSc, PhD; Charlotte Lee1, MSc, DPhil; Alice Hobson1, MSc; Georgina Harmer1, MSc; Sarah Payne Riches1, MPH, DPhil; Michaela Noreik1,3, MSc, PhD; Susan A Jebb1, PhD

1Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom
2Department of Biochemistry and Molecular Biology II, Center for Biomedical Research (CIBM), Institute of Nutrition and Food Technology (INYTA), University of Granada, Granada, Spain
3Faculty of Food and Nutrition Sciences, Hochschule Niederrhein, Mönchengladbach, Germany

Corresponding Author:
Carmen Piernas, MSc, PhD
Nuffield Department of Primary Care Health Sciences
University of Oxford
200 Woodstock Rd
Oxford, OX26GG
United Kingdom
Phone: 44 07752599266
Email: carmen.piernas-sanchez@phc.ox.ac.uk

Abstract

Background: Interventions targeting the nutritional quality of grocery shopping have the potential to help improve diet and health outcomes.

Objective: This study aims to assess the feasibility and acceptability of receiving advice on healthier food purchases through SwapSHOP, a behaviorally informed smartphone app that allows users to scan barcodes of grocery products from the United Kingdom, providing nutritional information and personalized swap suggestions to encourage healthier purchases.

Methods: We randomized adult volunteers in a 6-arm parallel-group controlled feasibility trial. Participants used the SwapSHOP app to record their grocery shopping during a 2-week run-in period and were individually randomized in a 3:1 ratio to either intervention or control arms within 3 strata related to a nutrient of concern of their choice: saturated fat (SFA), sugar, or salt. Participants randomized to the intervention received the SwapSHOP app with a healthier swap function, goal setting, and personalized feedback. Participants in the control group were instructed to use a simpler version of the app to log all their food purchases without receiving any guidance or advice. The primary outcome was the feasibility of progression to a full trial, including app use and follow-up rates at 6 weeks. The secondary outcomes included other feasibility outcomes, process and qualitative measures, and exploratory effectiveness outcomes to assess changes in the nutrient content of the purchased foods.

Results: A total of 112 participants were randomized into 3 groups: SFA (n=38 intervention and n=13 control), sugar (n=40 intervention and n=15 control), and salt (n=5 intervention and n=1 control, not analyzed). The 2 progression criteria were met for SFA and sugar: 81% (30/37) and 87% (34/39) of intervention participants in the SFA and sugar groups, respectively, used the app to obtain healthier swaps, and 89% (68/76) of intervention participants and 96% (23/24) of control participants completed follow-up by scanning all purchases over the follow-up period. The process and qualitative outcomes suggested that the intervention was acceptable and has the potential to influence shopping behaviors. There were reductions of −0.56 g per 100 g (95% CI −1.02 to −0.19) in SFA and −1 g per 100 g (95% CI −1.97 to −0.03) in total sugars across all food purchases in the intervention groups.

Conclusions: People were willing to use the SwapSHOP app to help reduce sugar and SFA (but not salt) in their grocery shopping. Adherence and follow-up rates suggest that a full trial is feasible. Given the suggestive evidence indicating that the intervention resulted in reductions in sugars and SFA, a definitive trial is necessary to target improvements in health outcomes.

Trial Registration: International Standard Randomised Controlled Trial Number ISRCTN13022312; https://doi.org/10.1186/ISRCTN13022312

(JMIR Mhealth Uhealth 2024;12:e45854) doi:10.2196/45854

https://mhealth.jmir.org/2024/12/e45854

JMIR Mhealth Uhealth 2024 | vol. 12 | e45854 | p.489

(page number not for citation purposes)
KEYWORDS
swaps; mobile app; supermarket; food purchases; diet; randomized controlled trial; mobile phone

Introduction

Background
Consumption of saturated fat (SFA), sugars, and salt in the United Kingdom and most high-income countries currently exceeds dietary recommendations for good health [1]. These nutrients of concern contribute to the burden of diabetes and cardiovascular disease, either directly or through their effects on blood cholesterol, blood pressure, insulin sensitivity, and body weight [2–8]. Despite decades of nutrition education and efforts to promote healthier behaviors, significant and sustained dietary changes are yet to be observed. Evidence suggests that people need more than general knowledge of dietary recommendations to change behavior [9–11]; however, there is limited evidence for individual-level interventions that are feasible and effective in supporting dietary change at the population level.

Food purchasing is a key antecedent of food consumption, and improving the nutritional quality of food purchases presents a clear opportunity to intervene. Grocery stores account for 71% of the weekly expenditure on food and drinks, including a substantial proportion of foods high in SFA, sugar, and salt [12]. For many foods, there are usually alternatives with less SFA, sugar, or salt, which are functionally similar. This variation can be attributed partly to the natural differences in recipes, such as those found in ready meals or variances in ingredient costs. In addition, it can also be linked to product reformulation driven by consumer demand, such as the introduction of low-fat yogurts, or policy-driven actions, such as changes in soft drink formulations [13].

Systematic reviews have identified some intervention components to support individual dietary change, including providing information, tailored dietary advice, self-monitoring, and personalized feedback [9,14,15]. Specifically, recommending healthier swaps at the point of purchase or as part of individually tailored regular feedback has shown potential to help improve the nutritional quality of grocery shopping [16–20]. Technological advances such as smartphone apps can facilitate this by providing scalable, lower-cost support to help people make healthier choices while shopping.

A systematic review of smartphone apps identified only 2 studies that tested apps that can provide healthier alternatives at the point of choice [21]. Although there is evidence of similar apps in the scientific literature, these previous apps mostly provide healthier swaps to consumers without the option of self-monitoring the quality of their grocery shopping, and many apps have limitations regarding how information is presented to consumers or the limited comprehensiveness of the databases covering the UK food market [22,23]. Our recent proof-of-concept study tested the functionality of the SaltSwap app to help find lower-salt foods when grocery shopping, which included behavioral components such as nutritional information, prompts to lower-salt options, goal setting on swaps, feedback on swaps and salt reduction, and history of swaps [20].

Objectives
For this study, we have further developed the SaltSwap app into SwapSHOP, a new app that also provides swaps for SFA, sugars, and salt. This study aimed to assess the feasibility and acceptability of receiving dietary advice through the SwapSHOP app. It is a stand-alone intervention that allowed users to scan barcodes of grocery products from major UK stores to obtain nutritional information and suggestions for personalized swaps with lower SFA, sugar, or salt and also enabled users to set goals and monitor their shopping behavior.

Methods

Study Design
This was a prospectively registered (International Standard Randomised Controlled Trial Number: ISRCTN13022312) randomized 6-arm parallel-group controlled feasibility trial conducted in the United Kingdom. After giving written informed consent to participate in the study and completing screening and baseline assessment, participants entered a 2-week run-in period where they used a basic version of the SwapSHOP app (no swap or behavioral functionality shown) to record their grocery shopping. They were individually randomized to 1 of the intervention arms or control following a 3:1 ratio (intervention: control) within 3 strata related to a nutrient of concern of their choice (SFA, sugar, or salt). Individuals participated in the study for 6 weeks from screening to the final follow-up.

Ethical Considerations
This study was reviewed and approved by the University of Oxford Medical Sciences Interdivisional Research Ethics Committee (R67216/RE001). Informed consent was obtained from all the participants.

Participants and Procedures
Participants were recruited between April and October 2021 through the community (eg, word of mouth) and through social networks and media, online newsletters and newspapers, and electronic mailing lists. Participants were eligible if they were aged ≥ 18 years, were willing and able to give informed consent, were English speaking and able to understand the demands of the study, were looking for support to improve their diet, owned a smartphone with access to the internet and an email account, were willing to download and use a smartphone app to scan and track their grocery shopping for the 6 weeks during their participation in the study, were responsible for at least some of their household grocery shopping, and shopped regularly (eg, at least once a week in a physical store or online at any of the 6 major food retailers in the United Kingdom: Tesco, Sainsbury’s, Waitrose, Asda, Morrisons, or Iceland). People were not eligible if they were already on a clinician-supervised diet or a restricted diet, were currently using apps to monitor the quality of their food shopping (excluding apps to track and monitor food intake), were currently participating in another study aimed at dietary change or asking them to change the way...
they shop for food, or were planning on going away from home (holiday or other) for >2 consecutive weeks following enrollment in the study.

Eligible participants were invited to complete a baseline web-based questionnaire that collected demographic information, relevant self-reported medical history, and details about their shopping behaviors. Participants also reported the nutrient (SFA, sugar, or salt) they were most concerned about in their diet, which was used as a stratification factor during the randomization process. Participants then entered a 2-week run-in period during which they were asked to record all their food purchases, either by scanning the barcode of purchased grocery products or manually inputting what was purchased using a simpler version of the SwapSHOP app. Only participants who completed this task with good engagement with the app were randomized. Good engagement was defined as scanning products in at least 2 shops, each with products from at least 3 different predefined food categories (eg, fresh meat, chilled ready meals, and soft drinks), and scanning products from ≥5 of the food categories across the 2 weeks.

Randomization and Blinding

After the 2-week run-in period, participants were individually randomized to either an intervention arm or control following a 3:1 ratio (intervention:control) within each stratum depending on their chosen nutrient of concern (sugar, salt, or SFA). Participants who did not complete the baseline data collection were not randomized. Randomization was performed with a computer using a random allocation sequence (concealed to the investigators) and was stratified by the nutrient of concern using block randomization with block sizes of 4 and 8.

It was not possible to blind the participants to the intervention because of the nature of the intervention; however, randomization was performed remotely via computer-generated randomization, and the researchers were not aware of the treatment group until after randomization was complete.

Intervention

Participants randomized to the intervention arms were able to access an enhanced version of the SwapSHOP app, which included a healthier swap function, goal setting, and personalized feedback. Participants scanned the barcode of grocery products to receive nutritional information about the product using the UK traffic light system and were presented with healthier alternatives or swaps that were lower in SFA, sugar, or salt (depending on group allocation). The swaps that appeared after scanning the original product were chosen from a list of products within the same food category, which were ranked from a larger to a smaller reduction in the specific nutrient that was initially chosen. Participants could also set goals for the number of swaps they wanted to make in each shopping trip and record the products they swapped in the app. The app had specific functions to provide feedback on the achievement of their goals and feedback on the overall nutrient reduction achieved by making swaps. The app also had a section to display the ranking of all purchased foods contributing the most to each nutrient of concern (Figure S1 in Multimedia Appendix 1).

The SwapSHOP app included a comprehensive database of >70,000 grocery products available in January 2021 within major UK grocery stores (including Morrisons, Sainsbury’s, Tesco, Waitrose, Iceland, and Asda) [24] and a bespoke system for categorizing products and selecting suitable alternative swap suggestions. SwapSHOP was based on a previous version developed exclusively for salt, and its theoretical basis is grounded in the Behavior Change Wheel [20,25], a framework that integrates the elements of capability, opportunity, and motivation, which are key for behavior change. A range of behavior change techniques from the taxonomy groups goals and planning and feedback and monitoring that were incorporated into this intervention have been associated with successful dietary change [9,26].

To enable assessment of changes in the nutritional composition of the shopping baskets, participants received weekly reminders to continue scanning and recording all their purchased products for the remaining 4 weeks of follow-up (with a minimum of 2 full weeks of grocery shopping data during the follow-up period).

Comparator

Participants randomized to the control arm were asked to continue using the simpler version of the app with no product nutrition information, swaps, or behavioral components, solely to record all their food purchases as part of the outcome assessment at the end of the 6-week follow-up.

Outcome Measures

Primary Outcomes

The primary objective of this study was to determine the feasibility of a larger study to evaluate the effectiveness of an app that offers healthier swaps while grocery shopping to improve the nutritional quality of food purchasing with respect to sugar, salt, or SFA. The primary outcomes were prespecified progression criteria as follows:

1. App use: at least 70% of participants in the active intervention group use the app to obtain swaps on at least 1 occasion by the end of the second week after randomization.
2. Follow-up rate: at least 60% of participants in total (intervention and control) complete follow-up by scanning all their purchases for a minimum of 2 weeks over the entire follow-up period (4 weeks).

Secondary Outcomes

Feasibility outcomes included (1) recruitment rates: total recruited (including number signed up, eligible, consented, and randomized), (2) time needed for recruitment of the final sample, (3) outcome reporting: number of participants who failed to scan their purchases for a minimum of 2 weeks during follow-up, and number of participants who failed to complete the end-of-study questionnaires.

Process Evaluation and Qualitative Outcome Measures

A summary of app-related use (within-app automatic recording) and acceptability measures was collected through the end-of-study questionnaires at follow-up: (1) average number
of shopping trips where the app was used to scan products to obtain a swap per week; (2) number of occasions the app was used to scan products for a swap per trip; (3) number of swaps made overall per week and per shopping trip; (4) nutrient (SFA, sugar, and salt) reduction per swap made; (5) use of specific functionality, for example, goal setting and feedback; (6) end-of-study questionnaires with free text to understand app use and acceptability of the swaps, app functionality (eg, if app scans most products), and if this prompted other behaviors such as reading nutrition labels.

**Exploratory Effectiveness Outcomes**

Measures included changes in the nutrient content (SFA and sugar in grams per 100 g) of household food purchases recorded in the app between baseline and follow-up in the intervention group compared with the control group.

**Statistical Analysis and Sample Size**

This study was planned as a feasibility study. A sample size of 120 (n=approximately 30 participants per intervention group and n=approximately 10 participants per control group) was prespecified as sufficient to enable testing of the trial methodology and outcome measures using parametric statistical models. The protocol was prospectively registered (International Standard Randomised Controlled Trial Number: ISRCTN13022312).

Baseline characteristics overall and by trial arm were summarized as means and SDs for continuous variables and counts and percentages for categorical variables. Baseline characteristics were coded as age (years), gender (man, woman, other, or not specified), ethnic group (Asian, Black, White, mixed, or other), education (no formal qualifications, secondary education, or higher education), income (≥£15,000, £15,000-£24,999, £25,000-£39,999, £40,000-£75,000, and >£75,000; GBP £1=US $1.38), household size (1, 2-4, and ≥5), grocery shopping habits (eg, spending ≥£25) on groceries more than once a week, once a week, once a fortnight, once a month, and less than once a month), and relevant self-reported health history (eg, concerns related to weight, high blood pressure, diabetes, and heart disease).

Descriptive statistics were used to present the primary and secondary outcomes using all available data, regardless of whether the participants completed the trial or withdrew.

For the exploratory effectiveness measures, we used data from products purchased over the 2 weekly periods recorded at the beginning and end of the trial, with available information on weight or volume as well as the nutrient content (grams per 100 g) to estimate changes in SFA and sugar for food purchases recorded in the app. The excluded items in this analysis included those categorized as fresh fruits, vegetables, or with no or minimal nutrient content (eg, sugar-free gum) as well as products that were manually entered in the app or those missing nutrient or volume information. We used linear regression models to investigate changes in the nutrient content of food purchases between baseline and follow-up in the intervention and control groups. Tests for linear regression assumptions were conducted and met. The main models included adjustment for baseline values of the dependent variable. Potential confounding because of imbalance in baseline characteristics was investigated, and the following variables were adjusted in sensitivity analyses: age (years), gender (man, woman, other, or not specified), ethnic group (Asian, Black, White, mixed, and other), and income (>£15,000, £15,000-£24,999, £25,000-£39,999, £40,000-£75,000, and >£75,000). StataSE (version 16; StataCorp) was used for all the analyses. A P value of <.05 was set to denote statistical significance.

A descriptive analysis of the qualitative outcome measures was conducted using the method by Braun and Clarke [27] for thematic analysis in NVivo 1 software (Lumiervo). Each response was line-by-line coded, and codes were inductively constructed based on the aim of the study. We then organized codes into subthemes using the One Sheet of Paper technique and produced top-level themes [28].

**Patient and Public Involvement**

The SwapSHOP app was based on a previous version developed exclusively for salt reduction [20], which was conceived and tested using input from a patient and public involvement panel. People told us that they would also value an app to help change other aspects of their diet. The current app was also beta tested by members of the public recruited from the community.

**Results**

**Participant Characteristics**

Study participants were recruited between April and October 2021. Of the 289 interested participants, 190 (65.7%) were eligible, provided consent, and entered the 2-week run-in period. Of these 190 participants, 78 (41%) were excluded because they did not download or use the app (n=49, 63%) or did not fulfill the progression criteria based on their engagement with the app (n=29, 37%). The final sample of 38.8% (112/289) of participants successfully completed the task and were randomized (Figure 1). A total of 51 participants were randomized to the SFA group (n=38 to intervention and n=13 to control), 55 participants were randomized to the sugar group (n=40 to intervention and n=15 to control), and 6 participants were randomized to the salt group (n=5 to intervention and n=1 to control). Of the randomized participants, 100 (89%) completed the study and were analyzed as follows: 49 (96%) in the SFA group and 51 (93%) in the sugar group. A total of 2 participants in the SFA group and 4 participants in the sugar group were withdrawn from the study and excluded from the analysis because their data indicated fraudulent activity (eg, fake phone numbers and implausible shopping patterns). Data from 6 participants who were randomized to the salt group were not analyzed, as this group did not reach the target sample by the end of the recruitment period.
Participants in the analytic sample were, on average, aged 35 (SD 12) years, mostly women (80/100, 80%) of White ethnicity (79/100, 79%), and from higher education backgrounds (73/100, 73%; Table 1). Participants mostly lived in households with 2 to 4 other members, shopped once (61/100, 61%) or more than once a week (26/100, 26%), shopped mostly in larger grocery stores (89/100, 98%), or shopped on the internet (39/100, 39%). Most participants looked at SFA or sugar information on nutrition labels always, often, or sometimes and approximately half of the sample (54/100, 54%) reported concerns related to weight among the relevant medical conditions.
## Table 1. Baseline characteristics of participants in the analytic sample.

<table>
<thead>
<tr>
<th></th>
<th>Total population (n=100)</th>
<th>Sugar group</th>
<th>Saturated fat group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control (n=12)</td>
<td>Intervention (n=39)</td>
<td>Control (n=12)</td>
</tr>
<tr>
<td>Age (y), mean (SD)</td>
<td>35 (12)</td>
<td>36 (8)</td>
<td>35 (10)</td>
</tr>
<tr>
<td>Gender (woman), n (%)</td>
<td>80 (80)</td>
<td>9 (75)</td>
<td>33 (85)</td>
</tr>
<tr>
<td>Ethnic group, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black or Asian</td>
<td>9 (9)</td>
<td>1 (8)</td>
<td>6 (15)</td>
</tr>
<tr>
<td>White</td>
<td>79 (79)</td>
<td>10 (83)</td>
<td>27 (69)</td>
</tr>
<tr>
<td>Mixed, other, or not specified</td>
<td>12 (12)</td>
<td>1 (8)</td>
<td>6 (15)</td>
</tr>
<tr>
<td>Education, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No qualifications or not specified</td>
<td>2 (2)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Secondary education</td>
<td>25 (25)</td>
<td>0 (0)</td>
<td>12 (31)</td>
</tr>
<tr>
<td>Higher education</td>
<td>73 (73)</td>
<td>12 (100)</td>
<td>27 (69)</td>
</tr>
<tr>
<td>Individual income**, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;£15,000</td>
<td>28 (28)</td>
<td>3 (25)</td>
<td>9 (23)</td>
</tr>
<tr>
<td>£15,000-£24,999</td>
<td>24 (24)</td>
<td>4 (33)</td>
<td>10 (26)</td>
</tr>
<tr>
<td>£25,000-£39,999</td>
<td>31 (31)</td>
<td>4 (33)</td>
<td>10 (26)</td>
</tr>
<tr>
<td>£40,000-£75,000</td>
<td>11 (11)</td>
<td>1 (8)</td>
<td>6 (15)</td>
</tr>
<tr>
<td>&gt;£75,000</td>
<td>1 (1)</td>
<td>0 (0)</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Household size, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>29 (29)</td>
<td>2 (17)</td>
<td>11 (28)</td>
</tr>
<tr>
<td>2-4</td>
<td>64 (64)</td>
<td>8 (67)</td>
<td>24 (62)</td>
</tr>
<tr>
<td>≥5</td>
<td>7 (7)</td>
<td>2 (17)</td>
<td>4 (10)</td>
</tr>
<tr>
<td>Frequency of grocery shopping (≥£25 per trip), n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>More than once a week</td>
<td>26 (26)</td>
<td>7 (58)</td>
<td>8 (21)</td>
</tr>
<tr>
<td>Once a week</td>
<td>61 (61)</td>
<td>4 (33)</td>
<td>26 (67)</td>
</tr>
<tr>
<td>Once a fortnight</td>
<td>11 (11)</td>
<td>1 (8)</td>
<td>3 (8)</td>
</tr>
<tr>
<td>Once a month</td>
<td>1 (1)</td>
<td>0 (0)</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Less than once a month</td>
<td>1 (1)</td>
<td>0 (0)</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Type of grocery shop usually visited, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supermarkets</td>
<td>98 (98)</td>
<td>12 (100)</td>
<td>38 (97)</td>
</tr>
<tr>
<td>Online supermarkets</td>
<td>30 (39)</td>
<td>5 (42)</td>
<td>11 (28)</td>
</tr>
<tr>
<td>Corner shop or convenience store</td>
<td>32 (32)</td>
<td>1 (8)</td>
<td>14 (36)</td>
</tr>
<tr>
<td>Greengrocers or fruit and vegetable shop</td>
<td>12 (12)</td>
<td>3 (25)</td>
<td>5 (13)</td>
</tr>
<tr>
<td>Butchers or meat market</td>
<td>11 (11)</td>
<td>0 (0)</td>
<td>6 (15)</td>
</tr>
<tr>
<td>Other fresh food markets</td>
<td>17 (17)</td>
<td>4 (33)</td>
<td>7 (18)</td>
</tr>
<tr>
<td>Looking at salt in nutrition labels, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Always or often</td>
<td>14 (14)</td>
<td>2 (17)</td>
<td>3 (8)</td>
</tr>
<tr>
<td>Sometimes</td>
<td>24 (24)</td>
<td>3 (25)</td>
<td>12 (31)</td>
</tr>
<tr>
<td>Rarely or never</td>
<td>62 (62)</td>
<td>7 (58)</td>
<td>24 (62)</td>
</tr>
<tr>
<td>Looking at sugar in nutrition labels, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Always or often</td>
<td>42 (42)</td>
<td>2 (17)</td>
<td>19 (49)</td>
</tr>
</tbody>
</table>
Looking at fat in nutrition labels, n (%)  

<table>
<thead>
<tr>
<th></th>
<th>Total population (n=100)</th>
<th>Sugar group</th>
<th>Saturated fat group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control (n=12)</td>
<td>Intervention (n=39)</td>
<td>Control (n=12)</td>
</tr>
<tr>
<td>Sometimes</td>
<td>30 (30)</td>
<td>6 (50)</td>
<td>10 (26)</td>
</tr>
<tr>
<td>Rarely or never</td>
<td>28 (28)</td>
<td>4 (33)</td>
<td>10 (26)</td>
</tr>
</tbody>
</table>

Relevant health conditions, n (%)  

<table>
<thead>
<tr>
<th></th>
<th>Control (n=12)</th>
<th>Intervention (n=39)</th>
<th>Control (n=12)</th>
<th>Intervention (n=37)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concerns related to weight</td>
<td>54 (54)</td>
<td>7 (58)</td>
<td>18 (46)</td>
<td>5 (42)</td>
</tr>
<tr>
<td>High blood pressure</td>
<td>5 (5)</td>
<td>0 (0)</td>
<td>2 (5)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>11 (11)</td>
<td>2 (17)</td>
<td>4 (10)</td>
<td>1 (8)</td>
</tr>
<tr>
<td>Heart disease</td>
<td>4 (4)</td>
<td>0 (0)</td>
<td>1 (3)</td>
<td>1 (8)</td>
</tr>
</tbody>
</table>

Relevant current medications, n (%)  

<table>
<thead>
<tr>
<th></th>
<th>Control (n=12)</th>
<th>Intervention (n=39)</th>
<th>Control (n=12)</th>
<th>Intervention (n=37)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High blood pressure</td>
<td>4 (4)</td>
<td>0 (0)</td>
<td>1 (3)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>6 (6)</td>
<td>0 (0)</td>
<td>2 (5)</td>
<td>1 (8)</td>
</tr>
<tr>
<td>Heart disease</td>
<td>1 (1)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>1 (8)</td>
</tr>
</tbody>
</table>

The number of participants recruited for the salt group did not reach the target sample size. Data from the participants in this group were not analyzed further.

GBP £1=US $1.38.

Primary Outcomes  

The number of participants recruited for the salt group did not reach the target sample size despite additional recruitment attempts, suggesting that a full-scale trial for salt reduction would not be feasible. Data from the participants in this group were not analyzed further. For participants randomized to the SFA and sugar groups, the 2 progression criteria were met above the set thresholds (Table 2). Overall, most intervention participants (30/37, 81% in the SFA group and 34/39, 87% in the sugar group) used the app to obtain swaps on at least 1 occasion by the end of the second week after randomization. In addition, 89% (68/76) of the intervention participants and 96% (23/24) of the control participants completed follow-up by scanning all purchases for a minimum of 2 weeks over the entire follow-up period.
Table 2. Primary and secondary outcomes—progression criteria and feasibility outcomes.­

<table>
<thead>
<tr>
<th>Total population, n (%)</th>
<th>Sugar group, n (%)</th>
<th>Saturated fat group, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total (N=100)</td>
<td>Control (n=24)</td>
<td>Intervention (n=76)</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>64 (84)</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>34 (87)</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>30 (81)</td>
</tr>
<tr>
<td>Control (n=12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intervetion (n=39)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11 (92)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>36 (92)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12 (100)</td>
<td>32 (86)</td>
</tr>
<tr>
<td>Total (N=100)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>23 (96)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>68 (89)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12 (100)</td>
<td></td>
</tr>
</tbody>
</table>

Primary outcomes—progression criteria

- Participants used the app to obtain healthier swaps on at least 1 occasion by the end of the second week after randomization: 91 (91) in the SFA group and 23 (96) in the sugar group.
- Participants completed follow-up by scanning all purchases for a minimum of 2 weeks over the entire follow-up period: 11 (92) in the SFA group and 36 (92) in the sugar group.

Secondary outcomes—feasibility outcomes

- Signed-up participants: 289 (100) in the SFA group and 226 (78) in the sugar group.
- Eligible participants: 197 (68) in the SFA group and 197 (68) in the sugar group.
- Consented participants: 141 (49) in the SFA group and 141 (49) in the sugar group.
- Completed baseline assessments: 106 (100) in the SFA group and 106 (100) in the sugar group.
- Randomized participants: 16 (15) in the SFA group and 16 (15) in the sugar group.
- Failed to complete follow-up or end-of-study questionnaire: 0 (0) in the SFA group and 0 (0) in the sugar group.

a Percentage of eligible, consented, and completed participants was calculated based on the number of signed-up participants originally; the number of randomized participants (total population) excluded those randomized to the salt group, as these were not analyzed.

b The number of participants recruited for the salt group did not reach the target sample size. Data from the participants in this group were not analyzed further.

c N/A: not applicable.

Secondary Outcomes

The secondary outcomes provided evidence that a larger trial aiming at reducing sugars or SFA would recruit enough participants that adhere to the trial procedures (ie, completing baseline assessments) within the set time frames and that 84% (16/100 failed to complete follow-up) of those randomized would complete follow-up assessments.

Process Evaluation and Qualitative Outcomes

Participants randomized to the SwapSHOP intervention used the app regularly (Table 3). The average number of shopping trips where the app was used to obtain a swap was 5 in the SFA group (92% of shopping trips) and 5.4 in the sugar group (83% of shopping trips). The average number of occasions when the app was used to scan products for a swap in each shopping trip was 2.5 times in the SFA group and 3.3 in the sugar group. Overall, participants set goals averaging approximately 2 swaps per shopping trip. The results showed that individual product swaps were associated with an average reduction in total sugars (−12.5 g of sugar per 100 g, SD 6.97) or in SFA (−4 g of SFA per 100 g, SD 2.16) in the sugar and SFA groups, respectively.
Table 3. Secondary outcomes—process evaluation measures among participants using the swap function^a,b.^  

<table>
<thead>
<tr>
<th>Measure</th>
<th>Sugar group (n=34), mean (SD)</th>
<th>Saturated fat group (n=30), mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average number of shopping trips where the app was used to obtain a swap</td>
<td>5.41 (5.14)</td>
<td>5.07 (3.39)</td>
</tr>
<tr>
<td>Percentage of total shopping trips where the app was used to obtain a swap</td>
<td>83.15 (24.74)</td>
<td>91.62 (18.83)</td>
</tr>
<tr>
<td>Occasions the app was used to scan products for a swap per shopping trip</td>
<td>3.31 (2.30)</td>
<td>2.49 (1.94)</td>
</tr>
<tr>
<td>Average swap goals set in the app per shopping trip</td>
<td>2.14 (1.15)</td>
<td>2.01 (1.54)</td>
</tr>
<tr>
<td>Sugar reduction per swap (grams per 100 g)</td>
<td>−12.47 (6.97)</td>
<td>−0.48 (6.70)</td>
</tr>
<tr>
<td>Saturated fat reduction per swap (grams per 100 g)</td>
<td>−1.62 (2.57)</td>
<td>−4.02 (2.16)</td>
</tr>
</tbody>
</table>

^a^Sugar and SFA reduction per swap per 100 g were calculated as the average change in nutrient per 100 g of product across all swaps made during the intervention period.

^b^The number of participants recruited for the salt group did not reach the target sample size. Data from the participants in this group were not analyzed further.

The results of the qualitative research relating to the acceptability of the app, the feedback and swaps provided through the app, and the usefulness and comprehensiveness of the app were summarized into 4 themes (Table S1 in Multimedia Appendix 1). Overall, most participants shared positive experiences of the intervention, noting that the app was helpful and the swap suggestions were acceptable. They valued the novelty of the app and the traffic light food labeling system, which encouraged them to read food labels. The key barriers to usability were that some grocery stores were not supported by the app and that there was poor product coverage within the app database. Although participants could manually enter their food purchases to find alternatives, they reported that this was time consuming and a barrier to engagement. Some participants also found it difficult to locate the suggested swaps in the store or were frustrated that the suggested swaps were not available in the store or that the product database in the app did not include own-label products in some stores. One significant obstacle to accepting the intervention was the specificity of certain swap suggestions. These suggestions were not always direct replacements or tailored to personal or household dietary preferences, which made it challenging for shoppers to act upon the prompts. In other instances, some participants found it challenging to reduce one nutritional component if it involved an increase in another, and they expressed a desire for a swap suggestion that recognized the overall healthiness of the product. Participants suggested that more information on the fiber composition of swaps, recommended portion sizes, and price comparison would inform their purchasing behavior. Most participants noted that other visual self-monitoring techniques would improve engagement with the intervention.

**Exploratory Effectiveness Outcomes**

Baseline and follow-up data on food purchases with available information on volume and nutrient content were available for 86% (44/51) of the participants in the sugar group and 82% (40/49) of the participants in the SFA group. This analysis included all food purchases recorded in the app but excluded purchases that were entered manually (22%) or those with no, minimal, or missing nutritional information (16%). The degree of missingness in the food purchasing data was comparable between the groups (manually entered products: 23% intervention vs 19% control group and missing nutrient information: 15% intervention vs 16% control group).

There was evidence of changes in the intended direction in both the intervention groups (Table 4). The sugar group reduced total sugars in their grocery purchases by −1 (95% CI −1.97 to −0.03) g/100 g, whereas the control group reduced total sugars in their grocery purchases by 0.32 (−1.47 to 2.11) g/100 g, though the differences between groups were not significant (−0.68, 95% CI −1.94 to 0.58 g/100 g; *P*=.28 adjusted for baseline values). The SFA group reduced total SFA in food purchases by −0.56 (95% CI −1.02 to −0.19) g/100 g, and the control group increased total SFA in food purchases by 0.52 (95% CI −0.19 to 1.22) g/100 g, with a significant between-group difference of −1.05 (95% CI −1.83 to −0.27) g/100 g and *P*=.009 adjusted for baseline values. These results were robust in sensitivity analyses adjusted for age, gender, ethnicity, and income (Table S2 in Multimedia Appendix 1).
Table 4. Secondary outcomes—exploratory effectiveness measures.\(^{a,b}\)

<table>
<thead>
<tr>
<th></th>
<th>Baseline, mean (SD)</th>
<th>Follow-up, mean (SD)</th>
<th>Change, mean (95% CI)</th>
<th>Change adjusted(^b), mean (95% CI)</th>
<th>Between-group difference(^b), intervention vs control Mean (95% CI)</th>
<th>(P) value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Purchased sugar (g/100 g)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sugar group (n=34)</td>
<td>5.13 (2.63)</td>
<td>4.13 (2.34)</td>
<td>−1.00 (−1.97 to −0.03)</td>
<td>−0.86 (−1.67 to −0.05)</td>
<td>−0.68 (−1.94 to 0.58)</td>
<td>.28</td>
</tr>
<tr>
<td>Control (n=10)</td>
<td>4.26 (1.49)</td>
<td>4.58 (1.86)</td>
<td>0.32 (−1.47 to 2.11)</td>
<td>−0.18 (−1.14 to 0.79)</td>
<td>N/A(^c)</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Purchased SFA(^d) (g/100 g)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SFA group (n=28)</td>
<td>2.13 (1.18)</td>
<td>1.58 (0.98)</td>
<td>−0.56 (−1.02 to −0.10)</td>
<td>−0.55 (−0.89 to −0.22)</td>
<td>−1.05 (−1.83 to −0.27)</td>
<td>.009</td>
</tr>
<tr>
<td>Control (n=12)</td>
<td>2.10 (0.84)</td>
<td>2.61 (1.22)</td>
<td>0.52 (−0.19 to 1.22)</td>
<td>0.50 (−0.20 to 1.20)</td>
<td>N/A(^c)</td>
<td>N/A</td>
</tr>
</tbody>
</table>

\(^{a}\)Linear regression adjusted for baseline values.

\(^{b}\)The number of participants recruited for the salt group did not reach the target sample size. Data from the participants in this group were not analyzed further.

\(^{c}\)Not applicable.

\(^{d}\)SFA: saturated fat.

**Discussion**

**Principal Findings**

Participants in this study were willing and able to use the SwapSHOP app as intended, and the research methods ran as planned, with high levels of adherence and follow-up recorded. This provided preliminary evidence of effectiveness to support dietary change to lower the intake of sugars and SFAs. There was little evidence that this general population sample was motivated to reduce salt intake using the SwapSHOP app. Further improvements to the app, especially enhancing the coverage of product data and the specificity of the swap algorithm, are needed to provide a higher quality user experience.

**Comparison With Prior Work**

Previous studies have shown the potential of instore swaps to support healthier choices, but there is very little evidence from interventions involving habitual shopping in physical stores. A quasi-experimental study evaluated the impact of the Change4Life Smart Swaps campaign to promote changes to lower fat or sugar foods when grocery shopping, showing that a higher percentage of participants in the intervention group reported choosing healthier options at the end of the study [29]. A total of 3 randomized controlled trials used a smartphone app to help reduce salt intake by promoting lower-salt swaps at the point of choice, showing changes in purchased salt in the intended direction but without evidence of changes in salt intakes [20,23,30]. Our previous study conducted in primary care settings provided individually tailored regular feedback on food shopping and offered lower-SFA swaps to patients with raised low-density lipoprotein cholesterol. The study showed modest but nonsignificant reductions in SFA consumption and SFA in purchased foods [18]. Other randomized controlled trials conducted in online retail environments, both real and experimental platforms, have shown that offering healthier swaps at the point of choice helps improve the nutritional quality of food shopping [16,17,19].

This SwapShop intervention showed promising signs of early effectiveness, with observed reductions of −1 g/100 g (95% CI −1.97 to −0.03) in total sugars and −0.56 (95% CI −1.02 to −0.19) in SFA in all food purchases in their respective intervention group, though this feasibility study was not powered to detect an effect in purchased nutrients. Data also show that smartphone app use is increasing across a wide range of demographic and age groups [31], and consumers are showing an increasing interest in healthier options [32]. However, although the app may be a useful tool for promoting healthier food choices through personalized advice and support, it is unlikely to be sufficient to achieve dietary recommendations. It is plausible that the impact could be enhanced by interventions to encourage product reformulation to offer greater availability of healthier alternatives. Moreover, price, promotions, positioning, and availability strategies within supermarkets have all been found to be major determinants of food choices [33-36], and these structural interventions are likely to be complementary.

**Strengths and Limitations**

This study’s strength lies in its randomized design and the process evaluation used to investigate the intervention’s pathway to impact. The qualitative components helped to provide context-specific information about the usability of the app and the acceptability of the swap suggestions. The SwapSHOP app had undergone extensive testing before this trial, given that it is an enhanced version of a previously tested app (SaltSwap) that was specifically designed to help people with hypertension choose lower-salt foods when grocery shopping in combination with face-to-face advice from a health care professional [20,37]. Furthermore, beta testing helped refine the app in line with the intended user feedback. SwapSHOP incorporated several behavioral elements with proven evidence to support dietary change [9,14,15], allowing people to set goals to swap to foods,
lower the nutrients of concern, identify major sources of SFA and sugar in their shopping, and provide feedback on achievement of their goal as well as on the overall nutrient reduction achieved through making swaps. A systematic review of interventions using apps to support dietary behavioral improvements suggested greater benefits of multicomponent interventions compared with single-component interventions [38]. Unlike SwapSHOP, the 2 other existing smartphone apps provided information on healthier swaps but did not offer the option of self-monitoring the nutritional quality of food shopping and had limitations regarding how information is presented to consumers or the limited comprehensiveness of the databases covering the UK food market [21,22].

Although the SwapSHOP app included the major UK grocery retailers, participants particularly reported issues related to the product database, the limited range of UK grocery stores, and the improvements needed in the swapping function to create a list of healthier options that are similar to the original product and to accommodate dietary preferences. These aspects must be addressed in future versions of the app to maximize usability and acceptability. Purchased products that were recorded manually by the participants or had missing nutrient information were excluded from the analysis of changes in total purchases, limiting the robustness of the exploratory effectiveness outcome results, although this should not differ by trial arm.

Another limitation is that the study recruited a small, self-selected sample of people who were motivated to take action to improve their diet quality. A large proportion of participants reported receiving higher education and living in less deprived geographical areas than the national average [39]. A lower socioeconomic status is related to poorer dietary quality and health outcomes; hence, the observed results may not represent a wider population with lower adherence to dietary recommendations [40]. Ethnicity is also related to food choice [41]. Although this study mostly included people of White backgrounds, the app database included a wide range of products available in the UK market, thus covering different dietary preferences that apply to other ethnic groups.

Conclusions
SwapSHOP is a behaviorally informed smartphone app that allows users to scan barcodes of grocery products from major UK supermarkets, providing tailored nutritional information and suggesting personalized swaps to support dietary change. This study provided evidence of feasibility as a stand-alone intervention to support motivated individuals wanting to reduce their SFA or sugar intake as well as preliminary evidence of effectiveness to support healthier food purchases. Given the low cost and scalability of this intervention, after further refinement of the app technology and expanded market coverage, a definitive trial is warranted to assess the potential of this tool to improve health outcomes.

Acknowledgments
This study was funded by the National Institute for Health Research (NIHR) Applied Research Collaborations Oxford. SAJ, MN, and AH were funded by the NIHR Oxford Biomedical Research Centre (BRC) Obesity, Diet and Lifestyle Theme. CP, GH, and SAJ were funded by the NIHR Applied Research Collaborations Oxford. CL was funded by the Engineering and Physical Sciences Research Council and the NIHR Oxford BRC. SPR was supported by a British Heart Foundation Clinical Research Training Fellowship (FS/16/34/32211). CP is currently funded by grant RYC2020-028818-I, MCIN/AEI/10.13039/501100011033 and “ESF Investing in your future” (Ministry of Science and Innovation, Spain). The funders had no role in designing the study, data collection, analysis, interpretation of data, writing the report, or the decision to submit the manuscript for publication. The authors would like to thank Peter Scarborough and Richard Harrington for their help and support in using the FoodDB database of food products included in the SwapSHOP app. The University of Oxford is the owner of the SwapSHOP app.

Data Availability
The data sets generated during and/or analyzed during this study are available on reasonable request. Access to the study data and the statistical code will be reviewed and granted upon request by the Nuffield Department of Primary Care PRimDISC committee (primdisc@phc.ox.ac.uk)

Authors’ Contributions
All authors contributed to the concept and design of the study. CP, AH, CL, MN, GH, and SPR developed the study protocols and materials, performed the experiments, and collected the study data. CP and CL analyzed and interpreted the data. CP wrote the manuscript and generated the tables and figures. All authors discussed the results and implications and commented on the final version of the manuscript.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Supplementary tables and figures.
[DOCX File, 692 KB - mhealth_v12i1e45854_app1.docx]
References


Abbreviations

SFA: saturated fat

© Carmen Piernas, Charlotte Lee, Alice Hobson, Georgina Harmer, Sarah Payne Riches, Michaela Noreik, Susan A Jebb. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 11.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Effectiveness of Metaverse Space–Based Exercise Video Distribution in Young Adults: Randomized Controlled Trial

Rami Mizuta¹, MSc, PT; Noriaki Maeda¹, PT, PhD; Tsubasa Tashiro¹*, PT, PhD; Yuta Suzuki², PT, PhD; Sayo Kuroda¹, BA, PT; Ayano Ishida¹, BA, PT; Sakura Oda¹, BA, PT; Tomoya Watanabe¹, BA, PT; Yuki Tamura¹, BA, PT; Makoto Komiya¹, PT, PhD; Yukio Urabe¹*, PT, PhD

¹Department of Sport Rehabilitation, Graduate School of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Japan
²Department of Physical Therapy, Faculty of Rehabilitation, Kyushu Nutrition Welfare University, Fukuoka, Japan
*these authors contributed equally

Corresponding Author:
Yukio Urabe, PT, PhD
Department of Sport Rehabilitation
Graduate School of Biomedical and Health Sciences
Hiroshima University
1-2-3 Kasumi, Minami-ku
Hiroshima, 734-8553
Japan
Phone: 81 082 257 5405
Email: yurabe@hiroshima-u.ac.jp

Abstract

Background: In response to the serious lack of physical activity among young adults, recent attempts have been made to encourage young people to exercise through exercise video distribution. However, merely distributing videos does not lead to improved physical activity levels. Metaverse space, which enables web-based interaction through avatars, allows users to watch exercise videos in the same space as other avatars.

Objective: This study explored whether exercise video distribution using metaverse space is effective in improving physical activity levels, along with mental health and locomotive function, among young people.

Methods: In this parallel-group randomized controlled trial participants were recruited using printed poster displays. A total of 48 young adults aged between 18 and 30 years were assigned to 3 groups of 16 each: the metaverse, YouTube, and control group. To encourage exercise, the metaverse group was given an exercise video each week with a load of around 4–8 metabolic equivalents of tasks (METs) for 8 videos delivered in the metaverse space. The YouTube group was sent a URL on YouTube every week to view exercise videos with the same content as the metaverse group. The control group was given no special instructions. The intervention period was 8 weeks. Pre- and postintervention physical activity, well-being, locomotive syndrome risk tests, and social capital were measured. Although this study was not blinded to the participants, the measurers did not know to which group the participants belonged. Mixed model repeated-measures analyses and a post hoc Wilcoxon signed rank sum test were performed to detect the effects of the intervention in all groups.

Results: The results of the mixed model repeated-measures analyses showed a significant interaction between groups and before and after the intervention for total physical activity (metaverse group: pre 737.1, SD 609.5 METs/week, post 1575.4, SD 1071.8 METs/week; YouTube group: pre 661.7, SD 710.7 METs/week, post 911.9, SD 1103.3 METs/week; and control group: pre 930.6, SD 665.1 METs/week, post 844.7, SD 701.8 METs/week; P= .04) but none for the indicators of well-being (P= .40), locomotive function scale (P= .17), and social capital (P= .23). A post hoc test showed a significant increase in physical activity in the metaverse group before and after the intervention (P= .006).

Conclusions: This study is the first to show that delivering exercise videos through metaverse space can increase physical activity in young adults by providing a gathering space for individuals similarly motivated for exercise practice. This way, the sense of isolation during exercise is reduced compared with merely distributing videos on YouTube. The use of metaverse space in health promotion is likely to spread, and this study provides a useful reference for its exploration.

Trial Registration: ClinicalTrials.gov NCT06019156; https://ichgcp.net/clinical-trials-registry/NCT06019156
Exercise video distribution; exercise; metaverse; physical activity; web-based intervention

Introduction

Physical inactivity among young people is a serious problem, with 40.3% of men and 66.1% of women in their 20s reporting not exercising even once a week, according to the National Health and Nutrition Survey, 2019 [1]. Surprisingly, a study of 100 college students aged between 18 and 23 years found that 65% of them were in the high-risk group for locomotive syndrome, a condition that reduces physical function and mobility such as standing and walking. A study of young Japanese adults aged between 18 and 20 years revealed that 45.9% of them had a prevalence of frailty-prefrailty [2]. As Japan’s aging society progresses, the number of people requiring nursing care also increases, and the reasons for this need are often locomotive disorders such as falls and joint diseases [3,4]. Therefore, establishing exercise habits at a young age and maintaining and improving locomotive function are important.

The spread of COVID-19 has led to a lack of exercise and a change to a more sedentary lifestyle, which has focused attention on exercise interventions over home-based web-based video delivery methods. This promotion of physical activity through exercise videos was intended to help address people’s lack of motivation to engage in physical activity when gymnasiums and recreation centers had to close and physical distance was required [5]. An exercise intervention study using YouTube reported that when a new exercise video of approximately 5 minutes was delivered once a week for 12 weeks, physical activity and frequency of muscle training increased in the exercise intervention group [6]. Hence, the distribution of exercise videos through YouTube, the most accessible tool for young people, has shown some effectiveness and has become commonplace even as of 2023, when COVID-19-related restrictions have eased.

However, this conventional promotion of physical activity through the distribution of exercise videos has elements that fail to encourage behavioral changes. Self-determination theory, as it relates to exercise implementation, begins with the fulfillment of 3 basic human psychological needs: competency (the experience of achievement and efficacy), autonomy (the experience of motivation), and relatedness (the experience of connection with others) [7]. However, traditional exercise video distribution using YouTube and other media has focused on individual video-watching, which fails to provide opportunities for connecting with other people. In fact, another study of an exercise intervention using YouTube reported improvements in mental health with no increase in physical activities [8]. Studies showed connection with others as an important factor in exercise implementation and that being in an exercise group was more effective in increasing physical activity than exercising alone [9,10]. Further, social encouragement by those around an individual and exercising together had a positive impact on the acquisition of exercise habits [11,12]. In addition, people who shared their step count information using a special app tended to exercise the next day when other people “liked” or commented on their step counts [13]. Moreover, the group that was provided with peer interaction and social support showed positive effects on locomotive functions, such as improved results in stand-up tests, compared with the group that lacked these supportive activities [14]. Thus, adding an element of social community when promoting exercise is more effective.

The metaverse is an internet-based 3D virtual world. For the generation of digital natives born in the 1990s and 2000s, it is expected that the metaverse will increasingly become a space where they spend part of their daily lives [15]. In recent years, the metaverse space has gained attention in the health care field [16-18], and a bibliometric analysis of virtual and augmented reality showed that the metaverse has been adapted for diagnosis, surgical treatment, and rehabilitation of pain, stroke, anxiety, depression, fear, cancer, and neurodegenerative diseases, with satisfactory results [16]. In a metaverse space, people can interact with and encourage others on the web through their avatars, which are their own alter egos, and create a community of people who gather in the same space. Social distance is known to interfere with mental health [19], and the metaverse’s ability to recreate a space for social interaction is expected to help improve the treatment of mental health symptoms such as anxiety, stress, and eating disorders [20-23]. Therefore, the use of the metaverse has the potential to improve mental health. Moreover, encouraging physical activity within the metaverse space has the potential to promote behavioral responses from users, such as motivating people to adopt a healthier lifestyle [24]. Therefore, exercise training using the metaverse is expected to develop as a method to promote physical activity among young people [16]; however, there is insufficient knowledge on whether physical activity can be promoted by distributing exercise videos in the metaverse space. This study verified whether exercise video distribution using an app that provides metaverse space is effective in improving physical activity levels, mental health, and locomotive function among young people. We hypothesized that the distribution of exercise videos in the metaverse space would increase physical activity and have positive effects related to mental health and locomotive function.

Methods

Study Design and Participants

This study was a parallel-group, randomized controlled trial. The participants were recruited among Hiroshima University students between August 10 and September 9, 2022. For recruitment, we displayed a poster explaining the study and indicated a contact person available for questions for those considering participation. Several people expressed their willingness to participate. Participants were offered an incentive of US $6.93. Written consent was obtained after we explained the study details face-to-face. All data were collected in person.
This study's specific participation criteria were as follows: individuals (1) aged 18 and 30 years at the time of obtaining consent, (2) who owned a smartphone or computer and had internet access, and (3) who did not belong to an exercise community such as a sports club. The exclusion criteria were as follows: individuals (1) who had a history of a disease that prohibited exercise, (2) whose physical activity was >3000 metabolic equivalents of tasks (METs)/week because a person with >3000 METs/week is defined as active [25-27], and (3) who were (or possibly) pregnant.

The sample size was calculated using G*power 3.1.9.2 (version 3.1.9.7; Heinrich-Heine-University Düsseldorf). When the effect size was calculated using raw data from a previous study with a similar design, which used a web-based exercise intervention with a physical activity measure as the main outcome [28], a large effect size of 0.34 was obtained with a partial η² of 0.108. When setting α error probability to .05, power (1–β error probability) to .8, number of groups to 3, and number of measurements to 2, the required sample size was 27 participants, with 9 participants per group. A total of 48 participants were required when the dropout rate was approximately 40% [8].

After completion of the baseline prequestionnaire and measurements, 3 participants who met the exclusion criteria and were physically active for >3000 METs/week were excluded. The participants were divided into 3 groups: “metaverse space-based exercise video distribution group (metaverse group),” “YouTube-based exercise video distribution group (YouTube group),” and “no video distribution group (control group)” in a 1:1:1 ratio. The allocation was performed by segregating blocks of size 3 to generate the sequence [29], and the allocation order was hidden until after each group was allocated. A total of 3 separate staff members performed each of the 3 tasks: generating the random allocation sequence, enrolling participants, and assigning participants to the interventions. We explained to all participants the nature of the 3 groups while obtaining their consent. After allocation, the metaverse and YouTube groups received an 8-week intervention in which the exercise videos were distributed, and a control group was established. The intervention period was from October 3 to November 27, 2022. Thereafter, a postquestionnaire survey and measurements were conducted for all 3 groups.

**Intervention**

**Procedure**

The recruitment ran for approximately a month, from August 10 to September 9, 2022, and the premeasurements were taken during a 2-week period from September 12 to 25, 2022. The 8-week intervention was conducted from October 3 to November 27, 2022. The postmeasurements were conducted over a period of 1 week (from December 5 to 11, 2022).

**Metaverse Group**

The participants in the metaverse group had access to Metaverse space (Spatial Systems Inc), where they could watch exercise videos whenever they wanted (Figure 1). The researcher created a metaverse space for watching exercise videos and set up new videos each week. The URL to access them was shared only with participants in the metaverse group. The unique feature of the metaverse was that users could enter a room to watch exercise videos with their avatars, which were their own alter egos and other avatars. The participants watched exercise videos, which were created by a physical therapist to increase their physical activity, for approximately 5 minutes [6,8]. Then, we instructed them to do the exercise while watching videos. The load setting was approximately 4-8 METs to train the trunk and the upper and lower extremities. A new video was released once a week [6], and the participants could watch 8 videos for 8 weeks. They were informed in advance that new videos would be updated every Monday and, as a reminder, were sent a SMS text message when a new video was available. A web-based survey regarding exercise video viewing frequency was administered, and adherence was tracked weekly.

Figure 1. The metaverse space in this study.
**YouTube Group**

The YouTube group was sent a message with the URL to access the exercise video every Monday. The YouTube group could watch videos at their preferred times by clicking on a link. The content and time of the video were identical to those of the metaverse group. As in the metaverse group, adherence to exercise implementation was confirmed through a web-based questionnaire, and video update reminders were sent every week.

**Control Group**

Participants were informed by email that they were assigned to the control group simultaneously with the start of the intervention in the Metaverse and YouTube groups. We asked the participants in the control group to spend 8 weeks as usual, without any special instructions or delivery of videos. This group was established as an indicator of the amount of physical activity that may be significantly affected by the infection status and government directives during the COVID-19 pandemic.

**Outcome Measurements**

**Basic Information**

The participants were asked and tested for age, height, weight, BMI, lifestyle, sex, living status, and frailty to characterize the basic information. The weight and frailty tests were measured by a physical therapist, and the other parameters were measured using a questionnaire.

**Lifestyle**

Breslow’s 7 health practice scores were used to assess lifestyle [30,31]. The 7 parameters were: smoking (presence or absence), alcohol intake (high or low: consumption of ≥5 or <5 alcoholic beverages), sleep (≥7 hours or <7 hours), breakfast (eating or not eating), eating between meals (eating or not eating), exercise (≥2 times, 30 minutes/week or <2 times, 30 minutes/week), and proper weight (BMI <25 kg/m² or ≥25 kg/m²).

**Living Status**

Regarding their living status, the participants were asked whether they lived alone or with others.

**Frailty Test**

The Japanese Cardiovascular Health Study criteria questionnaire was designed to indicate potential frailty, and its contents are as follows [32,33]: (1) have you lost 2 kg or more in the past 6 months? (2) Do you engage in moderate levels of physical exercise or sports aimed at improving health? (3) In the past 2 weeks, have you felt tired without reason? (4) Hand grip strength of <26 kg in men or <18 kg in women, and (5) gait speed of <1 m/second. Grip strength was measured using a hand dynamometer (TKK 5401; Takei). Each participant performed 2 trials, and the best value from both trials was used for the analysis. Walking speed was calculated by performing a 10-meter walk test and dividing the measured value by 10 to obtain the results per meter. All questions were rated 1 or 0, and higher scores indicated a higher possibility of frailty. Frailty, prefrailty, and robustness were defined by the number of applicable items: 3-5, 1-2, and 0.

**Primary and Secondary Outcomes**

The study assessed physical activity, well-being, locomotive syndrome risk tests, and social capital before and after the intervention for 8 weeks.

**Primary Outcome: Physical Activity**

The short form of the International Physical Activity Questionnaire was used to measure the effect of an intervention using exercise videos with different distribution methods (Metaverse and YouTube) on physical activity [34]. Total physical activity was measured as the average amount of vigorous physical activity, the amount of moderate activity, and the amount of walking (METs minutes/week). A MET is defined as the amount of energy required for a person to sit quietly. In the premeasurement, the participants answered for the most recent week, and in the postmeasurement, they responded for the week after the intervention was completed.

**Secondary Outcomes: Well-being, Locomotive Syndrome Risk Tests, and Social Capital**

**Well-Being**

To assess well-being and mental health, we used the Japanese version of the World Health Organization-Five Well-being Index [35]. Responses to each item were rated on a 6-point scale from 0 to 5 and consisted of 5 questions. The highest score is 25, with higher scores indicating better well-being.

**Locomotive Syndrome Risk Test**

We used the 25-question geriatric locomotive function scale, proposed by the Japanese Orthopedic Association [33,36], to measure levels of locomotive function based on normal daily activities and pain. All the items were answered on a scale of 0-4. The higher the score, the more impaired the motor function, which indicates a decline in mobility and interference with social life.

**Social Capital**

The social capital questions consisted of 11 items related to civic participation, social cohesion, and reciprocity, with a perfect score of 11 and higher scores indicating stronger social connections [37]. We included this question item because of the possibility that the sense of social participation could change in the metaverse group.

**Supplemental Outcomes: Impression of Each Intervention (Metaverse and YouTube Groups)**

We interviewed each of the metaverse and YouTube groups regarding their impressions of the intervention using exercise videos after the 8-week intervention. User experience is the experience of a user with the use of a product, system, or service. From the review of user experience, we selected the following list of questions: novelty, relatedness (connection with others), motivation, excitement, satisfaction, delight, comfort, attractiveness, expectation, and fulfillment [38]. The participants answered each question with a “good,” “rather good,” “rather poor,” or “poor” response.
Statistical Analysis

Primary and secondary outcomes were analyzed by performing a mixed model repeated-measures ANOVA. This analysis was used to detect the effect of interventions between the metaverse, YouTube, and control groups. We assessed the main effect as well as group and time interactions on the outcome measure. A mixed model repeated-measures analysis is an intention-to-treat analysis with unbiased estimates that considers all available data from participants enrolled in the study [39]. The effect size $r$ for the interaction effect of a mixed model repeated-measures analysis was calculated using F statistics. As a posttest, the Wilcoxon signed rank sum test was performed after confirming nonnormality to compare the pre- and postintervention outcomes of each group. The $z$-statistic was used to calculate the effect size $r$. Supplemental outcomes were analyzed by performing chi-square tests to compare the intervention impressions between the metaverse and YouTube groups. Note that to use the chi-square test, “rather good” was included in “good” and “rather poor” was included in “poor,” thus reorganizing responses into 2 groups. The effect size was calculated using Cramer $V$.

Statistical analyses were performed using SPSS (version 28.0; IBM Corp). The significance level was set at 5%.

Ethical Considerations

This study was approved by the Ethical Committee for Clinical Research at Hiroshima University (C2022-0004) and registered with UMIN (UMIN000048046). We followed the guidelines of the Consolidated Standards for Reporting Trials [40]. There were no significant changes in the method used after the study’s initiation.

Results

Participants

Figure 2 shows the participant flowchart. A total of 51 individuals expressed interest in the study; however, 3 were excluded because their physical activity was >3000 METs/week. All the participants in the metaverse and YouTube groups completed the intervention. Participants were also interviewed weekly during the 8-week intervention period to monitor their exercise implementation. The average frequency of exercise while watching the exercise video per week was 4.1 (SD 3.9) and 2.6 (SD 1.6) for the metaverse and YouTube groups, respectively (response rates: metaverse group 100% and YouTube group 99%). For the postmeasurement, a participant in the control group could not participate owing to illness.

Figure 2. Study flowchart. MET: metabolic equivalents of task.

Basic Information, Lifestyle, Living Status, and Frailty Test

Table 1 presents the participants’ demographic characteristics (mean 22.4, SD 2.4 years; 32/48, 67% women). Prefrailty accounted for 41/48 (85%) participants because we recruited those who were not members of a sports club and whose physical activity was <3000 METs/week.
<table>
<thead>
<tr>
<th>Variables</th>
<th>Total</th>
<th>Metaverse group (n=16)</th>
<th>YouTube group (n=16)</th>
<th>Control group (n=16)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>22.4 (2.4)</td>
<td>22.9 (1.4)</td>
<td>22.5 (2.0)</td>
<td>22.0 (3.4)</td>
</tr>
<tr>
<td>Height (cm), mean (SD)</td>
<td>164.0 (7.9)</td>
<td>162.4 (6.6)</td>
<td>163.9 (6.2)</td>
<td>165.7 (10.3)</td>
</tr>
<tr>
<td>Weight (kg), mean (SD)</td>
<td>57.7 (11.6)</td>
<td>56.7 (9.8)</td>
<td>57.3 (10.2)</td>
<td>59.0 (14.8)</td>
</tr>
<tr>
<td>BMI (kg/m²), mean (SD)</td>
<td>21.3 (2.9)</td>
<td>21.4 (2.5)</td>
<td>21.2 (2.5)</td>
<td>21.3 (3.6)</td>
</tr>
<tr>
<td>Lifestyle, mean (SD)</td>
<td>4.6 (0.9)</td>
<td>4.6 (0.9)</td>
<td>4.4 (0.7)</td>
<td>4.7 (1.1)</td>
</tr>
</tbody>
</table>

**Sex, n (%)**
- Male 16 (33) 6 (38) 6 (38) 5 (31)
- Female 32 (67) 10 (63) 10 (63) 11 (69)

**Living status, n (%)**
- Alone 31 (65) 11 (69) 10 (63) 10 (63)
- With others 17 (35) 5 (31) 6 (38) 6 (38)

**Frailty test, n (%)**
- Frailty 2 (4) 1 (6) 1 (6) 0 (0)
- Prefrailty 41 (85) 13 (81) 14 (88) 14 (88)
- Robust 5 (10) 2 (13) 1 (6) 2 (13)

### Physical Activity, Well-Being, Locomotive Syndrome Risk Test, and Social Capital

Table 2 shows the estimation results of the fixed effects from the mixed model repeated-measure analysis. Only total physical activity showed a significant interaction ($F_{2,45.042}=3.338, P=.04$; effect size=0.263). The main effects for the time showed a significant difference in vigorous activity ($F_{1,45.083}=6.921, P=.01$), locomotive function scale ($F_{1,44.895}=9.557, P=.003$), and social capital ($F_{1,44.650}=5.085, P=.03$).

Table 3 shows the results of the post hoc test comparison of the difference in total physical activity before and after the intervention in each group. In the metaverse group, a significant difference was detected between pre- and postintervention periods ($P=.006$; effect size=0.682).
Table 2. Estimate of the fixed effects from the mixed model repeated-measures analysis, (N=16).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Metaverse group</th>
<th>YouTube group</th>
<th>Control group</th>
<th>Main effect</th>
<th>Interaction effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time</td>
<td>Group</td>
<td>Time×Group</td>
<td>P value</td>
<td>F test (df)</td>
</tr>
<tr>
<td>Physical activity: total (METs (^a) minutes/week), mean (SD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre</td>
<td>5.153 (1, 45.051)</td>
<td>1.383 (2, 45.297)</td>
<td>3.338 (2, 45.042)</td>
<td>.03</td>
<td>.26</td>
</tr>
<tr>
<td>Post</td>
<td>6.921 (1, 45.083)</td>
<td>3.140 (2, 45.257)</td>
<td>1.917 (2, 45.075)</td>
<td>.01</td>
<td>.53</td>
</tr>
<tr>
<td>Physical activity: vigorous (METs minutes/week), mean (SD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre</td>
<td>2.004 (1, 44.498)</td>
<td>0.994 (2, 44.623)</td>
<td>0.748 (2, 44.490)</td>
<td>.16</td>
<td>.31</td>
</tr>
<tr>
<td>Post</td>
<td>0.013 (1, 45.119)</td>
<td>0.192 (2, 45.253)</td>
<td>1.488 (2, 45.110)</td>
<td>.91</td>
<td>.83</td>
</tr>
<tr>
<td>Physical activity: moderate (METs minutes/week)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre</td>
<td>2.449 (1, 44.726)</td>
<td>1.189 (2, 45.186)</td>
<td>0.934 (2, 44.718)</td>
<td>.12</td>
<td>.31</td>
</tr>
<tr>
<td>Post</td>
<td>9.557 (1, 44.895)</td>
<td>0.429 (2, 45.308)</td>
<td>1.821 (2, 44.887)</td>
<td>.3</td>
<td>.65</td>
</tr>
<tr>
<td>Physical activity: walking (METs minutes/week)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre</td>
<td>5.085 (1, 44.650)</td>
<td>2.348 (2, 45.142)</td>
<td>1.519 (2, 44.642)</td>
<td>.03</td>
<td>.11</td>
</tr>
<tr>
<td>Post</td>
<td>15.4 (3.9)</td>
<td>16.4 (4.1)</td>
<td>18.4 (4.1)</td>
<td>.26</td>
<td>.79</td>
</tr>
<tr>
<td>Well-being</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre</td>
<td>17.6 (5.5)</td>
<td>17.3 (3.9)</td>
<td>18.5 (4.8)</td>
<td>.55</td>
<td>.22</td>
</tr>
<tr>
<td>Post</td>
<td>3.5 (3.7)</td>
<td>3.5 (3.7)</td>
<td>3.5 (3.7)</td>
<td>.53</td>
<td>.24</td>
</tr>
<tr>
<td>Locomotive function scale</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre</td>
<td>3.9 (3.4)</td>
<td>3.8 (3.3)</td>
<td>3.5 (3.3)</td>
<td>.15</td>
<td>.65</td>
</tr>
<tr>
<td>Post</td>
<td>1.1 (1.9)</td>
<td>2.6 (2.9)</td>
<td>2.9 (3.3)</td>
<td>.21</td>
<td>.79</td>
</tr>
<tr>
<td>Social capital</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre</td>
<td>5.9 (1.9)</td>
<td>4.9 (1.8)</td>
<td>6.2 (1.3)</td>
<td>.03</td>
<td>.11</td>
</tr>
<tr>
<td>Post</td>
<td>7.1 (2.4)</td>
<td>5.7 (2.0)</td>
<td>6.2 (1.5)</td>
<td>.26</td>
<td>.79</td>
</tr>
</tbody>
</table>

\(^a\)MET: metabolic equivalents of task.
Table 3. Differences between pre- and postintervention outcome measures (N=48).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Physical activity: total (metabolic equivalents of tasks minutes/week)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Metaverse group</strong></td>
<td></td>
</tr>
<tr>
<td>Pre, mean (SD)</td>
<td>737.1 (609.5)</td>
</tr>
<tr>
<td>Post, mean (SD)</td>
<td>1575.4 (1071.8)</td>
</tr>
<tr>
<td>P value</td>
<td>.006</td>
</tr>
<tr>
<td>Effect size</td>
<td>0.682</td>
</tr>
<tr>
<td><strong>YouTube group</strong></td>
<td></td>
</tr>
<tr>
<td>Pre, mean (SD)</td>
<td>661.7 (710.7)</td>
</tr>
<tr>
<td>Post, mean (SD)</td>
<td>911.9 (1103.3)</td>
</tr>
<tr>
<td>P value</td>
<td>.23</td>
</tr>
<tr>
<td>Effect size</td>
<td>0.298</td>
</tr>
<tr>
<td><strong>Control group</strong></td>
<td></td>
</tr>
<tr>
<td>Pre, mean (SD)</td>
<td>930.6 (665.1)</td>
</tr>
<tr>
<td>Post, mean (SD)</td>
<td>844.7 (701.8)</td>
</tr>
<tr>
<td>P value</td>
<td>.57</td>
</tr>
<tr>
<td>Effect size</td>
<td>–0.142</td>
</tr>
</tbody>
</table>

Comparison of Impressions of the Intervention Between the Metaverse and YouTube Groups

Multimedia Appendix 1 and Figure 3 show the results of a chi-square test on impressions of the intervention between the metaverse and YouTube groups. Significant differences were observed between the 2 groups in novelty (P=.003; effect size=0.529), relatedness (P=.03; effect size=0.387), excitement (P=.01; effect size=0.433), delight (P=.002; effect size=0.539), attractiveness (P=.006; effect size=0.487), and expectation (P<.001; effect size=0.626). The results did not reveal significant differences in motivation, satisfaction, comfort, or fulfillment.

Figure 3. Impressions of intervention in the Metaverse and YouTube groups. *P<.05.
**Discussion**

This study examined whether an exercise intervention using exercise videos in the metaverse space has a positive impact on physical activity. The main results show a significant interaction between groups and time (pre- and postintervention) in total physical activity, with post hoc analysis showing a significant increase in total physical activity in the metaverse group after compared with before the intervention and no significant change in the YouTube and control groups. This study is the first to show that using the metaverse space to deliver exercise videos can promote increased physical activity.

The current results support the hypothesis that the use of the metaverse increases physical activity. The exercise videos used in the metaverse, and the YouTube groups are identical, and the metaverse group differs from the YouTube group in that the exercise videos can be viewed in a shared space with other avatars in the metaverse space. Studies have noted that interacting with others is more effective than exercising alone at increasing physical activity and sustaining exercise [9-13]. Metaverse characteristics include the ability to form communities within virtual reality and the persistence of virtual reality regardless of individual user access [15]. In this study, no change in social capital was detected before and after the intervention; however, the percentage of those who had an impression of relatedness (connection with others) was also significantly higher in the metaverse group (9/16, 56%) than in the YouTube group (3/16, 19%)—indicating that more people felt connected to others when using the metaverse than when using YouTube. Additionally, strategies to increase physical activity include improving the physical environment, such as its construction, and the social environment, including social support [41]. The use of the metaverse resulted in the persistence of a purposeful social space for exercise that could be accessed at any time. Therefore, the existence of the metaverse space is an aspect of environmental improvement for exercise for participants who previously had no exercise habits, leading to increased exercise implementation and physical activity.

The reason indicators other than physical activity did not improve in the metaverse group is also an important area to consider. This study asked about well-being as an item that could be related to mental health. Exercise is known to promote the secretion of estradiol and serotonin and regulate cortisol levels, an internal change that reduces depression and other depressive symptoms [42-44]. Additionally, reducing loneliness and encouraging social interaction decreases the risk of depression and mental health problems [45]. Therefore, we expected to observe mental health improvements in the metaverse group; however, there were no significant differences among the 3 groups. The standard for suspicion of mental problems is a cutoff value of ≤11 points for well-being [35,46]. However, the mean values for participant well-being did not correspond to the cutoff values. Thus, mental health changes were unlikely to be observed even if the amount of physical activity increased because the group did not initially have low well-being.

The locomotive function scale, a measure of body pain and locomotor function, showed an improving trend in the metaverse group; however, the difference was not significant. A previous study found pain relief after an 8-week exercise program, including strength training of the trunk, in participants with low back pain that appeared to be caused by inactivity [47]. Several studies have reported muscle hypertrophy occurring within 6-8 weeks [48]. These facts indicate that if training is managed meticulously in the right way, the body can be changed in a period of 8 weeks. However, it is possible that the unmonitored, 8-week intervention period did not result in sufficient changes in motor function, even though the amount of physical activity increased. To improve motor function, training to improve muscle strength and endurance should be continued at least 3 times per week [49].

As for the differences in impressions of the intervention between the metaverse and YouTube groups, many factors such as novelty, relatedness, excitement, delight, attractiveness, and expectation were more positive in the metaverse group than in the YouTube group. Regarding novelty, a previous study stated that sufficient novelty for the target audience leads to successful health support using the app [50]. Even for young adults who were relatively familiar with and used apps and social networking services on a regular basis, the metaverse seemed to be a novelty, which may have supported the implementation of the exercise in this study. Delight is an important factor in many areas of life [51]. Studies have reported that delight is the most important predictor of commitment and participation in general youth and elite sports [52,53]. This amplification of delight using a metaverse space to distribute videos is meaningful in promoting exercise. Interestingly, the metaverse group found the intervention more attractive. A previous study that used apps to promote healthy lifestyles, including regular exercise, found that a combination of different services, such as health-related quizzes and the ability to record health status, led to successful interventions [54]. Exercise video-delivered studies have also shown that simply delivering videos does not increase physical activity [8]. The essence of using the metaverse may have increased the attractiveness of exercise video distribution in this study.

This study examined the effectiveness of exercise video distribution using metaverse space by comparing the previously used YouTube and a control population. This study’s strength lies in making the first attempt to distribute exercise videos using metaverse space. Furthermore, it shows the potential of using metaverse space to improve the physical activity level of young adults.

However, this study has some limitations. First, selection bias may have occurred. We recruited participants who were not members of an athletic club and who did not exercise heavily. However, it is possible that the target population was composed of students who were interested in exercising, as suggested by their interest in participating in this study. Additionally, we recruited participants from a single university. The risk of selection bias may be reduced by conducting a study on a larger scale and involving more institutions. However, the 3 groups were randomly assigned in this study, and the required, precalculated sample size was secured. Second, it was difficult...
to accurately determine the time and frequency of interactions between participants in the metaverse space. If a system can be implemented to monitor the time of entry into the metaverse space, the frequency of interactions, and so on, the mechanism for promoting physical activity through the use of the metaverse space could be made more visible. Third, participants could receive the intervention at the time of their choosing; however, the start time of the intervention was not monitored. The effectiveness of the intervention could be moderated or altered by the time of the intervention, and future researchers should accurately monitor the start time of the intervention. We note that the main outcome, the calculation of physical activity, did not depend on the starting time of the intervention. Final, there was room for improvement in the comfort of the interventions in the metaverse group. This study used metaverse space in the manner in which the videos were arranged. Metaverse space (Spatial, Spatial Systems, Inc) has the advantage of being a service that anyone can use free of charge; however, if industry and academia collaborate to devise a system more suited to the distribution of exercise videos and the promotion of exercise, it could have a more positive effect on physical activity and other aspects.

In the near future, services allowing access to the metaverse will likely continue to expand. Therefore, the number of opportunities for the general public to experience the metaverse will further increase. The fact that 100% of the metaverse group in this study had high expectations is a hopeful part of the development of using the metaverse to improve physical activity. This research provides a basis for establishing further methods of using the metaverse to promote exercise.

Acknowledgments
We thank all the participants. We also thank Editage for the English language editing. This work was supported by the Japan Science and Technology Agency through the establishment of university fellowships toward the creation of science and technology innovation (JPMJFS2129).

Conflicts of Interest
None declared.

Multimedia Appendix 1
Comparison of impressions on intervention between the metaverse and YouTube groups. [PDF File (Adobe PDF File), 160 KB - mhealth_v12i1e46397_app1.pdf ]

Multimedia Appendix 2
CONSORT-eHealth (V 1.6.1). [PDF File (Adobe PDF File), 1312 KB - mhealth_v12i1e46397_app2.pdf ]

References


Mobile App Intervention of a Randomized Controlled Trial for Patients With Obesity and Those Who Are Overweight in General Practice: User Engagement Analysis Quantitative Study

Vera Helen Buss1*, PhD; Margo Barr1*, PhD; Sharon M Parker1*, MPH; Alamgir Kabir1*, PhD; Annie Y S Lau2*, PhD; Siaw-Teng Liaw3*, PhD; Nigel Stocks4*, MD; Mark F Harris1*, MD

1Centre for Primary Health Care and Equity, University of New South Wales, Sydney, Australia
2Australian Institute of Health Innovation, Macquarie University, Sydney, Australia
3School of Population Health, University of New South Wales, Sydney, Australia
4Adelaide Medical School, University of Adelaide, Adelaide, Australia

*all authors contributed equally

Corresponding Author:
Margo Barr, PhD
Centre for Primary Health Care and Equity
University of New South Wales
AGSM Building
High Street, Kensington Campus
Sydney, 2052
Australia
Phone: 61 290656041
Email: Margo.barr@unsw.edu.au

Abstract

Background: The Health eLiteracy for Prevention in General Practice trial is a primary health care–based behavior change intervention for weight loss in Australians who are overweight and those with obesity from lower socioeconomic areas. Individuals from these areas are known to have low levels of health literacy and are particularly at risk for chronic conditions, including diabetes and cardiovascular disease. The intervention comprised health check visits with a practice nurse, a purpose-built patient-facing mobile app (mysnapp), and a referral to telephone coaching.

Objective: This study aimed to assess mysnapp app use, its user profiles, the duration and frequency of use within the Health eLiteracy for Prevention in General Practice trial, its association with other intervention components, and its association with study outcomes (health literacy and diet) to determine whether they have significantly improved at 6 months.

Methods: In 2018, a total of 22 general practices from 2 Australian states were recruited and randomized by cluster to the intervention or usual care. Patients who met the main eligibility criteria (ie, BMI>28 in the previous 12 months and aged 40-74 years) were identified through the clinical software. The practice staff then provided the patients with details about this study. The intervention consisted of a health check with a practice nurse and a lifestyle app, a telephone coaching program, or both depending on the participants’ choice. Data were collected directly through the app and combined with data from the 6-week health check with the practice nurses, the telephone coaching, and the participants’ questionnaires at baseline and 6-month follow-up. The analyses comprised descriptive and inferential statistics.

Results: Of the 120 participants who received the intervention, 62 (52%) chose to use the app. The app and nonapp user groups did not differ significantly in demographics or prior recent hospital admissions. The median time between first and last app use was 52 (IQR 4-95) days, with a median of 5 (IQR 2-10) active days. App users were significantly more likely to attend the 6-week health check (2-sided Fisher exact test; P<.001) and participate in the telephone coaching (2-sided Fisher exact test; P=.007) than nonapp users. There was no association between app use and study outcomes shown to have significantly improved (health literacy and diet) at 6 months.

Conclusions: Recruitment and engagement were difficult for this study in disadvantaged populations with low health literacy. However, app users were more likely to attend the 6-week health check and participate in telephone coaching, suggesting that participants who opted for several intervention components felt more committed to this study.

https://mhealth.jmir.org/2024/1/e45942

JMIR Mhealth Uhealth 2024 | vol. 12 | e45942 | p.516
(page number not for citation purposes)
Considering other studies, the focus of this study was associated with successful weight loss in app users. Chin et al. [7] analyzed user data from a popular commercial weight loss app and found that in successful primary care patients who are particularly motivated to lose weight [6]. Patel et al. [5] reported that consistent weight self-monitoring during the study weeks [5]. Their study highlighted that consistent tracking was crucial, but only a minority of participants did so. Similarly, Laing et al. [6] found that providing access to a weight loss app to primary care patients who are overweight and those with obesity did not lead to significant weight loss compared to usual care. Only one-third of them logged into the app in the sixth month of the intervention, in which the median number of logins was 0 (IQR 0-2). The authors concluded that prescribing self-monitoring apps for caloric counting may be successful in primary care patients who are particularly motivated to lose weight [6].

### Rationale for the Study

Other research has shown that mobile app-based interventions can facilitate weight loss in individuals who are overweight and those with obesity, but it requires regular app use. For example, Patel et al. [5] reported that consistent weight self-monitoring via a mobile app could lead to clinically meaningful weight loss. However, the study classified only a quarter of participants as consistent trackers, which they defined as self-monitoring weight and diet on at least 6 days per week for at least 75% of the study weeks [5]. Their study highlighted that consistent tracking was crucial, but only a minority of participants did so. Similarly, Laing et al. [6] found that providing access to a weight loss app to primary care patients who are overweight and those with obesity did not lead to significant weight loss compared to usual care. Only one-third of them logged into the app in the sixth month of the intervention, in which the median number of logins was 0 (IQR 0-2). The authors concluded that prescribing self-monitoring apps for caloric counting may be successful in primary care patients who are particularly motivated to lose weight [6].

### Introduction

**Problem Statement**

Obesity is a major contributor to disease burden, increasing the risk of chronic conditions, including ischemic heart disease, stroke, diabetes mellitus, chronic kidney disease, and hypertensive heart disease [1]. According to the Global Burden of Disease Study 2017 [2], high BMI was the cause of 4.72 million deaths and 148 million disability-adjusted life-years worldwide, making it the fourth leading risk for mortality in 2017. In 2017 to 2018, an estimated 36% of the Australian adult population were overweight (ie, BMI 25.0-29.9) and 31% of them had obesity (ie, BMI ≥30.0) [3]. The proportion of people who are overweight or have obesity is higher in populations from lower socioeconomic backgrounds [3]. In 2017-2018, 72% of Australian adults residing in the lowest socioeconomic areas were overweight or had obesity compared to 62% from the highest, after adjusting for age [3]. People from the lowest socioeconomic areas were 1.9 times more likely to have diabetes in 2020 and 1.6 times more likely to have self-reported coronary heart disease in 2017-2018 than those from the highest socioeconomic areas [4].

### Rationale for the Study

Other research has shown that mobile app-based interventions can facilitate weight loss in individuals who are overweight and those with obesity, but it requires regular app use. For example, Patel et al. [5] reported that consistent weight self-monitoring during the study weeks [5]. Their study highlighted that consistent tracking was crucial, but only a minority of participants did so. Similarly, Laing et al. [6] found that providing access to a weight loss app to primary care patients who are overweight and those with obesity did not lead to significant weight loss compared to usual care. Only one-third of them logged into the app in the sixth month of the intervention, in which the median number of logins was 0 (IQR 0-2). The authors concluded that prescribing self-monitoring apps for caloric counting may be successful in primary care patients who are particularly motivated to lose weight [6].

### Description of the Intervention

The HeLP-GP trial was a behavior change intervention developed for implementation in Australian general practices aimed at Australians who are overweight and those with obesity from lower socioeconomic areas to help them reduce their weight. The intervention was based on the 5As framework (assess, advise, agree, assist, and arrange) [8]. It included health check visits with a practice nurse based on the 5As framework, the use of a purpose-built patient-facing mobile app called mysnapp, and referral to health coaching via the “Get Healthy” information and coaching service [9]. The mysnapp is based on a web-based platform developed by Lau et al. [10].

The trial was a pragmatic, 2-arm, unblinded cluster randomized controlled trial, which continued for 12 months. Primary outcomes included changes in weight, blood pressure, health literacy, and eHealth literacy [11,12]. Secondary outcomes included lipids, diet (fruit and vegetable intake), level of physical activity, quality of life, advice received, and referral for diet, physical activity, and weight loss [12]. Participants who received the intervention could choose to use the mobile app and access the telephone coaching program. The HeLP-GP trial assessed the intervention’s effectiveness [12]. The intervention led to significant improvements at 6 months compared to the controls for health literacy (mean DiD 0.22, 95% CI 0.01-0.44) and diet (mean DiD 0.98, 95% CI 0.50-1.47). There were no associations with any of the other outcomes [12].

### Objectives

The overall aim of this study, within the HeLP-GP trial, was to assess mysnapp app use, engagement, its association with other intervention components, and its association with study outcomes shown to have significantly improved (health literacy and diet) at 6 months.

Our objectives were to (1) explore differences in demographics and hospital admissions between participants who used the app and those who did not, (2) examine the duration and frequency of app use (app engagement) by participants overall and by module, (3) assess the association among app use, app engagement, and participation in other intervention components, and (4) examine the association between app use and app engagement on study outcomes that were shown to be significantly improved at 6 months (ie, health literacy and diet).
### Methods

#### Ethical Considerations

The University of New South Wales Human Research Ethics Committee (HC17474) approved the trial. The University of Adelaide Human Research Ethics Committee ratified this approval. All participants provided consent to take part in this study.

#### Intervention

The methodology of the randomized controlled trial, of which this study is a subanalysis, was published previously [13] and prospectively registered with the Australian New Zealand Clinical Trials Registry (ACTRN12617001508369). In 2018, a total of 22 general practices were recruited from 2 Australian states, New South Wales (South West and Central Sydney) and South Australia (Adelaide), and randomized by cluster to the HeLP-GP intervention (11 practices) or usual care (11 practices). General practices were recruited through the local Primary Health Networks. Practices were located in local government areas with Socio-Economic Indexes for Areas scores [14] equal to or below the eighth decile. The Australian Bureau of Statistics reported that these are usually associated with lower health literacy levels in the population [15], with health literacy being defined by the Australian Institute of Health and Welfare in their latest Health Literacy Report as “how people access, understand and use health information in ways that benefit their health” [16]. In total, 4 strata based on the practice size (<5 general practice [GPs] and ≥5 GPs) and ≥5 GPs) and the state were created and then we randomly allocated practices to each stratum’s intervention or usual care group. The intervention comprised a practice nurse-led health check; additionally, participants could choose whether to take up a lifestyle app, a telephone coaching program, or both. Potential participants were identified using the GPs’ software. The general practitioners of the intervention sites also assessed their patients for eligibility. The eligible patients were provided with trial information and consent forms by the reception staff. Recruitment occurred between October 2018 and September 2019.

At the baseline health check, the practice nurses helped participants with the mysnapp setup and access the coaching program. They entered the participant’s height, weight, waist circumference, and blood pressure into the app and set the health goals with the participant. For 6 weeks, the participants received a nutrition-related and a physical activity–related text message weekly. These were prepared to be sent automatically each week and provided direct advice and a web link for further information. In addition, the telephone coaching program provided free, confidential health support to participants to reach personalized lifestyle goals concerning diet, physical activity, alcohol, and body weight [17]. The coaching was available in multiple languages through an interpreter service. The practice nurses conducted a 6-week health check in which they reviewed and revised the participants’ health goals. Additionally, general practitioners conducted a 12-week health review. Text messages reminded participants to attend these follow-up visits.

#### Participants

Individuals were eligible for this study if they were aged 40–74 years, had a BMI of ≥28 and blood pressure levels recorded in the clinical software within the last 12 months, spoke English or Arabic, and had access to a smartphone or tablet. Potential participants were ineligible if they fulfilled any of the following exclusion criteria: recent weight loss (ie, >5% in the past 3 months), taking weight loss drugs (ie, orlistat or phentermine), diagnosed with insulin-dependent diabetes or cardiovascular disease (ie, angina, myocardial infarction, heart failure, heart valve disease, or stroke), cognitive impairment, or physical impairment disallowing them to perform moderate physical activity.

#### mysnapp Design

The mysnapp content was based on a web-based platform designed to help individuals control and maintain their health data and information to manage their health [10]. Research by Webb et al [18] and DiFilippo et al [19] into behavior change through mobile and electronic platforms informed the app design, including goal setting and self-monitoring, and additional methods to interact with individuals, mainly text messaging. The mysnapp app consisted of 4 core modules that allowed users to (1) set physical activity– and diet-based goals, (2) monitor their progress over the past 6 weeks, (3) take notes in a diary, and (4) learn about healthy eating and physical activity. Users could choose from the following goal options: set daily servings of fruits or vegetables or physical activity minutes; aim to drink fewer soft drinks, eat smaller portions, or eat fewer snacks or takeaway foods. In the self-monitoring module, they entered how many days of the week they achieved their goals. The educational material consisted of short text summaries and fact sheets about healthy foods, portion sizes, discretionary beverage consumption, physical activity benefits in English or Arabic, and links to simple exercise videos on YouTube.

#### Study Measures

##### App Use Measure

Data were collected on app use, specifically, when the study participants in the intervention group had an app account set up.

##### App Engagement Measures

Data were collected on the participants’ app use directly through mysnapp. Each month, a cumulative data report was created about app logins and interactions with the different app modules from each participant for 12 months. App engagement included active days, duration of app use, and frequency of accessing app modules.

##### Other Intervention Component Measures

The data from the 6-week health check with the practice nurses (ie, occurrence) and the telephone coaching (ie, occurrence and completion status) were the other intervention component measures.
Outcome Measures

The participants’ questionnaires at baseline and 6-month follow-up (i.e., self-reported fruit and vegetable intake, and health literacy) were used. Specifically, the diet questions were as follows: (1) How many servings of fruit do you usually eat each day? A serving is 1 medium-sized fruit such as an apple or 2 small-sized fruits or 1 cup of fruit pieces. (2) How many servings of vegetables do you usually eat each day? One serving is half a cup of cooked vegetables or 1 cup of salad vegetables. With the diet score being the portions of fruit intake (between 0 and a maximum of 2 per day) plus portions of vegetable intake (between 0 and a maximum of 5 per day) with a range of 0 to 7 based on the sum of fruit and vegetable scores. This diet measure has been validated against food frequency questionnaires [20].

Specifically, the Health Literacy Questionnaire domain 8 questions were the following [11]: please indicate how difficult or easy the following tasks are for you now: (1) find information about health problems; (2) find health information from several different places; (3) get information about health so you are up to date with the best information; (4) get health information in words you understand; and (5) get health information by yourself. There is a 5-point response option scale for each question (cannot do or always difficult, usually difficult, sometimes difficult, usually easy or always easy). The scores are reported as averages for the domain (with a range between 1 and 5) with high scores representing higher health literacy.

Table 1 contains definitions for study measures. Duration of app use, active days, and consistent use had preset maximum values (365 days or 52 weeks); the values were capped when they exceeded the maximum.

### Table 1. Measures and their definitions.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Type of variable</th>
<th>Explanation or definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>App use measures</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>App user</td>
<td>Binary and input variable</td>
<td>Study participants in the intervention group who had an app account set up</td>
</tr>
<tr>
<td><strong>App engagement measures</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration of app use</td>
<td>Continuous and input variable; maximum value: 365 days</td>
<td>Number of days between the first and last time a participant accessed the app</td>
</tr>
<tr>
<td>Active days</td>
<td>Continuous and input variable; maximum value: 365 days</td>
<td>Number of days a participant accessed the app</td>
</tr>
<tr>
<td>Consistent app use</td>
<td>Continuous and input variable; maximum value: 52 weeks</td>
<td>Number of consecutive weeks a participant accessed ≥1 time the app starting from the first app use</td>
</tr>
<tr>
<td>App module use</td>
<td>Binary and input variable</td>
<td>Participant accessed ≥1 the corresponding app module (goal setting, progress tracking, diary, or education)</td>
</tr>
<tr>
<td>Frequency of accessing app modules</td>
<td>Continuous and input variable</td>
<td>Number of times a participant accessed the corresponding app module (goal setting, progress tracking, diary, or education)</td>
</tr>
<tr>
<td><strong>Other intervention component measures</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Practice nurse-led health check</td>
<td>Categorical and input variable</td>
<td>Attended and not attended</td>
</tr>
<tr>
<td>Telephone coaching</td>
<td>Categorical and input variable</td>
<td>Completed, not completed, and not participated</td>
</tr>
<tr>
<td><strong>Outcome measures</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health literacy</td>
<td>Continuous and output variable</td>
<td>Health literacy, specifically the self-reported ability to find good quality health information, according to domain 8 of the Health Literacy Questionnaire [9], at baseline and 6-month follow-up. The scores were reported as averages for the domain (with a range between 1 and 5) with high scores representing higher health literacy.</td>
</tr>
<tr>
<td>Diet score</td>
<td>Continuous and output variable</td>
<td>Self-reported daily fruit and vegetable intake at baseline and 6-month follow-up. Diet score was the portions of fruit intake (between 0 and a maximum of 2 per day) plus portions of vegetable intake (between 0 and a maximum of 5 per day) with a range between 0 and 7 based on the sum of fruit and vegetable scores.</td>
</tr>
</tbody>
</table>

Data Analysis

Descriptive and inferential analyses in RStudio (with the programming language R; R Foundation for Statistical Computing) using a significance level of .05 for all statistical tests were conducted. Normally distributed continuous variables were summarized using the mean and SD, and nonnormally distributed continuous variables with median and IQR. Box plots compared continuous variables across the categories of nonnumerical variables [21]. Normality was tested using the Shapiro-Wilk normality test [22-24]. The 2-sided Welch t test was performed to compare the means of continuous variables.
between 2 subgroups (eg, participants using mysnapp versus those not using it) for normally distributed continuous variables [25]. Alternatively, the Wilcoxon signed rank test with continuity correction comparing the medians of nonnormally distributed continuous variables between 2 subgroups was used [26,27]. The Kruskal-Wallis rank-sum test was performed for more than 2 subgroups and nonnormally distributed continuous variables [28]. Pearson chi-square test with Yates continuity correction was used to test for associations between 2 categorical variables and the 2-sided Fisher exact test was used when there were less than 5 participants in any cell of the contingency table of expected frequencies [29-31].

For objective 4, we used 1-sided tests to assess whether app use versus nonapp use, or app engagement was associated with health literacy or diet between baseline and 6-month follow-up. Correlations between the app engagement and health literacy or diet score were measured with the Kendall rank correlation test (if variables did not follow a normal distribution) or Pearson product-moment correlation test (if they followed a normal distribution) [32,33].

Results

App Users

In total, 120 participants received the intervention, of which 62 (52%) people chose to use mysnapp. Among the 62 app users, 38 (61%) also opted for telephone coaching. Table 2 shows the results for the first objective, comparing the demographic characteristics of the participants who chose not to use mysnapp to those who decided to use it. There were no significant differences between app users and nonapp users.

Table 2. Demographic characteristics of participants in the intervention group (N=120).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Nonapp users (n=58)</th>
<th>App users (n=62)</th>
<th>Test statistics for differences between groups</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>58 (8)</td>
<td>61 (9)</td>
<td>$t_{115}=-1.56$</td>
<td>.12</td>
</tr>
<tr>
<td>Women, n (%)</td>
<td>28 (48)</td>
<td>32 (52)</td>
<td>$\chi^2 =0.1$</td>
<td>.86</td>
</tr>
<tr>
<td>Born in Australia, n (%)</td>
<td>27 (47)</td>
<td>39 (63)</td>
<td>$\chi^2 =2.6$</td>
<td>.11</td>
</tr>
<tr>
<td>Preferred language is English, n (%)</td>
<td>54 (93)</td>
<td>58 (94)</td>
<td>OR$^a$ 0.93, 95% CI 0.16-5.26</td>
<td>&gt;.99</td>
</tr>
<tr>
<td>Hospital admission in past 12 months, n (%)</td>
<td>15 (26)</td>
<td>12 (19)</td>
<td>$\chi^2 =0.4$</td>
<td>.53</td>
</tr>
<tr>
<td>Location New South Wales, n (%)</td>
<td>50 (86)</td>
<td>49 (79)</td>
<td>$\chi^2 =0.6$</td>
<td>.43</td>
</tr>
</tbody>
</table>

$^a$OR: odds ratio.

App Engagement

The median duration of app use was 52 (IQR 4-95) days. Further, 2 participants used mysnapp weekly throughout the 12 months (Table 3). Active days ranged from 1 to 117 days, with a median of 5 (IQR 2-10) days. The median number of weeks participants consistently used mysnapp from baseline was 1 (IQR 1-2). Of the 62 app users, 60 (97%) opened the goal setting module, 55 (89%) the education module, 39 (63%) the progress tracking module, and 25 (39%) the diary. Table 3 shows the consistency of app use and how many modules the app users accessed over the entire period of the intervention. Of the 19 app users who had opened 3 of the 4 modules, 17 (89%) had accessed the goal setting, progress tracking, and education modules. Among the 16 who had opened 2 modules, 14 (88%) had accessed the goal setting and education modules.
Table 3. Consistency of app use and frequency of accessing app modules (n=62).

<table>
<thead>
<tr>
<th>Variables and values</th>
<th>Participants, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistent app use (weeks)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>45 (73)</td>
</tr>
<tr>
<td>2-4</td>
<td>10 (16)</td>
</tr>
<tr>
<td>5-19</td>
<td>5 (8)</td>
</tr>
<tr>
<td>20-52</td>
<td>2 (3)</td>
</tr>
<tr>
<td>Number of modules accessed</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1 (2)</td>
</tr>
<tr>
<td>1</td>
<td>5 (8)</td>
</tr>
<tr>
<td>2</td>
<td>16 (26)</td>
</tr>
<tr>
<td>3</td>
<td>19 (31)</td>
</tr>
<tr>
<td>4</td>
<td>21 (34)</td>
</tr>
<tr>
<td>Frequency of accessing the goal setting module</td>
<td></td>
</tr>
<tr>
<td>0-3</td>
<td>54 (87)</td>
</tr>
<tr>
<td>4-7</td>
<td>6 (10)</td>
</tr>
<tr>
<td>8-15</td>
<td>2 (3)</td>
</tr>
<tr>
<td>&gt;15</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Frequency of accessing the progress tracking module</td>
<td></td>
</tr>
<tr>
<td>0-3</td>
<td>43 (69)</td>
</tr>
<tr>
<td>4-7</td>
<td>11 (18)</td>
</tr>
<tr>
<td>8-15</td>
<td>6 (10)</td>
</tr>
<tr>
<td>&gt;15</td>
<td>2 (3)</td>
</tr>
<tr>
<td>Frequency of accessing the diary module</td>
<td></td>
</tr>
<tr>
<td>0-3</td>
<td>48 (77)</td>
</tr>
<tr>
<td>4-7</td>
<td>5 (8)</td>
</tr>
<tr>
<td>8-15</td>
<td>4 (6)</td>
</tr>
<tr>
<td>&gt;15</td>
<td>5 (8)</td>
</tr>
<tr>
<td>Frequency of accessing the education module</td>
<td></td>
</tr>
<tr>
<td>0-3</td>
<td>41 (66)</td>
</tr>
<tr>
<td>4-7</td>
<td>12 (19)</td>
</tr>
<tr>
<td>8-15</td>
<td>6 (10)</td>
</tr>
<tr>
<td>&gt;15</td>
<td>3 (5)</td>
</tr>
</tbody>
</table>

Association With Other Intervention Components
The difference in telephone coaching uptake between the app and nonapp users was statistically significant (Freeman-Halton extension of 2-sided Fisher exact test \( P<.001 \), Table 4). The median number of days using mysnapp for the app users who completed the telephone coaching was 3.5 (IQR 1-7) days, for the app users who did not complete the telephone coaching it was 7 (IQR 2.5-9.5) days, and for the app users who did not undertake the telephone coaching it was 3.5 (IQR 2-9) days (Figure 1). The difference in median active days by telephone coaching completion status was not statistically significant \( \chi^2_{19}=13.2, P=.83 \).
Table 4. Association of app use with other intervention components (N=120).

<table>
<thead>
<tr>
<th>Other intervention components and status</th>
<th>Nonapp users (n=58), n (%)</th>
<th>App users (n=62), n (%)</th>
<th>Test for differences between groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telephone coaching program</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not participated</td>
<td>47 (81)</td>
<td>24 (39)</td>
<td>Freeman-Halton extension of Fisher exact test (2-tailed) P&lt;.001</td>
</tr>
<tr>
<td>Not completed</td>
<td>8 (14)</td>
<td>16 (26)</td>
<td></td>
</tr>
<tr>
<td>Completed</td>
<td>3 (5)</td>
<td>22 (35)</td>
<td></td>
</tr>
<tr>
<td>A 6-week health check</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not attended</td>
<td>54 (93)</td>
<td>46 (74)</td>
<td>Fisher exact test (2-tailed) P=.007</td>
</tr>
<tr>
<td>Attended</td>
<td>4 (7)</td>
<td>16 (26)</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Box plots of the number of days actively using mysnapp depending on the participation in telephone coaching; outliers excluded (two for no telephone coaching: 30 and 105 days, one for not completed: 117 days, and one for completed: 105 days).

The difference in the attendance rate of the 6-week health check between app users and nonusers was significant (2-sided Fisher exact test P=.007, Table 4). Those app users who attended the 6-week health check with the practice nurse did not have significantly more active days using mysnapp (median active days for 6-week health check attendees: 6, IQR 2-10 days, and for nonattendees: 4, IQR 2-10 days; W=374, P=.46).

Impact of App Use and App Engagement on Behavioral and Biomedical Outcome Measures

Differences in outcome measures between app users and nonusers, and app engagement were not significant (Tables 5 and 6) for study outcomes which were shown to be significantly improved at 6 months (ie, health literacy and diet).
Table 5. Health literacy and diet score at 2 time points for app and nonapp users, test for significant changes, and sensitivity analysis.

<table>
<thead>
<tr>
<th>Outcome variable and measure</th>
<th>App users (n=62)</th>
<th>Nonapp users (n=58)</th>
<th>Test statistica</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline 6 months</td>
<td>Baseline 6 months</td>
<td></td>
</tr>
<tr>
<td><strong>HLQb domain 8</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data available, n (%)</td>
<td>52 (84)</td>
<td>44 (76)</td>
<td>N/Ac</td>
</tr>
<tr>
<td>Median (IQR)</td>
<td>4 (4-5)</td>
<td>4 (4-5)</td>
<td>W=230.5, P=.10</td>
</tr>
<tr>
<td><strong>Diet score</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data available, n (%)</td>
<td>57 (92)</td>
<td>54 (93)</td>
<td>N/A</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>3 (2-5)</td>
<td>4 (4-5)</td>
<td>t36=0.32, P=.37</td>
</tr>
</tbody>
</table>

aTest for greater change in app users versus nonapp users from baseline to 6 months.
bHLQ: Health Literacy Questionnaire.
cN/A: not applicable.

Table 6. Correlation between app engagement and change in health literacy or diet score.

<table>
<thead>
<tr>
<th>Outcome variable and measure for app use</th>
<th>Test statistics for differences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HLQb domain 8</strong></td>
<td></td>
</tr>
<tr>
<td>Active days</td>
<td>z=-0.24, P=.81, τ=-0.03</td>
</tr>
<tr>
<td>Consistent app use</td>
<td>z=0.43, P=.67, τ=0.06</td>
</tr>
<tr>
<td><strong>Diet score</strong></td>
<td></td>
</tr>
<tr>
<td>Active days</td>
<td>z=0.55, P=.58, τ=0.07</td>
</tr>
<tr>
<td>Consistent app use</td>
<td>z=0.43, P=.67, τ=0.06</td>
</tr>
</tbody>
</table>

aHLQ: Health Literacy Questionnaire.

Discussion

Principal Results

The overall aim of this study was to assess mysnapp app use within the HeLP-GP trial and its association with study outcomes shown to have significantly improved (health literacy and diet) at 6 months. With regard to the specific objectives, (1) there were no significant differences in demographics between participants who used mysnapp and those who did not; (2) among app users, the median duration of app use was 52 days, with a median of 5 active days; (3) more participants who chose to use mysnapp also attended the 6-week health check with the practice nurse and opted for telephone coaching; and (4) there was no association between app use and study outcomes shown to have significantly improved (health literacy and diet) at 6 months.

Length and Frequency of App Use and Module Access

Turner-McGrievy et al [34] aimed to identify the best criteria for defining adherence to dietary self-monitoring with mobile devices when predicting weight loss. They found that adherence, defined as the number of days participants tracked at least 2 meal times, explained the most variance in weight loss at 6 months [34]. We were not able to measure this because the diary, available for recording meals, could also be used for other reasons such as activities, appointments, plans for the future, and thoughts about progress. In the study by Jacobs et al [35], they analyzed data from 7680 users of a commercial weight loss app; high adherence to self-monitoring (ie, logging at least 1 food event within a reasonable time after a meal) was associated with increased weight loss. However, they also found that app users with higher adherence rates had significantly lower body weight at baseline than those with lower adherence rates [35]. The analysis only comprised people who entered data in the app at least once a week for 12 weeks. In our study, 4.9% (n=3) of the app users were still entering data at week 12. Analyzing data from the same commercial app, Carey et al [36] found significant differences in 7 different engagement measures (ie, number of articles read, meals logged, steps recorded, messages to coach, exercise logged, weigh-ins, and days with 1 meal logged per week) between app users with moderate or high weight loss (ie, 5%-10% or >10% body weight loss, respectively) and individuals with no change in body weight (ie, ±1% body weight). Their analysis indicated that people with moderate to high weight loss engaged with all app sections [36]. In our study, only 34% (n=21) of the app users had accessed all of the modules.

Impact of App Use and App Engagement on Behavioral and Biomedical Outcome Measures

Other studies showed promising results for weight loss apps, for example, Carter et al [37], Patel et al [5], and Antoun et al [38]. Specifically, Carter et al [37] conducted a pilot study of 128 volunteers who are overweight comparing a smartphone app (My Meal Mate) with a website and paper diary. They found the mean weight loss over 6 months for the app was higher (4.6 kg, 95% CI 3.0-6.2) than for the diary group (2.9 kg, 95% CI
1.1-4.7) or the website group (1.3 kg, 95% CI 0.1-2.7). Antoun et al [38] in their review of 34 studies that evaluated the use of smartphones for weight loss found an overall mean loss of 2.8 kg (95% CI 2.6-3.0) at 6 months. Patel et al [5] found that consistent tracking was associated with greater weight loss than inconsistent tracking at 6 months (2.1 kg, 95% CI 0.3-4.0). A difference between these studies and ours was that they did not specifically target disadvantaged populations with low health literacy. Therefore, their apps were more complex than ours. In contrast, Lanpher et al [39] developed a weight loss intervention suitable for individuals with low health literacy. A computer algorithm automatically allocated the self-monitoring goals (eg, no sugary drinks, no snacking after dinner, eating 5 fruits and vegetables a week). Participants reported whether they achieved the goals via interactive voice response calls [39]. The algorithm decided which goals to assign next based on previous adherence to goals so that individuals would rather receive goals to which they were receptive [40]. They also received tailored skills training through verbal calls and materials, one-on-one counseling calls, and a membership at the gym [39]. The results showed that the intervention group maintained or lost weight over 12 months, independent of their level of health literacy [38].

Bennett et al [40] extended the intervention to comprise a mobile app. They evaluated its effectiveness in a randomized controlled trial including socioeconomically disadvantaged patients with increased cardiovascular risk by comparing the intervention to usual care [40]. The app used interactive voice responses or text messaging to simplify self-monitoring, like in the previous study. Additionally, participants received in-person coaching and personalized feedback messages immediately after entering data [40]. The intervention group achieved meaningful weight loss, with more than 40% of participants reducing their body weight by at least 5% compared to 17% of participants in the usual care group [39]. Comparing this intervention to ours raises the question of whether the way people had to select and track their goals in our app contributed to the low engagement and the nonsignificant findings. Locke and Latham [41] explained that goal commitment, goal importance, self-efficacy, feedback, and task complexity act as moderators between goals and performance. Potentially, the app did not sufficiently address all 5 moderators.

Association With Other Intervention Components

This study showed that mysnapp users were more likely to attend the 6-week face-to-face health check with the practice nurse and to participate in the telephone coaching program than nonusers. Potentially, these individuals were more motivated to lose weight and, therefore, more willing to engage in the other intervention components. Another explanation could be that participants who opted for several intervention components felt more committed to study participation and, therefore, made more use of the individual intervention components. Griausde et al [42] proposed a similar hypothesis in their mobile health–based prediabetes intervention study; they assumed that participants who received a more robust intervention were more committed to the study and subsequently more likely to complete the 12-week survey. Hutchesson et al [43] concluded that adding nondigital components, such as face-to-face visits and telephone coaching, to mobile health interventions can improve participants’ accountability even though these additional features may not be necessary for the intervention’s effectiveness.

Limitations

The plan for the randomized controlled trial was to recruit 800 study participants; however, only 215 individuals were able to be recruited (120 in the intervention and 95 in the control group) [13]. Further, despite targeting low socioeconomic areas, this study failed to recruit many participants with low health literacy. One needs to be cautious when interpreting the results of this study due to the small sample size and the high dropout. Despite considerable efforts and additional time to recruit participating practices and patients, the anticipated sample size was not achieved. Research by Perkins et al [44] has shown an ongoing issue with recruitment through Australian general practices. Another problem with the study was that the uptake of intervention components was determined by the clinician and patient. Thus, some chose to just have the app and others to just have the phone coaching. Additionally, the study may not be generalizable to other settings. Since recruitment was from 2 Australian urban areas, results could differ in rural areas or other urban areas. Diet score and health literacy level were self-reported, posing a risk of bias. Further, caution is required when interpreting the results in the context of low health literacy because the baseline health literacy levels were higher than anticipated [12]. According to data from the National Health Survey 2018, the health literacy level in this study’s sample was comparable to that of Australians who are overweight or those with obesity in the general population [45]. A potential explanation is that this study’s requirements (randomization, completing the questionnaire, and undertaking the health check) stopped people with low health literacy from participating. This rationale is in line with results from Krippalani et al [46], who found that people with low health literacy or numeracy were significantly less interested in participating in research.

Conclusions

There was no association between app use and study outcomes shown to have significantly improved (health literacy and diet) at 6 months. Recruitment and engagement were difficult for this study in disadvantaged populations with low health literacy. A potential explanation could be related to the self-selection of the goals and the weekly submission of the goal achievements. The practice nurses assisted participants at the beginning with the selection of goals. However, these may not have been relevant to participants, and nurses did not receive specific training in selecting meaningful goals for individuals.

However, app users were more likely to attend the 6-week health check and participate in telephone coaching, suggesting that participants who opted for several intervention components felt more committed to this study.
Acknowledgments
The authors would like to acknowledge the contribution to this research by An Tran, Carmel McNamara, Elizabeth Denney-Wilson, Katrina Paine, and Shoko Saito. We also acknowledge Louise Thomas for contributing to the trial protocol and early development. We are grateful for the partnership and support of the South Western Sydney, Adelaide and Nepean and Blue Mountains Primary Health Networks and the Australian Institute of Health Innovation. We would like to acknowledge the general practices and their staff and patients for participating in the research and consumers affiliated with Adelaide PHN for piloting mysnapp. This work was supported by the National Health and Medical Research Council of Australia (grant APP1125681, 2017).

Conflicts of Interest
None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1102 KB - mhealth_v12i1e45942_app1.pdf]

References
38. Antoun J, Itani H, Alarab N, Elsehmawy A. The effectiveness of combining nonmobile interventions with the use of smartphone apps with various features for weight loss: systematic review and meta-analysis. JMIR mHealth uHealth 2022;10(4):e35479 [FREE Full text] [doi: 10.2196/35479] [Medline: 35394443]


Abbreviations

GP: general practice
HeLP-GP: Health eLiteracy for Prevention in General Practice
The Association of Macronutrient Consumption and BMI to Exhaled Carbon Dioxide in Lumen Users: Retrospective Real-World Study

Shlomo Yeshurun1, PhD; Tomer Cramer1, PhD; Daniel Souroujon1,2, MD, MPH; Merav Mor1, PhD

1Metaflow Ltd, Tel-Aviv, Israel
2School of Public Health, Tel Aviv University, Tel-Aviv, Israel

Corresponding Author:
Shlomo Yeshurun, PhD
Metaflow Ltd
HaArba’a St 30
Tel-Aviv, 6473926
Israel
Phone: 972 37684062
Email: shlomoyesh@gmail.com

Abstract

Background: Metabolic flexibility is the ability of the body to rapidly switch between fuel sources based on their accessibility and metabolic requirements. High metabolic flexibility is associated with improved health outcomes and a reduced risk of several metabolic disorders. Metabolic flexibility can be improved through lifestyle changes, such as increasing physical activity and eating a balanced macronutrient diet. Lumen is a small handheld device that measures metabolic fuel usage through exhaled carbon dioxide (CO2), which allows individuals to monitor their metabolic flexibility and make lifestyle changes to enhance it.

Objective: This retrospective study aims to examine the postprandial CO2 response to meals logged by Lumen users and its relationship with macronutrient intake and BMI.

Methods: We analyzed deidentified data from 2607 Lumen users who logged their meals and measured their exhaled CO2 before and after those meals between May 1, 2023, and October 18, 2023. A linear mixed model was fitted to test the association between macronutrient consumption, BMI, age, and gender to the postprandial CO2 response, followed by a 2-way ANOVA.

Results: The model demonstrated significant associations ($P<.001$) between CO2 response after meals and both BMI and carbohydrate intake (BMI: $\beta=-0.112$, 95% CI $-0.156$ to $-0.069$; carbohydrates: $\beta=0.046$, 95% CI 0.034-0.058). In addition, a 2-way ANOVA revealed that higher carbohydrate intake resulted in a higher CO2 response compared to low carbohydrate intake ($F_{2,2569}=24.23; P<.001$), and users with high BMI showed modest responses to meals compared with low BMI ($F_{2,2569}=5.88; P=.003$).

Conclusions: In this study, we show that Lumen’s CO2 response is influenced both by macronutrient consumption and BMI. The results of this study highlight a distinct pattern of reduced metabolic flexibility in users with obesity, indicating the value of Lumen for assessing postprandial metabolic flexibility.

(JMIR Mhealth Uhealth 2024;12:e56083) doi:10.2196/56083

KEYWORDS
app; applications; association; BMI; body mass index; carbohydrate; carbon dioxide; consumption; correlate; correlation; diet; dietary; exhalation; exhale; food; Lumen; macronutrient; meal; metabolic flexibility; metabolic; metabolism; mHealth; mobile health; nutrient; nutrition; nutritional; obese; obesity; postprandial; prandial; retrospective; weight
Introduction

Background

The presence of obesity is the leading risk factor for metabolic disorders such as type 2 diabetes (T2D) and cardiovascular diseases (CVDs), and it has been linked to a reduced life expectancy [1,2]. Metabolic syndrome represents a collection of metabolic abnormalities that includes obesity as well as insulin resistance, hypertension, and dyslipidemia [3]. Changing the macronutrient distribution, such as low-fat or low-carbohydrate diets, has been proposed for the treatment and management of metabolic syndrome [4], and they have been shown to improve several clinical features of metabolic syndrome, including weight loss and cardiovascular risk [5,6].

Metabolic flexibility is the ability to switch between fuel sources (such as carbohydrates and fats) in response to their availability and metabolic demands, which is crucial for maintaining overall health and well-being [7,8]. Lifestyle modifications, such as exercise and a balanced diet, can therefore enhance metabolic flexibility, and high metabolic flexibility is associated with improved health outcomes, including a reduced risk of metabolic syndrome [7]. In addition, it was found to be impaired in individuals with obesity and T2D compared to lean and healthy individuals [9,10]. Assessment of metabolic flexibility can be conducted through postprandial changes (following a meal or any other insulin stimulation) in the respiratory exchange ratio (RER) from the metabolic cart [11,12], which estimates the body’s preference for macronutrient oxidation. Accordingly, several studies showed different RER responses between participants with high metabolic flexibility compared with participants with low flexibility for fasting, a high-carbohydrate diet, and a high-fat diet [13]. A recent study has shown that RER is greater after high-carbohydrate overfeeding than high-fat overfeeding, and impaired metabolic flexibility, as measured by ΔRER, is associated with greater weight gain over the following 6- and 12-month periods [14]. Furthermore, ΔRER is significantly elevated after high-carbohydrate intake compared with high-fat intake, indicating that metabolic flexibility is influenced by macronutrient composition [15]. These studies show that metabolic flexibility can be assessed based on RER from the metabolic cart.

However, metabolic carts are expensive, time-consuming, and only available in health care laboratories [16]. Thus, a small handheld device was developed to measure metabolic fuel usage by analyzing exhaled carbon dioxide (CO₂). The Lumen device is a portable breath-analyzer that measures metabolic fuel use through exhaled CO₂ and was found to be in agreement with the RER measured by the metabolic cart [17]. Furthermore, Lumen was found to be able to detect different metabolic responses to low- or high-carb lifestyles through CO₂ changes [18].

A variety of mobile health apps, including Lumen, which are designed for mobility and ubiquity, have demonstrated promising results in enhancing metabolism and facilitating weight management [19]. Using real-time data tracking, these tools empower individuals to make more informed health decisions [20]. As such, mobile apps such as Lumen have the potential to make a positive difference in people’s health and well-being [21].

Objective

This study examined data collected from Lumen users’ exhaled CO₂ measurements taken before and after logging a meal. The objective of this analysis was to investigate the postprandial response of Lumen’s CO₂ measurements to meals logged by Lumen users and to understand the association of this response with macronutrient consumption as well as users’ age, gender, and BMI.

Methods

Ethical Considerations

This study was determined to be exempt from institutional board review (IRB) under category 2, as detailed in 45 CFR 46.104(d) and the standard operating procedure of the Biomedical Research Alliance of New York (BRANY), by the BRANY Social, Behavioral, and Educational Research (SBER) IRB on May 9, 2023 (BRANY IRB File 23-119-1476).

Study Design

This is a retrospective observational study based on deidentified data collected from users of the Lumen device and app.

Participants

Participants’ data were collected retrospectively between May 1, 2023, and October 18, 2023. All users in the analysis were aged 18 years or older. Since only 12 users who were underweight (BMI<18.5) were found in the database, they were removed from the analysis.

Data Sources

As part of the onboarding process, users were required to specify their gender, date of birth, height, and weight. The use of the Lumen app includes an optional morning fasted measurement with the Lumen device, as well as recommended pre- and postmeal measurements throughout the day and a bedtime measurement before sleeping. The app provides nutritional recommendations based on the user’s personal preferences as well as their morning CO₂ measurement regarding the amount of macronutrients one should consume during the day. The app also allows users to log their meals—whether they are breakfast, lunch, dinner, or snacks. The macronutrient composition of these meals is determined by the nutritional database of Nutritionix [22], which includes all nutritional data available from the United States Department of Agriculture’s (USDA) Food Composition Database, restaurant chain data, and foods added by Nutritionix’s dietitians. An example of the food log feature in the app is shown in Figure 1.

Exhaled CO₂ measurements were obtained using the Lumen device (Metaflow Ltd). The Lumen mobile app guides the user through each phase of the Lumen maneuver (inhale, breath hold, and exhale), as previously described [17].

In this study, premeal and fasting measurements were defined as “premeal %CO₂,” while postmeal and bedtime measurements were defined as “postmeal %CO₂.”
were considered to be the “postmeal %CO$_2$” measure, with a maximum of 210 minutes between pre- and postmeal measurements selected for the analysis. The relative change in %CO$_2$ from premeal to postmeal was calculated for the final analysis. In addition, meals that were tagged as either breakfast, lunch, or dinner were selected for the analysis, while snacks were removed. Further data were excluded from the analysis using the Tukey method for outlier removal with an IQR of 1.5 for all macronutrients.

**Figure 1.** Lumen mobile app screenshots detailing the logged meals and their macronutrient distribution.

**Statistical Analysis**

Data were analyzed using Python JupyterLab (version 3.6.3; Project Jupyter), and all statistical analyses were conducted with the Python programming language, using custom scripts and the *statsmodels* package [23]. Figures were made with GraphPad Prism (version 10.1.0 for Windows, GraphPad Software) [24].

A linear mixed model (LMM) was fitted in order to test the relationship between the quantity of macronutrients (carbohydrates, fats, and proteins) and personal information (gender, age, and BMI) to the outcome variable of percentage of change in %CO$_2$ from premeal to postmeal, where users’ unique ID was adjusted for as a random effect [25]. A 2-way ANOVA was used to test the differences between different groups of statistically significant variables in the LMM. Among all the analyses performed, a 2-sided $P \leq 0.05$ was considered statistically significant.

**Results**

**Participants**

Overall, a total of 2607 users who completed 6671 coupled pre- and postmeal sessions from 6207 logged meals were used in the final analysis, with most users contributing only 1 session into the analysis (median 1, IQR 1-2 sessions per user).

**Descriptive**

A total of 81.97% (2137/2607) of the participants were women, and men and women did not differ in their ages or BMIs. A summary of the characteristics of the participants in the study is presented in Table 1.

Additionally, the macronutrient composition of the meals consumed by the participants was primarily carbohydrates, followed by proteins and fats, as detailed in Table 2.
Table 1. Sample characteristics of users (N=2607).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (female), n (%)</td>
<td>2137 (81.97)</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>47.6 (9.4)</td>
</tr>
<tr>
<td>BMI (kg/m²), mean (SD)</td>
<td>28.8 (6.0)</td>
</tr>
</tbody>
</table>

Table 2. Macronutrient composition of meals (N=6207).

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbohydrates (grams)</td>
<td>32.5 (20.5)</td>
</tr>
<tr>
<td>Proteins (grams)</td>
<td>29.7 (14.7)</td>
</tr>
<tr>
<td>Fats (grams)</td>
<td>22.5 (13.7)</td>
</tr>
</tbody>
</table>

Evaluation Outcomes

The overall CO₂ response after meal consumption at various time points is shown in Figure 2. As expected, %CO₂ measured by Lumen 30 to 210 minutes following meal intake increased significantly compared to premeal measurement.

The relationship between postprandial CO₂ response and macronutrient intake, as recorded in the Lumen app and users’ personal information, was examined using Linear Mixed Modeling (LMM). The users’ unique ID was incorporated as a random effect in the analysis. The model revealed a statistically significant relationship between the CO₂ response and the BMI of the users (P<.001), as well as their carbohydrate intake (P<.001). Lower BMI predicted a more significant increase in postprandial CO₂ response, and increased carbohydrate intake predicted a greater postprandial CO₂ response (Table 3). In contrast, their postprandial CO₂ response was not significantly influenced by their age, gender, fat intake, or protein intake (all P>.05; Table 3).

As the LMM analysis demonstrated significant associations between BMI, carbohydrate intake, and postprandial CO₂ response, these parameters were then divided into 3 categories each. BMI was classified as healthy (18.5-25 kg/m²), overweight (25-30 kg/m²), and obese (≥30 kg/m²), while carbohydrate intake was classified as low (0-30 grams), medium (30-60 grams), and high (≥60 grams).

To examine how the CO₂ response differed between each BMI category for each carbohydrate intake class, we conducted a 2-way ANOVA. In this analysis, only the last session of each user was used to eliminate the effect within the same user’s sessions. In agreement with the results from the LMM, this analysis revealed a significant effect of users’ BMI category (F_{2,2569}=5.88; P=.003) and level of carbohydrate intake (F_{2,2569}=24.23; P<.001), where users with obesity tended to have a lower postprandial CO₂ response than those with healthy BMI, and high consumption of carbohydrates resulted in a higher CO₂ response (Figure 3). Nevertheless, users with high BMI had similar elevated CO₂ response as overweight and healthy users when consuming meals high in carbohydrates.

Figure 2. Exhaled carbon dioxide (CO₂) response to meal at different time intervals (N=6671). Results represent mean (95% CI).
Table 3. Determinants of CO$_2$ response to food intake (N=6671).

<table>
<thead>
<tr>
<th>Variables</th>
<th>$\beta$ (95% CI)</th>
<th>z statistics</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>5.221 (3.242 to 7.199)</td>
<td>5.172</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>BMI</td>
<td>−0.112 (−0.156 to −0.069)</td>
<td>−5.026</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Age</td>
<td>−0.021 (−0.049 to 0.007)</td>
<td>−1.483</td>
<td>.14</td>
</tr>
<tr>
<td>Gender</td>
<td>−0.612 (−1.324 to 0.100)</td>
<td>−1.683</td>
<td>.09</td>
</tr>
<tr>
<td>Carbohydrates</td>
<td>0.046 (0.034 to 0.058)</td>
<td>7.496</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Fat</td>
<td>0.018 (−0.001 to 0.038)</td>
<td>1.816</td>
<td>.07</td>
</tr>
<tr>
<td>Protein</td>
<td>0.016 (−0.002 to 0.035)</td>
<td>1.755</td>
<td>.08</td>
</tr>
</tbody>
</table>

Figure 3. Carbon dioxide (CO$_2$) response to different carbohydrates intake in different BMI categories (N=2607). Results represent mean (95% CI).

Discussion

Principal Findings

This study showed how the Lumen device can detect the %CO$_2$ response to meals with a different macronutrient composition (mixed meals). Notably, this response is influenced by the carbohydrate consumption in each meal as well as the BMI of the user. Lumen’s measured exhaled %CO$_2$ was higher when carbohydrate intake was higher, in accordance with previous studies describing the association between RER and carbohydrates [11,14,15,18]. In addition, it reveals metabolic inflexibility in individuals with obesity as they show a reduced %CO$_2$ response compared with healthy individuals, which again is in accordance with the literature on metabolic flexibility [8,9,26].

Comparison With Previous Work

A recent prospective study has shown that a high-carbohydrate diet results in an elevation of Lumen’s %CO$_2$ compared with a low-carbohydrate diet [18]. In this retrospective analysis of Lumen users, we show similar results, as %CO$_2$ is elevated when carbohydrate consumption is increased, which highlights Lumen’s ability to detect changes in carbohydrate intake. Although previous studies that included high-fat and high-protein consumption resulted in elevated RER [14,27,28], we did not observe this in this study. Further study with a larger sample size is warranted to check for the effect of dietary fats and proteins on exhaled %CO$_2$.

The duration of %CO$_2$ elevation has shown a consistent plateau, spanning from 45 minutes postmeal to the last checkup at 210 minutes, where a subsequent decline was observed. Lumen users often analyze their postmeal breath between 60 and 120 minutes, so understanding how %CO$_2$ responds over longer periods is challenging. Woerle et al [29] showed carbohydrate oxidation up to 360 minutes after meal consumption, long after glucose and insulin levels in the plasma peaked. Future prospective studies should address when %CO$_2$ decreases after a mixed macronutrient meal and how it changes compared to other metabolites.

The impact of postprandial measurement of glucose levels on metabolic disease prevention has been discussed in previous studies [30]. Novel technological advancements have made it available to consumers, mostly with recent developments in continuous glucose monitoring (CGM) [31]. Other metabolic indicators, such as insulin, lipids, metabolomics, and indirect calorimetry, have also been used to assess metabolic flexibility and were found to be predictive of body weight, metabolic syndrome, T2D, and CVDs [11,25,32,33]. As of today, CGM devices need replacing every 2 weeks and are somewhat invasive, whereas most of the other aforementioned measurements can only be performed in a laboratory setting. In contrast, the Lumen device can be used at home and measures an individual’s metabolic state with a simple breath maneuver, which was found to be in agreement with the RER measured in the metabolic cart [17]. Moreover, compared to the metabolic cart, the portability of the Lumen device enables the collection of CO$_2$ measurements from a large number of participants after
consuming mixed macronutrient meals, which provides a higher level of external validity.

In addition to the effect of carbohydrate consumption, BMI was also found to be a key parameter in the postprandial %CO2 response, as a higher BMI resulted in a reduced response. These results are in accordance with the literature from the metabolic cart, where participants with obesity and T2D showed low ΔRER after insulin stimulation [9,10].

In this analysis, the majority of participants were women (2137/2607, 81.97%), in line with similar nutrition and weight management studies finding that women participate 3 times more often than men [34,35]. Age and gender were both statistically insignificant in their effect on the postprandial %CO2 response in the LMM, while some studies mention that metabolic flexibility is reduced with age [36]. Although progesterone is directly associated with CO2 [37,38], our analysis did not reveal any influence of gender on metabolic flexibility, possibly since women may have been in different phases of their menstrual cycle or at menopause. Future studies should investigate the effects of gender and the menstrual cycle on Lumen’s %CO2.

This study shows the potential of the Lumen device and mobile app to improve outcomes for people with metabolic disorders. There is increasing evidence showing that mobile health technologies can improve metabolic outcomes, in particular for lifestyle modifications in T2D and weight loss [39-41]. While most of these technologies incorporate a mobile app only, many studies have shown the benefits of an app accompanied by a device capable of providing real-time feedback on how those lifestyle modifications affect their measurements, which can improve engagement with them [42-45]. In a recent pilot clinical study, Lumen device and app usage improved several metabolic parameters in prediabetic patients after 3 months [46]. In light of these findings, the potential public health implications of mobile and ubiquitous health tools such as Lumen are noteworthy, showing great promise for addressing the epidemic of metabolic dysfunction across the globe in the near future.

Limitations
Despite the large data set of real-world evidence that was used in this analysis, several limitations need to be mentioned. First, due to its retrospective and observational nature, we cannot identify causal relationships between any of the variables in this study. Second, since this analysis is based on real-world data, the macronutrient consumption of each meal may be miscalculated or incomplete (recall bias), and a misuse of the device might occur as well. Moreover, the heterogeneous composition of the meals makes it difficult to determine if other factors in the meal might influence the %CO2 as well, in particular the type of carbohydrate and the fiber composition. In addition, Lumen users’ macronutrient intake might not be representative of most typical diets, as the Lumen app guides them toward a specific diet, primarily low in carbohydrate, and with most of them on a weight loss journey. As this study has specific user characteristics, caution should be exercised in interpreting these conclusions for a broader audience. Furthermore, this analysis did not consider the fastest %CO2 levels, which, together with the postprandial levels, could also be useful in the metabolic flexibility assessment [47]. Nonetheless, we believe that due to the nature of the outcome variable, which is the percentage of change from premeal to postmeal, it might have limited to no effect on the current results.

Lastly, postprandial %CO2 might be affected by other confounding variables, including preexisting conditions; medication use; dietary restrictions; the menstrual cycle; and lifestyle factors such as stress, sleep, and physical activity, which were not controlled for in this study. Future studies, both prospective and retrospective, should aim to address these limitations, particularly through controlled prospective investigations involving diverse populations and retrospective analyses that comprehensively consider potential covariates.

Conclusions
In conclusion, Lumen’s ability to evaluate metabolic flexibility following mixed meals was demonstrated through this retrospective analysis of postprandial exhaled %CO2, in which increased %CO2 was specific to carbohydrate consumption but not to fat consumption. Furthermore, a moderate %CO2 response was also observed among users with a high BMI, which suggests metabolic inflexibility among this group. As such, we propose Lumen as an accessible tool allowing individuals to make informed dietary choices conducive to metabolic health.

Acknowledgments
This study was funded by Metaflow Ltd. The authors would like to express their thanks to all app users who contributed their data to this study and to the Lumen team for their support. We would also like to thank all the reviewers for their timely and constructive inputs to this manuscript.

Authors’ Contributions
SY, TC, and MM led the design and conceptualization of the study. MM was the principal supervisor of the study. SY led the curation of the data and conducted data interpretation and formal analysis. SY wrote the first draft of the manuscript. TC and DS contributed to the data analysis. All authors critically revised the manuscript and gave their final approval.
Conflicts of Interest

MM is the chief scientific officer and the cofounder of MetaFlow Ltd. DS is a consultant at MetaFlow. SY and TC are paid employees at MetaFlow.

References


24. PRISM: analyze, graph and present your scientific work. GraphPad. URL: https://www.graphpad.com/ [accessed 2024-03-12]


37. PRISM: analyze, graph and present your scientific work. GraphPad. URL: https://www.graphpad.com/ [accessed 2024-03-12]


Abbreviations

- **BRANY**: Biomedical Research Alliance of New York
- **CGM**: continuous glucose monitoring
- **CO₂**: carbon dioxide
- **CVD**: cardiovascular disease
- **IRB**: institutional board review
- **LMM**: linear mixed model
- **RER**: respiratory exchange ratio
- **SBER**: Social, Behavioral, and Educational Research
- **T2D**: type 2 diabetes
- **USDA**: United States Department of Agriculture

©Shlomo Yeshurun, Tomer Cramer, Daniel Souroujon, Merav Mor. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 01.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
The Roles of Trust in Government and Sense of Community in the COVID-19 Contact Tracing Privacy Calculus: Mixed Method Study Using a 2-Wave Survey and In-Depth Interviews

Hyunjin Kang¹, PhD; Jeong Kyu Lee², PhD; Edmund WJ Lee¹, PhD; Cindy Toh³, BA

¹Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore, Singapore
²Department of Health and Exercise Science, University of Oklahoma, Norman, OK, United States
³Department of Anthropology, Columbia University, New York, NY, United States

Abstract

Background: Contact tracing technology has been adopted in many countries to aid in identifying, evaluating, and handling individuals who have had contact with those infected with COVID-19. Singapore was among the countries that actively implemented the government-led contact tracing program known as TraceTogether. Despite the benefits the contact tracing program could provide to individuals and the community, privacy issues were a significant barrier to individuals’ acceptance of the program.

Objective: Building on the privacy calculus model, this study investigates how the perceptions of the 2 key groups (ie, government and community members) involved in the digital contact tracing factor into individuals’ privacy calculus of digital contact tracing.

Methods: Using a mixed method approach, we conducted (1) a 2-wave survey (n=674) and (2) in-depth interviews (n=12) with TraceTogether users in Singapore. Using structural equation modeling, this study investigated how trust in the government and the sense of community exhibited by individuals during the early stage of implementation (time 1) predicted privacy concerns, perceived benefits, and future use intentions, measured after the program was fully implemented (time 2). Expanding on the survey results, this study conducted one-on-one interviews to gain in-depth insights into the privacy considerations involved in digital contact tracing.

Results: The results from the survey showed that trust in the government increased perceived benefits while decreasing privacy concerns regarding the use of TraceTogether. Furthermore, individuals who felt a connection to community members by participating in the program (ie, the sense of community) were more inclined to believe in its benefits. The sense of community also played a moderating role in the influence of government trust on perceived benefits. Follow-up in-depth interviews highlighted that having a sense of control over information and transparency in the government’s data management were crucial factors in privacy considerations. The interviews also highlighted surveillance as the most prevalent aspect of privacy concerns regarding TraceTogether use. In addition, our findings revealed that trust in the government, particularly the perceived transparency of government actions, was most strongly associated with concerns regarding the secondary use of data.

Conclusions: Using a mixed method approach involving a 2-wave survey and in-depth interview data, we expanded our understanding of privacy decisions and the privacy calculus in the context of digital contact tracing. The opposite influences of privacy concerns and perceived benefit on use intention suggest that the privacy calculus in TraceTogether might be viewed as a rational process of weighing between privacy risks and use benefits to make an uptake decision. However, our study demonstrated that existing perceptions toward the provider and the government in the contact tracing context, as well as the perception of the community triggered by TraceTogether use, may bias user appraisals of privacy risks and the benefits of contact tracing.

(JMIR Mhealth Uhealth 2024;12:e48986) doi:10.2196/48986
Introduction

Background

Contact tracing technology has been implemented in many countries to help identify, assess, and manage individuals exposed to persons infected with COVID-19 [1]. The main aims of a digital contact tracing system are not only to notify those who may have contracted the virus unknowingly but also to provide corresponding information and instructions from public health officials. Hence, some of contact tracing applications are designed to track the geolocations and mobility patterns of individuals using GPS and location-based solutions [2,3]. Digital contact tracing identifies individuals’ close contacts with people confirmed or suspected of having COVID-19 through Bluetooth technology [4] or even allows authorities to make decisions as to whether certain individuals are certified as healthy enough to enter public spaces [5]. The use and implementation of contact tracing apps during COVID-19 brings concerns of privacy intrusion and surveillance to the forefront, as the attempt to digitally map and trace the spread of the virus is conducted through apps that collect some of the most personal and sensitive data. Thus, privacy concerns were identified as a significant barrier to adopting the contact tracing system, significantly influencing the intention to participate in the program [6,7]. Applying the privacy calculus model, that is, the extent to which individuals are willing to disclose personal information after weighing the risks and benefits, this study examines how individuals’ perceptions of two important key entities—government and community members—involves in digital contact tracing play roles in their privacy calculus for adopting contact tracing. First, most contact tracing systems, including those in Singapore, Ireland, and France, were introduced and managed by the national government [6]. Derived from the social exchange concept, prior research on privacy suggests that trust in a service provider is a significant determinant of privacy calculus on the web [8,9]. Considering the managerial roles played by the government for contact tracing measures and the extensive personal information collected and analyzed for this effort, we postulated that individuals’ trust in the government would be the key influencer of their privacy calculus in digital contact tracing.

Community members are another key entity involved in the privacy calculus in digital contact tracing. The active participation of community members in the contact tracing program is crucial for the success of preventive efforts. However, digital contact tracing implies that the participants are connected through a web-based system for tracking, which may trigger privacy risks among users. Therefore, the sense of community triggered by the contact tracing measure may influence the relationship between privacy concerns and benefits, shaping one’s intention to participate in contact tracing efforts.

This study focuses on understanding privacy calculus in the context of using TraceTogether, a nationwide contact tracing program implemented in Singapore during the COVID-19 pandemic. The Government Technology Agency of Singapore, also known as GovTech, launched TraceTogether on March 20, 2020, followed by the launch of a token on June 28 to increase accessibility for those without a smartphone. In late December 2020, it became mandatory to check using TraceTogether before entering various places, including restaurants, workplaces, schools, places of worship, and shopping malls. As COVID-19 prevention measures eased, TraceTogether check-ins were no longer required in most settings from April 26, 2022.

Objective

This study investigates how trust in government and the sense of community during the early stage of contact tracing influence privacy calculus in digital contact tracings, as well as the ways and reasons why trust in government and sense of community impact privacy calculus in contact tracing for COVID-19 prevention.

Literature Review

Theoretical Framework: Privacy Calculus Theory

Overview

Privacy concerns about contact tracing programs during the COVID-19 pandemic have been raised because of the potential risks of extensive data collection, surveillance, and the possibility of unauthorized access [6,7], necessitating a careful balance between public health benefits and safeguarding individuals’ private information. Thus, privacy calculus theory has been used as one of the most robust theories within the field of communication [10] to unpack the underlying theoretical mechanisms in explaining the use of contact tracing apps through the lens of privacy. Having its theoretical roots and origins in economics and social exchange theory, individuals are assumed to be rational in making decisions and weighing trade-offs, and these decisions can be mathematically modeled as a function of the interaction between the benefits and risks of adopting a behavior [11]. For these reasons, privacy calculus theory offers a useful framework for investigating the influence of competing positive and negative beliefs on individuals’ willingness to engage with technologies by capturing the extent of trade-offs individuals are willing to make in relation to privacy [12]. For instance, despite knowing that companies such as Google or Waze could identify individuals’ precise geocoordinates, individuals who commute by driving may be willing to disclose this otherwise sensitive information (where they are and their mobility patterns) in exchange for the benefit of getting to their desired locations.
Privacy Concerns About Digital Contact Tracing

At the heart of privacy calculus theory is the act of self-disclosure of personal data, which is a product of two opposing forces: privacy concerns and the perceived benefits of information disclosure. Privacy concerns can be defined as individuals’ fears and apprehensions over the possible loss of privacy, as well as how their personal data could be used or abused [13,14]. Privacy concerns are multidimensional by nature, for example, [15,16]. Privacy concerns encompass 3 dimensions: perceived surveillance, perceived intrusion, and secondary use of information [16]. At a fundamental level, privacy concerns are the worries that people are being watched. Perceived surveillance encapsulates this notion and refers to individuals’ perceptions of the act of data collection pertaining to their lives—whether legitimate or illegal [17]. In the digital age, this could take the form of devices and technological platforms that capture individuals’ personal activities by recording their behaviors [18]. For instance, people living in Florida expressed concerns over the Florida Department of Health’s Healthy Together contact tracing app in collecting data such as contact lists, phone numbers, and medical conditions [19]. Relatedly, perceived intrusion is another dimension of privacy concerns and refers to the extent to which others are able to make independent decisions about possessing or soliciting information. For instance, in China, the government implemented a QR code system for contact tracing and assigned colors, such as green, yellow, or red, depending on the health status of individuals and the risk of exposure by triangulating data from public transport and health care systems [20], which determined whether they had access to public places [21]. The third dimension of privacy concerns is the secondary use of information, defined as the use of personal data collected from individuals for purposes other than what it was intended for [22]. As a large volume of data were collected by contact tracing apps, there were fears over how the government would use (or abuse) potentially sensitive data [20,21].

In the context of contact tracing app use, a recent study showed that when people are concerned about the extent of surveillance and intrusion of the implementation of contact tracing apps, this may negatively influence their intention to use, as fears of public health surveillance could be transformed into routine monitoring of populations extending beyond the purpose of infectious disease management [23]. This is consistent with prior evidence of privacy concerns and different facets of technology use or behaviors to engage in privacy protection behaviors [13,24]. In Singapore, after the government announced that police would have access to TraceTogether data for criminal investigations and after correcting perceptions that the data were only collected for contact tracing and no other purpose, some members of the public expressed concerns and disappointment and indicated their intention to use it less [25].

Perceived Benefits

In this study, perceived benefits were conceptualized as the health benefits that individuals would receive by using contact tracing apps, such as protecting them from COVID-19, making them more informed, or improving the overall public health capacity in managing and keeping infections low [12]. Hence, the use of contact tracing during the COVID-19 pandemic addresses a common problem by providing a crucial tool for communal health management, where the benefits extend beyond individual interests.

Although the loss of personal privacy is a major concern for different populations during the use of COVID-19 contact tracing apps, some have recognized the potential benefits of digital surveillance. First, advocates of “big data systems” would argue that tracking and mapping the spread of infectious diseases through digital means (eg, electronic health records) is more efficient than the traditional means of calling and asking people to remember their close contacts during the day and would reduce the burden on health care workers [26]. The data collected could enable public health agencies to quickly map potential hotspots in real time [27] and take decisive actions to curb the spread of COVID-19. Second, the use of contact tracing technology may help users feel empowered. Research has shown that individuals may gain psychological relief and peace of mind by using contact tracing apps as they provide a sense of empowerment and certainty—having data and knowledge about whether they were in close proximity with the infected, as they coped with the ambiguity of the pandemic by monitoring and interacting with their own health data [19]. A study conducted in Sweden [7] found that the benefits of contact tracing apps were positively associated with willingness to use them. They found that the perceived prosocial usefulness of contact tracing apps was positively associated with willingness to use contact tracing apps. Research has also underscored the prevalence of the privacy paradox that, despite concerns over privacy loss, people are willing to accept and use apps as long as they are perceived to be beneficial in lowering infection rates [7].

Trust in Government and Sense of Community as Antecedents of Privacy Calculus in TraceTogether

Overview

Communication privacy management theory, built upon the privacy calculus model, uses the metaphor of “boundaries” to explain the motivation to reveal or withhold information [28]. According to this theory, boundary openness or closure for information flow is largely dependent on how individuals perceive institutional privacy assurances. In other words, the cognitive processes of risk control assessment and privacy concerns are context specific. Thus, privacy-relevant beliefs are “better related to individuals’ own information experiences and social contexts rather than regarded as a global consequence of technology use per se” [16]. This approach aligns with the notion that privacy is a dynamic process [29,30]. This study also investigated the antecedents that shape privacy calculus in the context of digital contact tracing.

Specifically, we posit that the service provider (ie, government) of the tracing system and other community members who use the device to collectively fight against COVID-19 would be the key entities that would influence individual users’ intentions to use a contact tracing device. This study examines how (1) trust in government and (2) the sense of community triggered by the use of contact tracing technology are associated with one’s privacy calculus regarding digital contact tracing, thereby shaping one’s intention to adopt digital tracing technology.
Trust in Government

Social exchange theory suggests that users disclose personal information to gain intangible benefits from a relationship, given that these perceived benefits outweigh the perceived risks [31]. The benefits of a social exchange might differ from an economic exchange in that the former may rely on existing social ties and involve intangible values, including emotions and social power [31]. Similar to other contact tracing programs, TraceTogether is government-initiated. Therefore, user acceptance and adoption are deeply linked to Singaporeans’ trust and confidence in their government [26].

Trust is established when individuals or groups harboring positive perceptions of one another enable the relationship to achieve anticipated results [32]. When an individual places trust in another person, group, or organization, they find themselves liberated from the burden of anxiety and obligation to constantly monitor the actions of the other party [33]. Thus, it can be expected that people who use the contact tracing system will perceive social exchange as more reliable if they trust the government. Supporting this notion, prior studies on web-based privacy have demonstrated that trust in web-based service providers or vendors significantly influences privacy management decisions by decreasing privacy concerns [8,9].

A study in Singapore [34] found that political trust could mediate Singaporeans’ privacy concerns, and those who exhibited trust in their government tended to have more positive attitudes toward digital contact tracing technology. This finding is echoed in multiple studies on contact tracing apps in countries including France [35], Japan [36], Germany [37], and the United Kingdom [38].

Trust in service providers can also be critical to increasing users’ perceived disclosure benefits [39]. Building on the concept of privacy calculus [40], one study [41] argues that gaining users’ willingness to disclose personal information requires the exchange to be based on a fair social contract. In other words, users find the benefits more attractive when they perceive the provider as trustworthy, leading to a positive bias in their privacy calculus. Taken together, we hypothesized that the effects of trust in the government on privacy calculus in TraceTogether would be as follows:

- **H1**: Trust in government increases the perceived benefit of using contact tracing apps.
- **H2**: Trust in government will decrease privacy concerns when using contact tracing apps.

Sense of Community Triggered by Contact Tracing

The effectiveness of the contact tracing program for infectious disease prevention largely depends on the participation rate of individuals in the community. Moreover, recognizing the collective action nature of contact tracing underscores its significance in navigating the social dilemma posed by the virus, emphasizing the shared responsibility in curbing the spread for the greater well-being of the community. To enhance the uptake rate of TraceTogether, one way in which the government has promoted the application to Singaporeans is by framing digital contact tracing as a community effort and responsibility. The TraceTogether interface included various elements designed to encourage feelings of closeness and connectedness with other users. For example, the TraceTogether app indicates the number of activated devices nearby and shows how many Bluetooth signals are exchanged among them. The language used in an official promotional video for TraceTogether also reinforced this, depicting the tool as being for “‘community-driven’ contact tracing suggesting that this is a ‘grassroots’ mechanism of fighting the virus” [42]. Therefore, group membership and collective benefit have been leveraged to encourage user adoption and acceptance of TraceTogether.

However, the contact tracing program connects those in the vicinity via Bluetooth technology to estimate the proximity and duration of encounters, which may also trigger privacy concerns. Communication privacy management theory [28] suggests that privacy management is not just about deciding to disclose or withdraw personal information with others (personal boundaries). When one’s personal information is shared, those who can access it become co-owners of the information, suggesting that privacy management also involves co-managing collective boundaries. Thus, the disruption of the synchronized management of collective privacy boundaries among co-owners is a significant source of privacy risk perception. Although individual participants in the contact tracing program cannot access others’ information directly, contact tracing implies that people encountered may be notified of one’s infection. Moreover, signaling connectivity with others through interface cues may trigger privacy risk heuristics, heightening privacy concerns among users [43].

Collectively, we expect that the perception of “the greater good” for the improved safety of loved ones and the community is a significant factor influencing the privacy calculus of being part of the contact tracing program [44,45]. This study hypothesizes a paradoxical function of the sense of community such that the sense of community triggered by digital contact tracing will enhance perceived benefits while simultaneously presenting privacy concerns at the same time:

- **H3**: The sense of community triggered by contact tracing apps increases the perceived benefits of contact tracing apps.
- **H4**: The sense of community triggered by contact tracing apps will increase the privacy concerns of contact tracing apps.

This study also examined the interplay between one’s perception of the government and the community on privacy calculus in digital contact tracing. Although trust in government has developed over time, a sense of community can be promoted through various design and marketing elements of digital contact tracing. As discussed above, TraceTogether was promoted to Singaporeans by framing contact tracing as a community effort and responsibility [46]. The user interface of the application was also designed to imbue users with a sense of community through various design elements. Thus, the study attempts to test whether one’s trust in government factors in the privacy calculus in digital contact tracing would vary as a function of the larger community that they perceive through participation in the collective contact tracing effort. However, given the absence of literature documenting the moderating role of sense of
community in the relationship between trust in government and privacy calculus, this study investigates the interplay between sense of community and trust in government, proposing the following research question:

RQ1. Will the sense of community generated by contact tracing apps moderate the influence of trust in government on privacy concerns and perceived benefits of using contact tracing apps?

Finally, based on the basic tenet of privacy calculus theory discussed earlier, we also hypothesize that privacy concerns and perceived benefits will have opposite associations with the intention to use TraceTogether. Given that the use of a contact tracing system itself implies the disclosure of personal information, we tested the intention to use TraceTogether as the outcome of the privacy calculus. Checking through TraceTogether was mandatory to enter most venues, including shopping malls, offices, workplaces, and schools, at the time of the second wave of data collection when use intention had been measured. Therefore, the measurement was designed to assess their intention to use TraceTogether voluntarily until the pandemic ended, rather than merely assessing their intention to use the device for a certain duration. The study model and methodology are illustrated in Figure 1.

H5: Perceived benefits are positively associated with the intention to actively participate in contact tracing programs.

H6: Privacy concerns are negatively associated with the intention to actively participate in the contact tracing programs.

Figure 1. Summary of the study model and methodology. H: hypothesis; RQ: research question.

Study 1: Two–Wave Longitudinal Survey

Methods

Recruitment

To test our hypotheses and research questions, we used a mixed method approach involving a 2-wave survey and in-depth interviews. The first survey was conducted from January to February 2021 with panel members provided by a local research company. English-speaking Singaporean citizens or permanent residents aged 21 years or above participated in the survey (N=1198). The quota sampling procedure was employed to ensure that the study sample represented the proportion of the ethnic structure of the national population [47]. The first study was conducted in an omnibus survey format to understand the various perceptions and behaviors related to COVID-19 and contact tracing practices. Among all participants who completed the survey, 94.91% (n=1137) responded that they had downloaded the TraceTogether mobile app or collected the token.

A follow-up second-wave survey was carried out in August 2021, approximately 6 months after the first survey. This survey assessed privacy calculus factors: privacy concerns, perceived benefits, and adoption intention. A second-wave survey invitation was sent to those who completed the first survey. In total, 674 participants completed the second-wave survey (response rate: 59.3%). Data from those who either downloaded the TraceTogether mobile app or received the token and who completed both studies at the point of the first wave of data collection were included in the study sample (n=626; see Table 1 for the descriptive statistics of the study sample). Our sample comprised 88.7% (n=555) Chinese, 4.3% (n=27) Malay, and 7% (n=44) Indian individuals. In comparison to the ethnic structure of the national population in Singapore, where Chinese people constitute 74%, Malay people 13%, and Indian people 9% [48], our sample overrepresented the Chinese group and underrepresented the Malay group. Moreover, the sample included only a small portion of participants aged older than 65 years (n=25; 4%), thereby underrepresenting the older population. During the first wave of data collection (January-February 2021), the TraceTogether app or token could be used for checking at selected venues where people were likely to be in close contact for prolonged periods or where high human traffic is expected. These venues included schools, educational institutes, shopping malls, restaurants, and workplaces. People can still choose to scan the QR code.
displayed at the entrance instead of using TraceTogether. From June 2021, the use of TraceTogether became mandatory when entering these venues.

Table 1. Descriptive statistics.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Values (N=626)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>332 (53)</td>
</tr>
<tr>
<td>Female</td>
<td>294 (47)</td>
</tr>
<tr>
<td><strong>Age group (y), n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>21-34</td>
<td>130 (20.8)</td>
</tr>
<tr>
<td>35-49</td>
<td>281 (44.9)</td>
</tr>
<tr>
<td>50-64</td>
<td>190 (30.3)</td>
</tr>
<tr>
<td>65 or older</td>
<td>25 (4)</td>
</tr>
<tr>
<td>Mean age (y), mean (SD)</td>
<td>44.9 (11)</td>
</tr>
<tr>
<td><strong>Ethnicity, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>555 (88.7)</td>
</tr>
<tr>
<td>Malay</td>
<td>27 (4.3)</td>
</tr>
<tr>
<td>Indian</td>
<td>44 (7)</td>
</tr>
<tr>
<td><strong>Education level, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Secondary and below</td>
<td>13 (2)</td>
</tr>
<tr>
<td>Preuniversity</td>
<td>270 (43.1)</td>
</tr>
<tr>
<td>University degree or above</td>
<td>343 (54.8)</td>
</tr>
<tr>
<td><strong>Yearly household income (Singaporean $\text{a}$), n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;2000</td>
<td>37 (5.9)</td>
</tr>
<tr>
<td>2000-6000</td>
<td>185 (29.6)</td>
</tr>
<tr>
<td>6000-10,000</td>
<td>241 (38.5)</td>
</tr>
<tr>
<td>≥10,000</td>
<td>163 (26)</td>
</tr>
<tr>
<td><strong>Self-assessed health status, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Poor</td>
<td>18 (3)</td>
</tr>
<tr>
<td>Fair</td>
<td>163 (26)</td>
</tr>
<tr>
<td>Good</td>
<td>293 (46.8)</td>
</tr>
<tr>
<td>Very good</td>
<td>126 (20.1)</td>
</tr>
<tr>
<td>Excellent</td>
<td>26 (4)</td>
</tr>
<tr>
<td><strong>TraceTogether adoption</strong></td>
<td></td>
</tr>
<tr>
<td>Have downloaded TraceTogether app (Yes)</td>
<td>Wave 1: 494 (78.9); wave 2: 586 (93.6)</td>
</tr>
<tr>
<td>Have collected TraceTogether token (Yes)</td>
<td>Wave 1: 452 (72.2); wave 2: 549 (87.7)</td>
</tr>
</tbody>
</table>

\(\text{a}\)Singaporean $1=US$0.75.

**Measurement**

The two exogenous variables (government trust, sense of community) and control variables (demographic information and TraceTogether use) were measured in the first-wave survey. Privacy calculus factors—privacy concerns and perceived benefits—and TraceTogether use intention were measured in the second-wave survey.

Trust in government was measured using 3 items adapted from Wu et al [49]. The sense of community was measured using 5 items adapted from 2 studies [50,51]. The sense of community captured the extent to which users felt a sense of belonging and emotional ties with other TraceTogether users. Privacy concerns were assessed using 10-item measures which captured 3 aspects of privacy concerns, namely perceived surveillance, perceived intrusion, and secondary use of personal information [16]. The perceived benefits of using TraceTogether were measured with 5 items that assessed the extent to which the participant perceived the health and information benefits they received from using TraceTogether [52]. Finally, use intention was
measured using three items. Checking TraceTogether was mandatory to enter most public venues. Therefore, the measurement assessed their intention to use TraceTogether voluntarily until the pandemic ended, rather than asking about their intention to use the device within a certain period. All items used a 7-point Likert scale (1=strongly disagree to 7=strongly agree) unless otherwise indicated. Table 2 lists the measurement items and their factor loadings.

<table>
<thead>
<tr>
<th>Variable and items</th>
<th>Values, mean (SD)</th>
<th>Factor loading, β</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Trust in government</strong></td>
<td>4.80 (1.53)</td>
<td>.91</td>
</tr>
<tr>
<td>Even if not monitored, I would trust the government to do the right thing.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I trust the government to protect my personal information.</td>
<td></td>
<td>.96</td>
</tr>
<tr>
<td>I believe that the government is trustworthy.</td>
<td></td>
<td>.95</td>
</tr>
<tr>
<td><strong>Sense of community (sentences complete the phrase “When I use TraceTogether...”)</strong></td>
<td>4.51 (1.62)</td>
<td>.89</td>
</tr>
<tr>
<td>It makes me feel like I am part of a community.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel a high sense of attachment with other users.</td>
<td></td>
<td>.94</td>
</tr>
<tr>
<td>I feel an emotional connection with other users.</td>
<td></td>
<td>.93</td>
</tr>
<tr>
<td>It reminds me of the people around me.</td>
<td></td>
<td>.90</td>
</tr>
<tr>
<td>It makes me feel a sense of belonging.</td>
<td></td>
<td>.94</td>
</tr>
<tr>
<td><strong>Privacy concerns</strong></td>
<td>4.71 (1.50)</td>
<td>.83</td>
</tr>
<tr>
<td>I believe that my location is being monitored in real time when using TraceTogether.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am concerned that TraceTogether is collecting too much information about me.</td>
<td></td>
<td>.88</td>
</tr>
<tr>
<td>I am concerned that TraceTogether may monitor my activities.</td>
<td></td>
<td>.91</td>
</tr>
<tr>
<td>I am concerned that my activities are being monitored.</td>
<td></td>
<td>.91</td>
</tr>
<tr>
<td>I feel that as a result of using TraceTogether, others know more about me than I am comfortable with.</td>
<td></td>
<td>.88</td>
</tr>
<tr>
<td>I believe that as a result of using TraceTogether, information about me that I consider private is now more readily available to others than I prefer.</td>
<td></td>
<td>.90</td>
</tr>
<tr>
<td>I feel that as a result of using TraceTogether, information about me is out there that, if used, will invade my privacy.</td>
<td></td>
<td>.90</td>
</tr>
<tr>
<td>I am concerned that TraceTogether may use my personal information for other purposes without notifying me or getting my authorization.</td>
<td></td>
<td>.90</td>
</tr>
<tr>
<td>I am concerned that TraceTogether may use my information for other purposes.</td>
<td></td>
<td>.87</td>
</tr>
<tr>
<td>I am concerned that TraceTogether may share my personal information with other entities without getting my authorization.</td>
<td></td>
<td>.87</td>
</tr>
<tr>
<td><strong>Perceived benefits</strong></td>
<td>5 (1.39)</td>
<td>.86</td>
</tr>
<tr>
<td>Using TraceTogether would improve my access to my health information related to COVID-19.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using TraceTogether would improve my access to my health information related to COVID-19.</td>
<td></td>
<td>.87</td>
</tr>
<tr>
<td>Using TraceTogether would improve my ability to manage my health.</td>
<td></td>
<td>.93</td>
</tr>
<tr>
<td>Using the COVID app would improve the quality of health care.</td>
<td></td>
<td>.91</td>
</tr>
<tr>
<td>I would manage my health more effectively using TraceTogether.</td>
<td></td>
<td>.90</td>
</tr>
<tr>
<td><strong>TraceTogether use intention</strong></td>
<td>5.23 (1.40)</td>
<td>.86</td>
</tr>
<tr>
<td>I will use TraceTogether even if TraceTogether becomes optional when entering public venues.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am willing to use TraceTogether until the COVID pandemic ends.</td>
<td></td>
<td>.74</td>
</tr>
<tr>
<td>I plan to use TraceTogether even if I am not monitored.</td>
<td></td>
<td>.86</td>
</tr>
</tbody>
</table>

**Ethical Considerations**

The study procedure and questionnaire were reviewed and approved by Nanyang Technological University Institutional Review Board (IRB-2022-213). A consent form was displayed on the first page of the survey, and the participants were required to provide their consent to complete the study.
Results

Measurement Model

Confirmatory factor analysis was conducted using Mplus (version 8.8; Muthen & Muthen) statistical software with a maximum likelihood estimation method. The confirmatory factor analysis results indicated a reasonable fit to data: $\chi^2 = 937.9$ ($P < .001$), root mean square error of approximation=0.06 (90% CI 0.056-0.065), comparative fit index=0.97, and Tucker-Lewis index=0.96. The indicators reflected their respective latent variables, as evidenced by the high factor loadings. The magnitude of all factor loadings and Cronbach $\alpha$ were equal to or greater than 0.74 and 0.86, respectively. The measurement model showed robust convergent and discriminant validity: composite reliability and average variance extracted (AVE) were greater than 0.70 and 0.50, respectively. The square roots of the AVEs of all observed variables were greater than the intercorrelations between variables. All factors were significantly correlated with each other. See Table 3 for composite reliability, AVE, and Cronbach $\alpha$, and see Table 4 for the correlation matrix.

Table 3. Validity and reliability.

<table>
<thead>
<tr>
<th></th>
<th>CR $^a$</th>
<th>AVE $^b$</th>
<th>Cronbach $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government trust</td>
<td>0.958</td>
<td>0.885</td>
<td>0.958</td>
</tr>
<tr>
<td>Sense of community</td>
<td>0.964</td>
<td>0.844</td>
<td>0.964</td>
</tr>
<tr>
<td>Privacy concerns</td>
<td>0.973</td>
<td>0.783</td>
<td>0.974</td>
</tr>
<tr>
<td>Perceived benefits</td>
<td>0.952</td>
<td>0.798</td>
<td>0.952</td>
</tr>
<tr>
<td>Behavior intention</td>
<td>0.862</td>
<td>0.676</td>
<td>0.859</td>
</tr>
</tbody>
</table>

$^a$CR: composite reliability.

$^b$AVE: average variance extracted.

Table 4. Correlation matrix (Pearson $r$).

<table>
<thead>
<tr>
<th></th>
<th>Government trust</th>
<th>Sense of community</th>
<th>Privacy concerns</th>
<th>Perceived benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sense of community</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r$</td>
<td>0.63</td>
<td>— $^a$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P value</td>
<td>&lt;.001</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Privacy concerns</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r$</td>
<td>-0.27</td>
<td>-0.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P value</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived benefits</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r$</td>
<td>0.48</td>
<td>0.54</td>
<td>-0.17</td>
<td></td>
</tr>
<tr>
<td>P value</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>Behavior intention</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r$</td>
<td>0.44</td>
<td>0.47</td>
<td>-0.21</td>
<td>0.63</td>
</tr>
<tr>
<td>P value</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

$^a$Not applicable.

Model Testing

The structural equation modeling (SEM) result without the interaction term in the model revealed a good fit: $\chi^2 = 1382.9$ ($P < .001$), root mean square error of approximation=0.048 (90% CI 0.045-0.052), comparative fit index=0.957, and Tucker-Lewis index=0.952. As these model fit indices have not been developed for the SEM models assessing latent interaction effects, we used a log-likelihood (~2 LL) ratio test to compare the SEM model where the interaction was not estimated (parsimonious model) with the model where the interaction was estimated (more complex model) [53]. The test was statistically significant, meaning that the model without the interaction term represented a significant loss of fit relative to the more complex model. Thus, we conclude that the model with the interaction term is well-fitted. In addition, the model with the interaction term had lower values of the 2 fit indices, Akaike information criteria and Bayesian information criteria, in comparison to the model without the interaction term.

Our analysis revealed that trust in the government significantly reduced privacy concerns of TraceTogether ($\beta = -0.39; P < .001$), whereas sense of community did not have a significant impact on privacy concerns ($\beta = 0.02; P = .86$). However, both government trust and sense of community significantly increased the perceived benefits of using TraceTogether (government trust: $\beta = 0.28; P < .001$; sense of community: $\beta = 0.54; P < .001$; H1, H2
and H3 were supported, while H 4 was not supported. Regarding RQ1, the results indicated that trust in the government and sense of community interacted only with perceived benefits (β = −.14; P = .004). The results are summarized in Figure 2. The interaction pattern (Figure 3) shows that the triggering sense of community has a larger impact on those who display lower trust in the government than on those with high trust. This finding suggests that triggering a sense of community can buffer the negative impact of skepticism in the government on their perceptions of the benefits they can obtain using TraceTogether.

**Figure 2.** Summary of structural equation model results. *P<.05, **P<.01, ***P<.001. Age, gender, ethnicity, education, income, and use frequency of TraceTogether use measured at time 1 were controlled.

Finally, supporting the privacy calculus hypothesis, privacy concerns decreased TraceTogether use intention (β = −.07; P = .04) whereas perceived benefit increased use intention (β = .53; P < .001). Hence, H5 and H6 were supported.

---

**Study 2: After In-Depth Interviews**

**Overview**

The survey results illustrated the effects of trust in the government and a sense of community on TraceTogether privacy calculus and, ultimately, TraceTogether users’ intentions. However, with the survey data, we could not identify which aspects of privacy concerns were factored in the privacy
calculus, influencing their intention to use TraceTogether. In addition, given the significant roles of trust in government and sense of community in TraceTogether and privacy calculus identified by the survey results, we conducted in-depth interviews to further understand the underlying reasons for these effects. The purpose of this study was two-fold: 1) to understand the reasons underlying the effects of government and sense of community on the privacy calculus of TraceTogether use, and 2) to obtain a deeper understanding of different aspects of privacy concerns among TraceTogether users and the implications for their privacy calculus.

Methods

Participants and Procedure

To obtain additional insights into the privacy calculus of TraceTogether, we conducted in-depth interviews with 12 Singaporeans between April 13 and May 11, 2022. We recruited Singaporeans (mean age 38, SD 13.28; 7 female participants, 5 male participants; see Table 5) who used TraceTogether during the COVID-19 pandemic. Among them, 7 were Chinese, 3 Malay, and 2 Indian. A trained research assistant conducted the interviews using a web-based conference program (ie, Zoom) in English. Participants were given a link to the consent form before starting the interview, which began once their consent was received. Each interview followed a preconstructed list of questions to understand their perceptions and use of TraceTogether. After completing the 4 interviews, the TraceTogether check-in was no longer used, starting from April 26. Subsequently, for the remaining 8 interviews, the participants were asked to answer questions reflecting their prior experiences with TraceTogether. Interview recruitment was stopped when we reached a state of theoretical saturation [54].

Table 5. Demographics of in-depth interview participants.

<table>
<thead>
<tr>
<th>ID</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Age (y)</th>
<th>Occupation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Male</td>
<td>Indian</td>
<td>24</td>
<td>Student</td>
</tr>
<tr>
<td>2</td>
<td>Male</td>
<td>Chinese</td>
<td>23</td>
<td>Student</td>
</tr>
<tr>
<td>3</td>
<td>Female</td>
<td>Chinese</td>
<td>23</td>
<td>Student</td>
</tr>
<tr>
<td>4</td>
<td>Female</td>
<td>Chinese</td>
<td>49</td>
<td>Manager</td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
<td>Chinese</td>
<td>56</td>
<td>Teacher</td>
</tr>
<tr>
<td>6</td>
<td>Female</td>
<td>Chinese</td>
<td>49</td>
<td>Teacher</td>
</tr>
<tr>
<td>7</td>
<td>Female</td>
<td>Malay</td>
<td>50</td>
<td>Teacher</td>
</tr>
<tr>
<td>8</td>
<td>Female</td>
<td>Malay</td>
<td>31</td>
<td>Teacher</td>
</tr>
<tr>
<td>9</td>
<td>Male</td>
<td>Malay</td>
<td>31</td>
<td>Clinician</td>
</tr>
<tr>
<td>10</td>
<td>Male</td>
<td>Chinese</td>
<td>51</td>
<td>Worker in the Singapore armed forces</td>
</tr>
<tr>
<td>11</td>
<td>Female</td>
<td>Chinese</td>
<td>49</td>
<td>Administrator</td>
</tr>
<tr>
<td>12</td>
<td>Female</td>
<td>Indian</td>
<td>24</td>
<td>Social media and project manager</td>
</tr>
</tbody>
</table>

Data Analysis

We used the 3-step protocol [55], which provides a structured framework for qualitative researchers to systematically analyze and derive meaningful interpretations from their data, involving 3 stages of data analysis: open-coding, axial and hierarchical coding, and theoretical interpretation, outlined as follows: First, 2 research assistants, including the interviewer, reviewed and coded the transcribed interviews independently. They read the transcribed interviews line by line and assigned the initial codes according to the themes of the original conversations (eg, concerns about using TraceTogether and adoption barriers). The second step was the axial and hierarchical coding stage; 2 researchers identified the second-level interpretive themes based on the theoretical framework and hypotheses (eg, surveillance, intrusion, and secondary use of data) by analyzing and grouping the initial coding performed by the 2 research assistants. Prior literature on web-based privacy calculus and privacy issues of digital contact tracing guided this coding stage. Finally, the researchers discussed how to interpret the themes that surfaced in the second stage in order to answer the research questions.

Ethical Considerations

Participants were given a link to the consent form before starting the interview, which began once their consent was received. The consent form, interview questions, and protocols were reviewed and approved by the primary investigator’s institute (IRB-2022-213).

Results

Overview

In a scenario where policies on and implementation of a contact tracing system were rolled out urgently and comprehensively, privacy became a significant barrier to adopting the program. Our study revealed that trust in the government serves as a social lubricant, effectively mitigating privacy concerns arising from the inherent risks associated with the tracing program, thereby facilitating smoother implementation of the program. Our interview results indicated that trust in the government translated into the belief that the government would possess adequate security measures to safeguard citizens’ privacy and prevent the misuse of their information. However, considering other community members while participating in the contact tracing program does not necessarily alleviate privacy concerns. The
The Role of Trust in Government in TraceTogether Privacy

The interview responses also suggested a significant influence of trust in the government on both perceived benefits and privacy concerns. Consistent with the findings of our web-based survey, the results indicated that individuals who trust the government were more likely to perceive TraceTogether as beneficial while exhibiting lower privacy concerns. For example, participant 7 (female, age 50 y) did not understand why the app should track user locations; therefore, she felt that she did not want to use it. However, she explained that she downloaded the app because she trusted the government, stating, “I trust that my government knows what they’re doing to the people. I trust that it is for the good of us.” Conversely, participant 12 (female, age 24 y), who indicated distrust in the government, did not perceive any additional benefits of using TraceTogether beyond using QR codes for checking-in, stating, “I feel like using the QR code would still work.” Furthermore, she voiced several privacy concerns, remarking, “What if this thing is tracking me 24/7? What if I am at home with my constant location at home, and my address is being transmitted?” She also referenced the public backlash at the start of 2021, when Foreign Minister Vivian Balakrishnan revealed that the police could access TraceTogether data for criminal investigations. This announcement contradicted his statement in June of the previous year that TraceTogether data would only be used for contact tracing [56]. She mentioned that she felt betrayed by government. Regarding the same incident, however, participant 5 (male, age 56 y) felt indifferent, stating: “There’s nothing to hide, so I am not concerned. Those who are criminals, they should be afraid. Good for them. They are better afraid than they should track user locations; therefore, she felt that she did not want to use it. However, she explained that she downloaded the app because she trusted the government, stating, “I trust that my government knows what they’re doing to the people. I trust that it is for the good of us.” Conversely, participant 12 (female, age 24 y), who indicated distrust in the government, did not perceive any additional benefits of using TraceTogether beyond using QR codes for checking-in, stating, “I feel like using the QR code would still work.” Furthermore, she voiced several privacy concerns, remarking, “What if this thing is tracking me 24/7? What if I am at home with my constant location at home, and my address is being transmitted?” She also referenced the public backlash at the start of 2021, when Foreign Minister Vivian Balakrishnan revealed that the police could access TraceTogether data for criminal investigations. This announcement contradicted his statement in June of the previous year that TraceTogether data would only be used for contact tracing [56]. She mentioned that she felt betrayed by government. Regarding the same incident, however, participant 5 (male, age 56 y) felt indifferent, stating: “There’s nothing to hide, so I am not concerned. Those who are criminals, they should be afraid. Good for them. They are better afraid than they should do anything wrong or I am a criminal... It’s just that... I don’t like to know that my personal particulars are being shared.” Participant 4 (female, age 49 y) also shared, “I can put it [token] at home...They cannot even trace me to home. They can trace the data on my phone.” These deliberate choices can be understood as small acts of resistance to mandated surveillance.

The respondents also shared that the digital contact tracing system, in fact, collected more information beyond what was necessary, such as how participant 4 (female, age 49 y) expressed her concerns. “Government is going to trace me where I go, what I do, who I meet. I feel very invaded by my privacy.” From the responses, we discovered that perceived intrusion becomes prominent because they do not know what information is to be tracked and used and the purpose of tracking clearly. Participant 2 (male, age 23 y) said, “I never know what they use the information for... They want to check whether we stay at home. It’s still not right because it’s an invasion of privacy.” Participant 4 (female, age 49 y) also raised concerns that TraceTogether would gather information beyond what has been told: “[if] I go to the shopping mall, and then they start to want to know how much you spend and say how much you spend in that shopping mall or what exact activities are being conducted.”

In addition, several respondents cited distrust in the government as a factor that increased privacy concerns, specifically with regard to the secondary use of data. Regarding the government’s announcement that the police could use TraceTogether data for criminal investigations, participant 3 (female, age 23 y) expressed her concerns, saying, “I thought it was a bit annoying because there was a lack of transparency about whatever they were going to use the app. If they wanted to do it for this stuff, they should have said it earlier and not let it be like, ‘Oops, I said it as an accident.’” The government’s transparency was revealed as an important factor in determining users’ privacy concerns about the secondary use of data. Transparency on how contact tracing apps work can not only affect user confidence, but potentially enhance user trust and acceptance.

Dissecting the Multiple Aspects of Privacy Concerns in TraceTogether

As discussed earlier, privacy concerns are not a single-faceted concept, but encompass multiple aspects, including perceived surveillance, perceived intrusion, and secondary use of data [16]. Through the postinterviews, we attempted to understand the different aspects of privacy concerns among TraceTogether users.

Surveillance stood out as the most prominent privacy concern, as mentioned by 7 of 12 respondents. Respondents explained that lack of control contributes to their fear and anxiety about being watched, even if they feel that they have nothing to hide. Some respondents preferred using a TraceTogether token to the mobile app to reclaim a sense of control over where and when their locations were traced. For example, participant 7 (female, age 50 y) mentioned, “I am not afraid because I don’t have anything to hide and that I am doing anything wrong or I am a criminal... It’s just that... I don’t like to know that my personal particulars are being shared.” Participant 4 (female, age 49 y) also shared, “I can put it [token] at home...They cannot even trace me to home. They can trace the data on my phone.” These deliberate choices can be understood as small acts of resistance to mandated surveillance.

The respondents also shared that the digital contact tracing system, in fact, collected more information beyond what was necessary, such as how participant 4 (female, age 49 y) expressed her concerns. “Government is going to trace me where I go, what I do, who I meet. I feel very invaded by my privacy.” From the responses, we discovered that perceived intrusion becomes prominent because they do not know what information is to be tracked and used and the purpose of tracking clearly. Participant 2 (male, age 23 y) said, “I never know what they use the information for... They want to check whether we stay at home. It’s still not right because it’s an invasion of privacy.” Participant 4 (female, age 49 y) also raised concerns that TraceTogether would gather information beyond what has been told: “[if] I go to the shopping mall, and then they start to want to know how much you spend and say how much you spend in that shopping mall or what exact activities are being conducted.”

In addition, several respondents cited distrust in the government as a factor that increased privacy concerns, specifically with regard to the secondary use of data. Regarding the government’s announcement that the police could use TraceTogether data for criminal investigations, participant 3 (female, age 23 y) expressed her concerns, saying, “I thought it was a bit annoying because there was a lack of transparency about whatever they were going to use the app. If they wanted to do it for this stuff, they should have said it earlier and not let it be like, ‘Oops, I said it as an accident.’” The government’s transparency was revealed as an important factor in determining users’ privacy concerns about the secondary use of data. Transparency on how contact tracing apps work can not only affect user confidence, but potentially enhance user trust and acceptance.
but also determine ethical considerations. Participant 2 (male, age 23 y) shared his point, stating, “If they ask for your consent and you decide to give it (tracked information), then sure. If they do it on their own, then that’s not okay.” Participant 1 (male, age 24 y) shared a similar point, explaining: “It’s just that I’m perfectly fine with them in using this app for in terms of national security or intelligence or what, but what I want them to do is be upfront and open.”

**Discussion**

**Theoretical Implications**

Using a mixed method approach involving a 2-wave survey and in-depth interview data, we expanded our understanding of privacy decisions and calculus in the context of digital contact tracing. Our study has several notable findings. First, we found that privacy concerns had a marginal negative association with use intention, whereas perceived benefits had a strong positive association, consistent with previous research [57,58]. The opposite influences of privacy concerns and perceived benefit on use intention suggest that the privacy calculus in TraceTogether might be viewed as a rational process of weighing privacy risks and use benefits to make an uptake decision. Second, although our findings on the relationships between trust in government and sense of community and privacy calculus are largely consistent with existing research [59], our study demonstrated that their existing perceptions toward the provider and the government in the contact tracing context, as well as the perception of the community triggered by TraceTogether use, may bias user appraisals of privacy risks and benefits of contact tracing. Finally, our research extends the existing knowledge of privacy calculus by delineating how the sense of community moderated the relationship between perceived benefit and trust in government, such that individuals’ sense of community elevated perceived benefit, even for those with low trust in government. The study results provides several important theoretical implications: First, the study illuminated how people’s trust in the system provider plays a role in privacy calculus in digital contact tracing systems. The digital contact tracing system, including TraceTogether, is a special form of a web-based system built and managed by government organizations during an unprecedented pandemic. Digital contact tracing entails delegating the right to track and process location, contact, and health-related information to the government’s discretion. Therefore, we posit that trust in government is a powerful factor in determining social exchange decisions.

As expected, trust in the government significantly influenced privacy calculus at the later stage when the use of TraceTogether became mandatory for entering most venues and workplaces. The results support prior research, which suggests the importance of trust in the government for the successful implementation of the contact tracing program for infectious disease prevention (eg, [32-36]). Trust in the government becomes pivotal in government-citizen relationships, especially during the implementation of new policies. This is largely because trust can reduce the transaction costs in relationships [60]. Our interview responses offer deeper insights into how trust in the government lowers transaction costs, specifically addressing privacy risks in the context of participating in the contact tracing program. We found that one’s privacy calculus can be biased according to the level of trust displayed by individuals, and an interviewee showing distrust in the government tends to exaggerate the privacy risk of TraceTogether while doubting its effectiveness without specific evidence. Conversely, another interviewee, who started to use TraceTogether without knowing its purpose of tracking one’s location, shared that she downloaded the app simply because she trusted the government, and thus trusted that the app was good for them.

Our results also revealed that the sense of community elevated by TraceTogether use positively influenced users’ benefit appraisal, even though it did not reduce privacy concerns. This finding is in line with research [61], which found that in largely collectivist societies (eg, China), there is a greater acceptance of contact tracing apps compared with more individualistic societies (eg, the United States, Germany), and people are generally willing to accept these technologies, especially if they are considered effective. In our study, the results might be because Singaporeans accept the spirit of the “greater good” underlying the contact tracing measure in the pandemic situation in the community, supporting the extant literature. A prior study compared Singapore and Switzerland and attributed the former’s higher rates of acceptance and adoption to its norms that prioritize the “interests of the community” [46]. Interestingly, privacy concerns regarding the use of TraceTogether persisted even when a sense of community was fostered. The follow-up interviews revealed these findings. For instance, our interviews indicated that considering fellow community members while participating in the contact tracing program motivated individuals to endure the discomfort of sacrificing privacy for the greater good rather than directly alleviating the concerns themselves. This aligns with prior studies on the role of sense of community in coping behaviors; they showed that a stronger sense of community motivates individuals to focus on problem-solving coping behaviors [62,63]. By extending these findings to our study, it can be inferred that strengthening the sense of community compels individuals to concentrate on addressing the larger community issue while accepting potential risks associated with participating in the program to tackle this issue.

Our postinterview responses provided an in-depth understanding of privacy calculus in TraceTogether use by dissecting different aspects of privacy concerns [16] that TraceTogether users exhibited. The interviewees responded that they felt they were being surveyed through TraceTogether, which often required excessive personal information. However, interviewees were generally willing to bear surveillance and intrusion for the greater good of the community and their own safety. However, the results indicated that having a sense of control over information is an important factor in the privacy calculus in TraceTogether. Most interviewees also called for greater transparency in data collection and secondary use of data. They actively sought strategies to better control their personal information. For example, some participants indicated that they prefer to use the token rather than the app so that they can take the token only when TraceTogether checking-in is required.
Hence, our study results support the notion that privacy management is a dynamic process involving “selective control of access to the self or to one’s group” [64].

Another notable finding of our study was how the sense of community could elevate the benefits of contact tracing despite low trust in the government. This is important considering that norms-based approaches [65] are effective in modifying health behaviors in diverse contexts. Existing research has identified several moderating factors that influence users’ intention to use contact tracing apps. For instance, demographic factors, such as gender, racial and ethnic identity, and education level, are significant moderators of contact tracing app design on people’s intention to install apps [66]. Separately, another study found that self-efficacy was a significant moderator of trust and intention to use contact tracing apps [67]. Although these findings are important for technology developers, they view technology use and acceptance through a largely technocentric and individualistic approach, even though public health scholars [68] have argued for the need to consider societal factors as driving factors of adoption, as evidenced by research [69] on how norms could drive app use, even among marginalized communities.

Practical Implications

Our study offers valuable insights into the policy and design considerations concerning digital contact tracing programs. The interview results revealed that, although Singaporeans are willing to accept a certain level of surveillance for preventive measures via contact tracing, issues arise when data use lacks transparency, potentially causing individuals to reconsider their positive intentions to participate in the greater good. Moreover, because digital contact tracing was made mandatory, they also felt that they had no control over their information, which may overturn their privacy calculus, and thus opted out from participating in the program if it became optional. The results suggest that for the successful implementation of the contact tracing program, it is crucial to implement policy measures to ensure clear and transparent communication regarding the use of data collected from contact tracing practices. In addition, providing as many control options as possible would encourage users to maintain active participation in the contact tracing program.

The moderating effect of a sense of community prompted by contact tracing has valuable practical implications. The interaction result indicated that the negative impact of distrust in government on contact tracing program participation can be overcome by promoting the “social good” nature of the contact tracing program via campaign or application design. According to our study results, the campaign for TraceTogether that framed contact tracing as a community effort and responsibility, with the campaign slogan being “Protect your community,” might have enhanced community participation, especially among those with lower trust in the government. If, unfortunately, another pandemic emerges in the future, and a similar program is needed, the contact tracing program can actively incorporate social- and community-related features. For example, in the early stage of implementation, the application can be promoted through social networks, allowing users to invite and encourage their family and friends to opt for the program.

Additionally, the contact tracing app interface can be designed to cue a sense of community when using it. Illustrations of the family and community for the application interface design, as TraceTogether has done, would be an effective strategy to enhance participation in the program. TraceTogether also displays the number of Bluetooth signals shared by others. However, such information can be framed as emphasizing collective efforts for the greater good of the community. For example, information can be shown through a visual illustration of people forming networks.

Limitation and Future Research Direction

This study had some limitations. Although the longitudinal survey design allowed us to establish the temporal sequencing of the variables of interest, the dependent variable (i.e., behavioral intention) was assessed in the follow-up survey without obtaining the baseline measurement. Hence, we cannot fully leverage the longitudinal nature of the data to establish the direction of the association or causal relationships between the variables. Future studies may use longitudinal data with repeated measures of the variables (two or more time points) to address them. Additionally, the data provide limited potential for generalizability. Survey participants with a higher socioeconomic status (e.g., education) were overrepresented. In addition, the use of self-reported data in surveys may pose the risk of social desirability and recall biases. Finally, we acknowledge that the mandatory use of TraceTogether between waves 1 and 2 of our survey might have an impact on the use intention of participants, as existing research [70] on a Singapore sample showed high adoption due to the mandatory use of the app for entering public venues. To circumvent this, we modified our items such that they captured participants’ volitional use in contexts when TraceTogether became voluntary, rather than capturing actual use behavior as the dependent variable due to social desirability bias. Despite the above limitations, we are confident that this study has presented valid evidence of the relationships among trust, privacy concerns, perceived benefits, and TraceTogether use intentions.

Conflicts of Interest

None declared.

References


### Abbreviations

- **AVE**: average variance extracted
- **SEM**: structural equation modeling
Patient-Centered Chronic Wound Care Mobile Apps: Systematic Identification, Analysis, and Assessment

Tassilo Dege¹,*, MSc, MD; Bernadette Glatzel¹,*, MSc, MD; Franziska Grü¹, MD; Simon Goller¹, MD; Caroline Glatzel¹, MD; Matthias Goebeler¹, MD; Astrid Schmieder¹, MD

¹ University of Wuerzburg, Germany

*these authors contributed equally

Corresponding Author:
Tassilo Dege, MSc, MD

Abstract

Background: The prevalence of chronic wounds is predicted to increase within the aging populations in industrialized countries. Patients experience significant distress due to pain, wound secretions, and the resulting immobilization. As the number of wounds continues to rise, their adequate care becomes increasingly costly in terms of health care resources worldwide. eHealth support systems are being increasingly integrated into patient care. However, to date, no systematic analysis of such apps for chronic wounds has been published.

Objective: The aims of this study were to systematically identify and subjectively assess publicly available German- or English-language mobile apps for patients with chronic wounds, with quality assessments performed by both patients and physicians.

Methods: Two reviewers independently conducted a systematic search and assessment of German- or English-language mobile apps for patients with chronic wounds that were available in the Google Play Store and Apple App Store from April 2022 to May 2022. In total, 3 apps met the inclusion and exclusion criteria and were reviewed independently by 10 physicians using the German Mobile App Rating Scale (MARS) and the System Usability Scale (SUS). The app with the highest mean MARS score was subsequently reviewed by 11 patients with chronic wounds using the German user version of the MARS (uMARS) and the SUS. Additionally, Affinity for Technology Interaction (ATI) scale scores were collected from both patients and physicians.

Results: This study assessed mobile apps for patients with chronic wounds that were selected from a pool of 118 identified apps. Of the 73 apps available in both app stores, 10 were patient oriented. After excluding apps with advertisements or costs, 3 apps were evaluated by 10 physicians. Mean MARS scores ranged from 2.64 (SD 0.65) to 3.88 (SD 0.65) out of 5, and mean SUS scores ranged from 50.75 (SD 27) to 80.5 (SD 17.7) out of 100. WUND APP received the highest mean MARS score (mean 3.88, SD 0.65 out of 5) among physicians. Hence, it was subsequently assessed by 11 patients and achieved a similar rating (uMARS score: mean 3.89, SD 0.4 out of 5). Technical affinity, as measured with the ATI scale, was slightly lower in patients (score: mean 3.62, SD 1.35 out of 6) compared to physicians (score: mean 3.88, SD 1.03 out 6).

Conclusions: The quality ratings from physicians and patients were comparable and indicated mediocre app quality. Technical affinity, as assessed by using the ATI scale, was slightly lower for patients. Adequate apps for patients with chronic wounds remain limited, emphasizing the need for improved app development to meet patient needs. The ATI scale proved valuable for assessing technical affinity among different user groups.

(JMIR Mhealth Uhealth 2024;12:e51592) doi:10.2196/51592

KEYWORDS
chronic wounds; chronic leg ulcers; mobile applications; evaluation; mental health; Mobile Application Rating Scale; System Usability Scale; affinity for technology interaction; ATI; teledermatology; disease management; health app; skin; eHealth; telemedicine; mHealth; mobile health; app; apps; applications; quality; rating; wound; wounds; chronic; ulcer; ulcers; sore; sores; dermatology; chronic wound

Introduction

Background

Aging societies in industrialized nations are experiencing an increasing prevalence of chronic wounds, resulting in growing challenges in patient care. The high costs of therapy and personnel often limit proper treatment. Chronic wounds significantly impact patients’ quality of life, requiring intensive therapy multiple times per week, along with regular medical checkups. Moreover, they have substantial economic
implications, including hospitalizations, personnel costs (e.g., wound care nurses, home health care services, physicians, wound managers, day clinics, etc), material costs, and management (e.g., medical transportation). Purwins et al [1] identified hospitalizations, nursing staff, and material costs as the main contributors to the overall expenses.

Mobile health (mHealth) apps hold promise for bridging gaps in health care. However, a lack of evidence exists for the effectiveness of available mHealth apps [2], and high-quality trials are needed to examine their effects [3]. So far, only reviews of the use of mHealth apps without systematic searches and evaluations have been performed [4]. To address this gap in evidence for patients with chronic wounds, a systematic analysis is necessary to gather reliable data in this area.

A study by Svendsen et al [5] demonstrated that patient-centered smartphone apps can significantly improve treatment adherence in cohorts of patients with psoriasis and patients with rheumatic diseases. Participants expressed favorable views toward medical apps, indicating a willingness to use such apps if available.

To subjectively assess app quality, the Mobile App Rating Scale (MARS) [6] was developed, which evaluates engagement, functionality, aesthetics, and information. Additionally, the System Usability Scale (SUS) is a 10-item questionnaire for assessing the usability of a system. It has been effectively applied to evaluate mobile apps related to dementia, depression, pediatric obesity, and smoking cessation [7]. Moreover, the Affinity for Technology Interaction (ATI) scale [8] provides an easy and reliable means to quantify an individual’s technology affinity. An aim of this study was to conduct a focused analysis on the core functionalities and core features of the included apps; therefore, we excluded apps with advertisements in order to prevent disruptions in users’ experiences and advertisements’ effects on usability [9]. Advertisements can be confounding variables and make it difficult to compare apps’ performance or users’ experiences. Advertisements can frustrate or annoy users and reduce the clarity and intuitiveness of an app’s interface [10]. We also excluded paid apps to prevent bias toward users with financial means, maintain analysis inclusivity, and ensure that preferences align with intrinsic app quality rather than financial considerations, thereby enhancing the analysis’ fairness and validity.

**Aim of This Study**

The primary objectives of this study were to identify and evaluate publicly available smartphone apps designed for patients with chronic wounds. The assessment aimed to provide subjective quality ratings for these apps, while also collecting data on the technical affinity of this specific patient group. To date, a systematic review and assessment of smartphone apps tailored for patients with chronic wounds has not been conducted.

**Methods**

**Ethical Considerations**

This study was conducted in accordance with the Declaration of Helsinki, and ethical approval was waived by the local ethics committee of the University of Würzburg.

**App Selection**

A systematic search of the German Google Play Store and the Apple App Store was performed from April 2022 to May 2022. The search terms used were “wound,” “pressure ulcer,” “ulcus,” “Wunde,” “Wunddokumentation,” and “wound documentation.” Two independent reviewers searched each app store. The inclusion criteria were that apps had to (1) be available in both app stores, (2) be available in English or German, and (3) be specifically designed for patients. Apps that were not available free of charge and apps that contained advertisements were excluded.

The following information, when available in the app stores and on the associated app websites, was collected: app name, target group (eg, patients and medical personnel), cost, platform, advertisements, features, and search term used to identify the app.

**Evaluation of App Quality**

The MARS [11] was developed for professional raters to evaluate mobile apps, and it is a validated and reliable scale. The user version of the MARS (uMARS) [12] was designed for users to evaluate the quality of mHealth apps. Both scales are based on a 5-point Likert scale for the following four sections: “Engagement,” “Aesthetics,” “Functionality,” and “Information.” Additionally, there is a “Subjective” section. Studies using the MARS have already been performed for several chronic diseases and apps related to breast cancer. The uMARS [12] has been broadly applied to evaluate apps for rheumatic diseases, weight loss, nutrition tracking, and menstrual tracking. The SUS is a simple, 10-item attitude Likert scale that provides an overview of subjective assessments of usability [13].

Prior to the evaluation, suitable apps were selected based on the inclusion and exclusion criteria, resulting in a total of 3 apps. The quality of these three apps was then evaluated by 10 physicians using the MARS and SUS. Before the assessment, the physicians first watched a short training video that explained the MARS and were then asked to use the apps for more than 10 minutes.

Finally, the best app—the one with the highest mean MARS score among the physicians—was evaluated by 11 patients with chronic wounds using the uMARS and SUS.

**Evaluation of Technical Affinity**

Many studies that evaluate the quality of mobile apps lack information on cohorts’ technical affinity, which is necessary to assess and interpret the results. Therefore, the ATI scale [8] was used to gather information on physicians’ and patients’ technical affinity. The ATI scale was designed to quantify a tendency to actively engage in intensive technology interaction or a tendency to avoid technology interaction. In both the patient group and the physician group, ATI scale scores were collected. A Pearson correlation was used to correlate the ages of patients and ATI scale scores.
Comparative Analysis of Patients’ and Physicians’ Data

The MARS results represented the physicians’ evaluations, as stated in the Evaluation of App Quality section, whereas uMARS results corresponded to the patients’ evaluations. As a next step, after normalization, the Mann-Whitney U test was used to analyze whether there was a significant difference between MARS and uMARS scores and between patients’ and physicians’ SUS and ATI scale median scores. A comparison of the five subcategories within the MARS was performed. \( P \) values of < .05 were considered significant. The data analysis was performed with SPSS 23 (IBM Corp).

Results

App Selection

A total of 118 apps were identified—95 in the Apple App Store and 96 in the Google Play Store—of which 73 were available in both app stores (Figure 1). Of these 73 apps, 10 were specifically designed for patients. Of these 10 apps, 1 contained advertising, 4 were not free of charge, and 3 met both exclusion criteria, resulting in a total of 7 apps that were excluded from further analysis.

Figure 1. Flowchart illustrating the screening process for identifying suitable mobile apps.
Evaluation of App Quality

A total of 3 apps—WoundEducation, APD Skin Monitoring, and WUND APP—met all inclusion and exclusion criteria and were evaluated by 10 physicians. WoundEducation provides an overview of different wound types and treatment options. APD Skin Monitoring includes an automated area calculation function to quantify wound area. WUND APP offers advice and information on chronic wounds and contains a diary function to track patient-related outcomes, such as pain or the level of wound secretion.

The mean MARS scores (and SDs) of the physicians, including subcategory scores and SUS scores, are provided in Table 1.

<table>
<thead>
<tr>
<th>Mobile app name</th>
<th>MARS score, mean (SD)</th>
<th>Engagement score, mean (SD)</th>
<th>Functionality score, mean (SD)</th>
<th>Aesthetics score, mean (SD)</th>
<th>Information score, mean (SD)</th>
<th>Psychotherapy score, mean (SD)</th>
<th>SUS score (%), mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WUND APP</td>
<td>3.88 (0.65)</td>
<td>3.36 (0.89)</td>
<td>4.38 (0.66)</td>
<td>4.13 (0.76)</td>
<td>3.67 (0.56)</td>
<td>2.98 (0.36)</td>
<td>80.5 (17.7)</td>
</tr>
<tr>
<td>WoundEducation</td>
<td>3.01 (0.5)</td>
<td>2.14 (0.44)</td>
<td>3.93 (0.99)</td>
<td>2.73 (0.75)</td>
<td>3.25 (0.58)</td>
<td>2.57 (0.46)</td>
<td>72.75 (18.5)</td>
</tr>
<tr>
<td>APD Skin Monitoring</td>
<td>2.64 (0.65)</td>
<td>2.36 (0.56)</td>
<td>2.65 (0.83)</td>
<td>2.77 (0.69)</td>
<td>2.79 (0.82)</td>
<td>2.47 (0.63)</td>
<td>50.75 (27)</td>
</tr>
</tbody>
</table>

WUND APP had the highest mean MARS score (mean 3.88, SD 0.65 out of 5) and was subsequently analyzed by 11 patients with chronic wounds. WUND APP had a similar mean uMARS score (mean 3.89, SD 0.4 out of 5) when analyzed by the patients.

Evaluation of Technical Affinity

The ATI scale scores ranged from 1.56 to 6 out of 6 for patients and from 2.11 to 5.67 out of 6 for physicians (Figure 2). An ATI scale score of >3 indicates average technology affinity, and a score of >4 indicates high technology affinity [8]. Mean ATI scale scores were slightly lower for patients (mean 3.62, SD 1.35 out of 6) than those for physicians (mean 3.88, SD 1.03 out of 6), but the difference was not statistically significant (P=.43). Patients ranged in age from 28 to 70 years; 7 were male, and 4 were female. Physicians ranged in age from 26 to 43 years; 4 were male, and 6 were female. The Pearson correlation coefficient between ATI scale scores and the ages of patients was −0.706 (P=.02), indicating a strong negative correlation.
Comparative Analysis of Patients’ and Physicians’ Data

MARS and uMARS scores were almost identical. The subcategory “Information” was rated worse by physicians, while the subcategory “Functionality” was rated worse by patients (Figure 3). For the subcategories “Aesthetics” and “Engagement,” no differences between physicians and patients were found. No statistically significant difference (Mann-Whitney U test) was found between MARS and uMARS scores ($P=.76$) or between patients’ and physicians’ SUS scores ($P=.39$; Multimedia Appendix 1).
**Discussion**

**Principal Findings**

This study represents the first systematic examination that aimed to identify and assess smartphone apps specifically designed for patients with chronic wounds. The quality of these apps was evaluated by independent professional reviewers (physicians) and patients using well-validated scoring systems—the MARS, uMARS, and SUS. Using subjective measures in the assessment of eHealth is relevant and essential, and it is crucial to recognize this fact. Agreement among the assessors, even for subjective measures, is a good sign of the reliability of the assessment [14]. The overall findings indicate that the apps received moderate ratings. Among all apps evaluated, WUND APP achieved the highest mean MARS score (mean 3.88, SD 0.65) when assessed by physicians. Similarly, it received a mean uMARS score of 3.89 (SD 0.4) when evaluated by patients. Given that the preselection was exclusively conducted by physicians, patients’ assessments might have differed.

WUND APP includes a diary function for documenting wound photos and patient-related outcomes, such as pain or wound secretion. Furthermore, it has a reminder function for necessities such as physician appointments and dressing changes. Its user interface is clear, and the app is user-friendly. Consequently,
physicians assigned high scores for the subcategories “Engagement” (mean 3.36, SD 0.89), “Functionality” (mean 4.38, SD 0.66), and “Aesthetics” (mean 4.13, SD 0.76). Additionally, the app provides information on the causes, diagnostics, and treatments of the different wound types, resulting in a mean MARS score of 3.67 (SD 0.56) for the subcategory “Information.”

WoundEducation presents information in a straightforward structure that resembles an article with embedded links, delivers medical information, and illustrates various wound types through examples. The app received high ratings for overall functionality (mean 3.93, SD 0.99) and information provision (mean 3.25, SD 0.58). However, it demonstrated limited interactivity, as reflected by having the lowest score for the “Engagement” subcategory (mean 2.14, SD 0.44). Additionally, its aesthetics (mean 2.73, SD 0.75) were rated the lowest among the three apps.

APD Skin Monitoring [15] uses a coin as a reference for calculating the area of a wound, proving particularly useful in assessing wound progression and healing status regardless of whether the wounds are irregularly shaped. In addition, the coloration of wounds can be analyzed and tracked over time. However, the app received the lowest ratings for the subcategories “Functionality” (mean 2.65, SD 0.83) and “Engagement” (mean 2.36, SD 0.56).

Our findings revealed that a limited number of wound apps were specifically tailored for patients, accounting for only 14% (1073) of all wound apps that were available in both app stores. Overall, both physicians and patients rated WUND APP similarly; however, although not statistically significant, it is noteworthy that differences emerged in 2 subcategories (“Information”: P=.97; “Functionality”: P=.56). Physicians rated the information content lower, which could be attributed to their expert medical knowledge. As professionals, they may have had higher expectations regarding the app’s information content and may have found certain aspects less informative. In a recent study, it was shown that patients trusted recommendations and reviews from medical organizations and health care professionals when selecting apps [16]; their motivation to continue using apps was driven by features that supported goal setting and tracking, data sharing, decision-making, and empowerment.

On the other hand, patients rated the functionality of WUND APP lower, which was possibly due to their limited involvement in the app’s development process. A systematic review reported that health care professionals were engaged in the development process for only 35% of the 7 analyzed apps that were specifically designed for patients with rheumatoid arthritis [17]. Another systematic review revealed that patients were only engaged in the development process for 15% of the 32 analyzed apps that were designed for individuals with rheumatic and musculoskeletal diseases [18]. The inclusion of patients in the development process of future wound apps could help to ensure that the apps meet their specific needs and preferences. Our study highlights the scarcity of wound apps designed explicitly for patients and the importance of involving patients in the app development process. Tailoring apps to meet patients’ specific requirements and involving them in the design process would likely result in improved app functionality and overall user satisfaction.

To the best of our knowledge, this study represents the first systematic study to collect data on the technical affinity of patients with wounds. Surprisingly, no statistically significant difference in technical affinity was observed between wound care patients and physicians, even with the inclusion of patients aged up to 70 years (P=.43). Nevertheless, future studies and app development projects should aim to include older patients and comprehensively assess and address their specific needs. The low adoption and use of mHealth apps among older patients are frequently attributed to inadequate designs [19].

In 2019, Germany introduced a digital health app (DiGA) directory that includes scientifically validated apps. Physicians can prescribe DiGAs in a manner similar to how they prescribe medications [20]. However, to date, no DiGAs specifically tailored for patients with chronic wounds are available in the directory. It is worth noting that physicians with a higher technical affinity and those who are female hold significantly more positive attitudes toward DiGAs [21].

The absence of wound care apps highlights the unmet potential for innovative digital solutions to address the needs of patients with chronic wounds. The inclusion of validated and effective wound care apps in the DiGA directory could significantly improve patient outcomes and health care management in this specific area. It is imperative for future app development initiatives to focus on developing and validating apps that cater to the unique requirements of patients with chronic wounds, to provide them with accessible and effective digital health care resources.

Multiple attempts have been made to integrate mHealth apps into wound care [22], including apps for wound care measurements [23], wound care dressing decision support systems [24], and home-based self-management systems [25]. An Australian study assessed an artificial intelligence app for wound assessment, involving 166 patients in the standard group and 124 in the intervention group. The intervention group demonstrated significantly improved wound documentation, along with positive outcomes such as enhanced patient adherence, efficient digital care provision, and substantial reductions in wound size [26].

A cultural shift toward greater technology affinity has been accelerated by the COVID-19 pandemic [27]. Patients are increasingly using mobile apps when they perceive clear benefits to their use, such as reducing social contact during the COVID-19 pandemic. For patients with chronic wounds, these benefits may include using an app as a diary, while for others, apps may offer the advantage of saving time and transportation costs through telemedicine services. Telemedical approaches could significantly alter and improve the current wound care landscape, bearing the potential to improve the efficiency, accuracy, and accessibility of both diagnoses and treatments. Telemedicine in chronic wound management was shown to be noninferior to conventional standard care in a systematic review and meta-analysis [28]. The earlier diagnosis of complications,
such as wound infections, can reduce the use of antibiotics and lower health care costs by preventing hospital stays.

**Limitations**

Due to the limited number of apps that met our strict inclusion and exclusion criteria, only a small subset could be analyzed in this study. The selection of apps was conducted within a relatively short time frame, which might have further restricted the available options for evaluation.

The nonsignificant differences between physicians’ and patients’ ratings in this study may be attributed to the small sample of only 11 patients and 10 physicians, which resulted in limited statistical power. Therefore, we emphasize the importance of cautious interpretation and the consideration of larger sample sizes for future research.

Another factor that influenced this study’s results was the inclusion of patients who agreed to evaluate **WUND APP**. This approach potentially introduced selection bias, as our patient sample may not represent the true technical affinity of all patients with chronic wounds. The overall technical affinity of a larger and more diverse patient population might be lower than what was observed in this study. Future studies with larger patient cohorts might provide a more comprehensive understanding of the technical affinity and usability experience of patients with chronic wounds who use mHealth apps. Further, the assessment relied on a qualitative survey—a method that may be susceptible to various biases. However, to truly ascertain the effectiveness of **WUND APP**, a randomized controlled trial is needed.

**Conclusions**

Patient involvement is crucial in app development. By involving all stakeholders, including physicians, wound care experts, and patients, throughout the development process, apps can be tailored to meet the specific needs and preferences of the end users, resulting in increased user satisfaction and improved health outcomes.

The validated ATI scale proved to be a valuable tool for evaluating an individual’s technical affinity. In future studies and app evaluations, technical affinity should be determined to generalize outcomes to specific patient and consumer cohorts.

**Acknowledgments**

TD is supported by the Clinician Scientist Program (project number: Z-2/CSP-25) of the Interdisciplinary Center of Clinical Research (IZKF) of the Medical Faculty, University of Würzburg.

**Conflicts of Interest**

None declared.

**Multimedia Appendix 1**

Patient and physician characteristics (age and sex) and the Affinity for Technology Interaction scale, System Usability Scale, Mobile App Rating Scale (MARS), and user version of the MARS scores (including subcategory scores) of all 11 patients and 10 physicians who participated.

**References**


uMARS: user version of the Mobile App Rating Scale

Edited by L Buis; submitted 04.08.23; peer-reviewed by C Jacob, M Zhou; revised version received 17.12.23; accepted 25.01.24; published 21.03.24.

Please cite as:
Dege T, Glatzel B, Borst V, Grän F, Goller S, Glatzel C, Goebeler M, Schmieder A
Patient-Centered Chronic Wound Care Mobile Apps: Systematic Identification, Analysis, and Assessment
JMIR Mhealth Uhealth 2024;12:e51592
URL: https://mhealth.jmir.org/2024/1/e51592
doi:10.2196/51592

© Tassilo Dege, Bernadette Glatzel, Vanessa Borst, Franziska Grän, Simon Goller, Caroline Glatzel, Matthias Goebeler, Astrid Schmieder. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 21.3.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Mobile Apps for Common Noncommunicable Disease Management: Systematic Search in App Stores and Evaluation Using the Mobile App Rating Scale

Khang Jin Cheah1,*, MSc; Zahara Abdul Manaf1,*, PhD; Arimi Fitri Mat Ludin1,*, PhD; Nurul Huda Razalli1,*, PhD; Norfilza Mohd Mokhtar2,*, PhD; Sawal Hamid Md Ali3,*, PhD

1, 2, 3

*all authors contributed equally

Corresponding Author:
Zahara Abdul Manaf, PhD

Abstract

Background: The success of mobile apps in improving the lifestyle of patients with noncommunicable diseases through self-management interventions is contingent upon the emerging growth in this field. While users of mobile health (mHealth) apps continue to grow in number, little is known about the quality of available apps that provide self-management for common noncommunicable diseases such as diabetes, hypertension, and obesity.

Objective: We aimed to investigate the availability, characteristics, and quality of mHealth apps for common noncommunicable disease health management that included dietary aspects (based on the developer’s description), as well as their features for promoting health outcomes and self-monitoring.

Methods: A systematic search of English-language apps on the Google Play Store (Google LLC) and Apple App Store (Apple Inc) was conducted between August 7, 2022, and September 13, 2022. The search terms used included weight management, obesity, diabetes, hypertension, cardiovascular diseases, stroke, and diet. The selected mHealth apps’ titles and content were screened based on the description that was provided. Apps that were not designed with self-management features were excluded. We analyzed the mHealth apps by category and whether they involved health care professionals, were based on scientific testing, and had self-monitoring features. A validated and multidimensional tool, the Mobile App Rating Scale (MARS), was used to evaluate each mHealth app’s quality based on a 5-point Likert scale from 1 (inadequate) to 5 (excellent).

Results: Overall, 42 apps were identified. Diabetes-specific mHealth apps accounted for 7% (n=3) of the market, hypertension apps for 12% (n=5), and general noncommunicable disease management apps for 21% (n=9). About 38% (n=16) of the apps were for managing chronic diseases, while 74% (n=31) were for weight management. Self-management features such as weight tracking, BMI calculators, diet tracking, and fluid intake tracking were seen in 86% (n=36) of the apps. Most mHealth apps (n=37, 88%) did not indicate whether there was involvement of health care professionals in app development. Additionally, none of the apps reported scientific evidence demonstrating their efficacy in managing health. The overall mean MARS score was 3.2 of 5, with a range of 2.0 to 4.1. Functionality was the best-rated category (mean score 3.9, SD 0.5), followed by aesthetics (mean score 3.2, SD 0.9), information (mean score 3.1, SD 0.7), and engagement (mean score 2.9, SD 0.6).

Conclusions: The quality of mHealth apps for managing chronic diseases was heterogeneous, with roughly half of them falling short of acceptable standards for both quality and content. The majority of apps contained scant information about scientific evidence and the developer’s history. To increase user confidence and accomplish desired health outcomes, mHealth apps should be optimized with the help of health care professionals. Future studies on mHealth content analysis should focus on other diseases as well.

(JMIR Mhealth Uhealth 2024;12:e49055) doi:10.2196/49055

KEYWORDS
mHealth apps; mobile health; health apps; chronic diseases; self-management; app quality; apps; app; application; applications; quality; MARS; Mobile App Rating Scale; mHealth; chronic; review methods; review methodology; review of apps
**Introduction**

Globally, noncommunicable diseases (NCDs) account for 74% of all deaths yearly (41 million people), which highlights the global health burden [1]. In Malaysia, the latest National Health and Morbidity Survey reported that two-thirds of the Malaysian population have at least 1 of 3 common NCDs, namely diabetes, hypertension, or hypercholesterolemia, and about half of the population (50.1%) are overweight or obese [2]. There are barriers to care access, delivery, and self-management for the management of NCDs, such as being unable to visit a clinic in a timely manner and long consultation waiting times [3], prompting policy makers to improve the health care system. The introduction of mobile health (mHealth) technology presents an opportunity for patient self-monitoring, helping health care providers personalize the management of patients [4] and increase cost-effectiveness throughout health management [5].

According to a definition by the World Health Organization, mHealth is the use of mobile and wireless devices such as mobile phones, tablets, and personal digital assistants to support health care management [6]. In the rapidly growing mHealth app market, the presence of these apps could facilitate the health care management system. A recent review of the literature revealed that researchers have recognized that mHealth could be an effective tool in chronic disease management [4] and improve patients’ self-management behavior [7]. A growing body of research demonstrates the health benefits of mHealth interventions for patients with NCDs in terms of enhancing patient self-monitoring and health outcomes in NCDs such as type 2 diabetes [8], obesity [9], and cardiovascular diseases [10].

Self-management is crucial in the daily management of chronic diseases to improve quality of life and reduce management costs [11]. However, poor self-management among patients with chronic disease has been observed [12]. Technologies such as mHealth have the potential capacity to empower patients requiring support in their self-management efforts. A review by Cruz-Ramos et al [10] demonstrated that many mHealth apps for cardiovascular diseases support self-management features such as medical advice, reminders, and self-monitoring notifications [10]. Moreover, research has also found that self-monitoring of weight and dietary intake is associated with positive outcomes for weight loss [13]. As the key to person-centered care, mHealth apps enhance self-management for chronic conditions by providing personalized goal setting, active reminders, social interaction, and support [14]. Hence, mHealth apps could help build decision-support systems that bridge the gap between self-management and conventional health care management.

As of 2022, there were nearly 2.67 million mobile apps available on the Google Play Store. Of these, more than 130,000 apps were health care or health and fitness apps [15]. The number of mHealth apps available on the Google Play Store and Apple App Store continues to grow [16-18]. Globally, it was estimated that 6.6 billion individuals own a smartphone, and the number is expected to grow to 7.7 billion by the year 2027 [19], which allows mHealth technology to be more accessible to individuals.

However, caution must be taken regarding this growth, as the evidence related to its efficacy and benefits for chronic health disease management is not well identified.

Previous research on the perception and usability of mHealth apps for NCD management has demonstrated a growing interest in user-centric health-tracking mobile apps among the population with chronic illness [20,21]. An empirical study conducted to predict patients’ intentions to continue using mHealth services as part of self-managing their chronic conditions revealed that the participants had high intentions to continue the use of mHealth services [22]. The use of mHealth apps is highly encouraged, as it has been linked to higher rates of health-promoting behavior among people with chronic medical conditions [23]. Self-management using mHealth apps could be a part of health management, as people living with NCDs have the autonomy to take responsibility for their health. Relevant content analysis studies have been conducted in different geographical areas [24-26], but limited studies have been conducted among Southeast Asian countries. As such, there is an emerging need to bridge this research gap by initiating more content analysis studies in Malaysia to contribute to more holistic development of mHealth apps.

This paper provides a review of the current landscape of mHealth apps, with an emphasis on common chronic disease management. Understanding trends in mHealth apps and their relevant features will benefit users in developing informed decisions, as well as help health care providers improve the quality of mHealth. To make better-informed decisions, the reliability of currently existing mHealth apps should be explored. This study aims to describe mHealth apps available in conventional app stores for common NCDs, determine their health categories, explore the features they focus on, identify neglected areas, and evaluate their quality using the Mobile App Rating Scale (MARS) assessment tool.

**Methods**

This review involved a systematic search of apps available in mobile phone app stores. The protocol adhered to the 5-step framework outlined by Arksey and O’Malley [27], which includes (1) identifying the research question; (2) identifying relevant apps; (3) selecting apps; (4) charting the data; and (5) collating, summarizing, and reporting the results.

**Identifying the Research Question**

This review aimed to answer the following questions: “What are the available mHealth apps in the Malaysian market for common NCDs?” “What are the app features available in the mHealth apps for NCDs in Malaysia?” and “What is the quality of the mHealth apps for NCDs in Malaysia?”

**Identifying Relevant Apps**

The search was conducted from July 7, 2022, to August 14, 2022, on the Apple App Store (Apple Inc) and Google Play Store (Google LLC) using the search terms weight loss, obesity, diabetes, hypertension and cardiovascular diseases, stroke, weight management, and diet. The search terms were identified using appropriate Medical Subject Headings terms as well as the free text of keywords. The selected NCDs were chosen...
because they are among the most prevalent chronic diseases in Malaysia and around the world.

**Selecting Apps**

To be eligible for inclusion in this review, the title and content of the identified mHealth apps were screened and filtered based on the descriptions provided by the app developer. An app met the inclusion criteria if it (1) was developed in English; (2) had self-monitoring feature(s); and (3) was developed for chronic disease management, namely obesity, type 2 diabetes, hypertension, and cardiovascular disease. An app was excluded based on the following criteria: it was a heart rate tracker, exercise tracker, or game; or it was a medicine-delivery, appointment-based, recipe-sharing, or research study app.

**Charting the Data**

The data were extracted based on feature categories, which included (1) type of health management, (2) number of downloads, (3) country of the developers, and (4) app features. A data extraction table with the mHealth apps’ basic information, such as country of origin, name of app developer, and number of downloads, is shown in Table 1. The details of the mHealth apps’ features and characteristics are outlined in Multimedia Appendix 1.
<table>
<thead>
<tr>
<th>Name of app</th>
<th>Number of downloads</th>
<th>Availability of in-app purchase</th>
<th>Cost of in-app purchase (US $)</th>
<th>Country of app developer</th>
<th>App developer</th>
<th>Operating system</th>
</tr>
</thead>
<tbody>
<tr>
<td>MyFitness Pal: Calorie Counter</td>
<td>100,000,000</td>
<td>Yes</td>
<td>0.76-0.70</td>
<td>US</td>
<td>MyFitnessPal, Inc</td>
<td>Android</td>
</tr>
<tr>
<td>Health &amp; fitness tracker with calorie counter</td>
<td>5,000,000</td>
<td>Yes</td>
<td>0.65-2.90</td>
<td>India</td>
<td>DROID INFINITY</td>
<td>Android</td>
</tr>
<tr>
<td>Lifesum: Healthy eating &amp; diet</td>
<td>10,000,000</td>
<td>Yes</td>
<td>3.95-50.30</td>
<td>Sweden</td>
<td>Lifesum</td>
<td>Android</td>
</tr>
<tr>
<td>Withings Health Mate</td>
<td>1,000,000</td>
<td>No</td>
<td>0.00</td>
<td>France</td>
<td>Withings</td>
<td>Android</td>
</tr>
<tr>
<td>Fitbit</td>
<td>50,000,000</td>
<td>Yes</td>
<td>7.20-290.70</td>
<td>US</td>
<td>Fitbit LLC</td>
<td>Android</td>
</tr>
<tr>
<td>HealthifyMe</td>
<td>10,000,000</td>
<td>Yes</td>
<td>7.65-72.10</td>
<td>Singapore</td>
<td>HealthifyMe</td>
<td>Android</td>
</tr>
<tr>
<td>Noom: Weight loss</td>
<td>10,000,000</td>
<td>Yes</td>
<td>0.76-142.10</td>
<td>US</td>
<td>Noom Inc</td>
<td>Android</td>
</tr>
<tr>
<td>Personal Health Monitor</td>
<td>100,000</td>
<td>No</td>
<td>0.00</td>
<td>Ukraine</td>
<td>Extrawest</td>
<td>Android</td>
</tr>
<tr>
<td>Life Extend: Healthy Habits</td>
<td>100,000</td>
<td>Yes</td>
<td>3.10-203.30</td>
<td>US</td>
<td>LifeOmic</td>
<td>Android</td>
</tr>
<tr>
<td>Qardio Heart Health</td>
<td>100,000</td>
<td>No</td>
<td>0.00</td>
<td>Canada</td>
<td>Qardio Inc</td>
<td>Android</td>
</tr>
<tr>
<td>FitTrack My-Health: Track Scale</td>
<td>100,000</td>
<td>Yes</td>
<td>9.61-92.90</td>
<td>US</td>
<td>Fittrack</td>
<td>Android</td>
</tr>
<tr>
<td>Calorie counter by lose it!</td>
<td>10,000,000</td>
<td>Yes</td>
<td>4.25-161.75</td>
<td>US</td>
<td>FitNow Inc</td>
<td>Android</td>
</tr>
<tr>
<td>One Drop: Better Health Today</td>
<td>1,000,000</td>
<td>Yes</td>
<td>20.10-20.35</td>
<td>US</td>
<td>One Drop</td>
<td>Android</td>
</tr>
<tr>
<td>Calorie Counter - MyNetDiary</td>
<td>1,000,000</td>
<td>Yes</td>
<td>3.75-60.10</td>
<td>US</td>
<td>MyNetDiary.com</td>
<td>Android</td>
</tr>
<tr>
<td>Healthi: Personal Weight Loss</td>
<td>500,000</td>
<td>Yes</td>
<td>1.30-52.45</td>
<td>Sunshine Health Studios</td>
<td></td>
<td>Android</td>
</tr>
<tr>
<td>Health Diet Foods Fitness Help</td>
<td>500,000</td>
<td>Yes</td>
<td>2.20-21.90</td>
<td>India</td>
<td>RecoveryBull.com</td>
<td>Android</td>
</tr>
<tr>
<td>Unimeal: Healthy Diet&amp;Workouts</td>
<td>100,000</td>
<td>Yes</td>
<td>5.50-95.10</td>
<td>Cyprus</td>
<td>Uniwell</td>
<td>Android</td>
</tr>
<tr>
<td>Health Click Away</td>
<td>10,000</td>
<td>Yes</td>
<td>3.10-21.85</td>
<td>US</td>
<td>HealthClickAway</td>
<td>Android</td>
</tr>
<tr>
<td>Possible-Nutrition Weight Loss</td>
<td>1,000,000</td>
<td>Yes</td>
<td>0.00</td>
<td>India</td>
<td>Truweight Wellness</td>
<td>Android</td>
</tr>
<tr>
<td>Health Club-Home workouts&amp; Fitness-calorie tracker</td>
<td>50,000</td>
<td>No</td>
<td>0.00</td>
<td>___</td>
<td>Health Club Group</td>
<td>Android</td>
</tr>
<tr>
<td>Smart Diet Planner weight loss</td>
<td>100,000</td>
<td>Yes</td>
<td>9.2-28.40</td>
<td>India</td>
<td>Appneurons Technologies Private Limited</td>
<td>Android</td>
</tr>
<tr>
<td>Heart Care Health &amp; Diet Tips</td>
<td>10,000</td>
<td>Yes</td>
<td>2.2-21.90</td>
<td>India</td>
<td>RecoveryBull.com</td>
<td>Android</td>
</tr>
<tr>
<td>Calorie Counter + (Nutracheck)</td>
<td>1,000,000</td>
<td>Yes</td>
<td>1.85-36.10</td>
<td>England</td>
<td>Nutracheck</td>
<td>Android</td>
</tr>
<tr>
<td>Health Mate - Calorie Counter &amp; Weight Loss App</td>
<td>500,000</td>
<td>Yes</td>
<td>0.90-5.69</td>
<td>India</td>
<td>PIXEL BYTES</td>
<td>Android</td>
</tr>
<tr>
<td>Doctor2u- One Stop Healthcare</td>
<td>500,000</td>
<td>Yes</td>
<td>2.20</td>
<td>Malaysia</td>
<td>BP Healthcare Group</td>
<td>Android</td>
</tr>
<tr>
<td>Name of app</td>
<td>Number of downloads</td>
<td>Availability of in-app purchase</td>
<td>Cost of in-app purchase (US $)</td>
<td>Country of app developer</td>
<td>App developer</td>
<td>Operating system</td>
</tr>
<tr>
<td>-------------</td>
<td>---------------------</td>
<td>---------------------------------</td>
<td>---------------------------------</td>
<td>--------------------------</td>
<td>---------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>BookDoc - Go Active Get reward</td>
<td>500,000</td>
<td>Yes</td>
<td>1.75-17.50</td>
<td>Malaysia</td>
<td>BookDoc</td>
<td>Android</td>
</tr>
<tr>
<td>Health Pal - Fitness, Weight loss coach, Pedometer</td>
<td>1,000,000</td>
<td>Yes</td>
<td>1.65</td>
<td>India</td>
<td>Digit Grove</td>
<td>Android</td>
</tr>
<tr>
<td>Creda - manage chronic condition</td>
<td>100,000</td>
<td>Yes</td>
<td>1.75-18.40</td>
<td>US</td>
<td>KnowYourMeds Inc</td>
<td>Android</td>
</tr>
<tr>
<td>Mhealth</td>
<td>—</td>
<td>No</td>
<td>0.00</td>
<td>—</td>
<td>mutifun LLC</td>
<td>iOS</td>
</tr>
<tr>
<td>my Mhealth</td>
<td>—</td>
<td>No</td>
<td>0.00</td>
<td>England</td>
<td>my mhealth</td>
<td>iOS</td>
</tr>
<tr>
<td>Zero: Fasting &amp; Health Tracker</td>
<td>—</td>
<td>Yes</td>
<td>10.10-69.95</td>
<td>US</td>
<td>Zero Longevity Science Inc</td>
<td>iOS</td>
</tr>
<tr>
<td>Lose Weight at Home in 30 Days</td>
<td>—</td>
<td>Yes</td>
<td>5.25-54.65</td>
<td>Hong Kong</td>
<td>ABishKING LIMITED</td>
<td>iOS</td>
</tr>
<tr>
<td>BodyFast Intermittent Fasting</td>
<td>—</td>
<td>Yes</td>
<td>3.95-61.2</td>
<td>Germany</td>
<td>BodyFast GmbH</td>
<td>iOS</td>
</tr>
<tr>
<td>BetterMe - Health Coaching</td>
<td>—</td>
<td>Yes</td>
<td>4.60-36.70</td>
<td>Cyprus</td>
<td>BetterMe Limited</td>
<td>iOS</td>
</tr>
<tr>
<td>Weight Loss Running by Slimkit</td>
<td>—</td>
<td>Yes</td>
<td>8.75-38.25</td>
<td>UK</td>
<td>MONTIBUS LTD</td>
<td>iOS</td>
</tr>
<tr>
<td>My Diet Coach - Weight Loss</td>
<td>—</td>
<td>Yes</td>
<td>2.20-8.75</td>
<td>US</td>
<td>Easy Tiger Apps LLC</td>
<td>iOS</td>
</tr>
<tr>
<td>Fitness Coach &amp; Diet: FitCoach</td>
<td>—</td>
<td>Yes</td>
<td>9.20-57.95</td>
<td>Cyprus</td>
<td>A.L. AMAZING APPS LIMITED</td>
<td>iOS</td>
</tr>
<tr>
<td>Argus: Calorie Counter &amp; Step</td>
<td>—</td>
<td>Yes</td>
<td>9.20-27.35</td>
<td>US</td>
<td>Azumio Inc</td>
<td>iOS</td>
</tr>
<tr>
<td>Speedoc - Care Comes to You</td>
<td>—</td>
<td>No</td>
<td>0.00</td>
<td>Singapore</td>
<td>Speedoc</td>
<td>iOS</td>
</tr>
<tr>
<td>Glucose Buddy Diabetes Tracker</td>
<td>—</td>
<td>Yes</td>
<td>3.70-54.65</td>
<td>US</td>
<td>Azumio Inc</td>
<td>iOS</td>
</tr>
<tr>
<td>DOC2US - Trusted Online Doctor</td>
<td>—</td>
<td>No</td>
<td>0.00</td>
<td>Malaysia</td>
<td>Doc2Us</td>
<td>iOS</td>
</tr>
<tr>
<td>Foodvisor - Nutrition &amp; Diet</td>
<td>—</td>
<td>Yes</td>
<td>17.9-80.90</td>
<td>France</td>
<td>Foodvisor</td>
<td>iOS</td>
</tr>
</tbody>
</table>

*aNot available (information was not found in app stores)*.

**Collating, Summarizing, and Reporting the Results**

After obtaining the screening results, we performed a descriptive analysis, comparison, and functionality assessment based on the information provided by the app developers. In addition, we analyzed the mHealth apps’ category, as well as whether health care experts were involved, whether they were based on scientific testing, and whether they had self-monitoring, based on the description provided by the app providers. Additional information, such as star ratings and the presence of a privacy policy, was also tabulated. A star rating offers a quick overview of an app’s overall user satisfaction, making it a valuable component of quality assessment. The presence of a privacy policy reveals how an app manages user data, ensuring the protection of users’ private health information, which is crucial for informed decision-making.

**Quality Assessment of mHealth Apps**

The quality of the mHealth apps was assessed using the validated MARS evaluation tool, which has demonstrated excellent internal consistency (α=.90) [28]. The MARS has 4 sections for objective evaluation: engagement (eg, the level of entertainment provided, interactivity, and appropriateness of app content), functionality (eg, app performance, ease of use, and navigation), aesthetics (eg, layout, graphics quality, and overall visual appeal), and information (eg, accuracy of the app description, source of information, and quality of information). The subjective quality evaluation section of the MARS subscale has 4 items. However, we excluded this section in this study as the aim was to assess the apps’ quality objectively. Each item was evaluated using a Likert scale with a score range from 1 (inadequate) to 5 (excellent). The overall quality score was calculated based on the mean scores for each of the 4 sections. A mean score of 3 was considered the minimum acceptable score, whereas a score greater than 4 of 5 was preferable. Before
the evaluation, 2 authors independently used each of the apps and conducted the quality assessment in agreement with each other; disagreements were resolved through discussion with a third author.

Results

Search Results

Our search found a total of 1156 apps through keyword retrieval from the Apple App Store and Google Play Store (Figure 1). After removing duplicates, 323 apps were screened; 150 apps met the inclusion criteria as apps focusing on selected chronic disease management (iOS: n=103; Android: n=47) and were included for eligibility assessment. Among them, 42 were included in this study for analysis.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for the systematic search and selection of mobile health (mHealth) apps.
Sample Distribution and App Characteristics

A total of 42 apps (n=28 Android and n=14 iOS apps) were included in this review. Among these, 31 (74%) were for weight management while 11 (21%) were dedicated to chronic disease health management. The chronic diseases commonly targeted by the apps included cardiovascular disease (n=1, 2%), type 2 diabetes (n=3, 7%) and hypertension (n=5, 12%). However, there were 9 apps (21%) that did not specify which chronic diseases they targeted. Most of the apps received recent updates, in the year 2022. The general characteristics of the mHealth apps are described in Table 2. The download count is exclusive to Android apps, as this information was not obtainable for iOS apps.

Table. Sample distribution of chronic disease management apps (n=42).

<table>
<thead>
<tr>
<th>Category</th>
<th>Apps, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight management</td>
<td>31 (74)</td>
</tr>
<tr>
<td>Type 2 diabetes mellitus</td>
<td>3 (7)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>5 (12)</td>
</tr>
<tr>
<td>Cardiovascular disease</td>
<td>1 (2)</td>
</tr>
<tr>
<td>General NCDs&lt;sup&gt;a&lt;/sup&gt;</td>
<td>9 (21)</td>
</tr>
</tbody>
</table>

<sup>a</sup>NCD: noncommunicable disease (this included hypertension, type 2 diabetes mellitus, and cardiovascular diseases).

As shown in Table 1, about 81% (n=34) of the apps provided in-app purchases for live-chat subscriptions, ad removal, and premium feature subscriptions ranging from weekly to yearly, among others. The price for each purchase was between US $0.66 to US $203.26. All the apps included in this review can be freely downloaded by users. Most of the apps (n=15, 36%) were developed in the United States, followed by India (n=7, 17%), Malaysia (n=3, 7%) and Singapore (n=2, 5%).

Star Ratings and Privacy Policies of the Included mHealth Apps

Two apps on the Google Play Store and 6 apps on the Apple App Store received no reviews. The number of users (more than 1 million) who provided ratings for Android apps was significantly greater than the number for iOS apps. Most apps received a user rating of more than 4 stars of a total of 5. In this study, all the apps had a privacy policy. Table 3 shows the star ratings and privacy policies of the included apps.

Table. Star ratings and privacy policies of the included mobile health apps.

<table>
<thead>
<tr>
<th></th>
<th>Android&lt;sup&gt;a&lt;/sup&gt;</th>
<th>iOS&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star rating (1-5 stars), mean (SD)</td>
<td>4.3 (0.3)</td>
<td>4.6 (0.3)</td>
</tr>
<tr>
<td>Privacy policy, n (%)</td>
<td>28 (100)</td>
<td>14 (100)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Apps without ratings: n=2.
<sup>b</sup>Apps without ratings: n=6.

App Feature Assessment

The most common feature that was available on the mHealth apps was self-monitoring (n=35, 83%), which allows users to track their body weight, food intake, fluid intake, and other health indicators. Approximately 16 of 42 apps provided online consultation or personalized feedback via the app as well as goal-setting features (n=35, 83%). Meanwhile, there were no evidence-based apps that used scientific testing or multidisciplinary team involvement in app development, based on the descriptions provided by the developers. About 62% of apps (n=26) introduced at least 1 health care professional (eg, health coach, nutritionist, or dietitian) in the health management of the app. The overview of the functionality characteristics of the apps is presented in Table 4.
Table . Overview of functionality assessment of selected mobile health apps (n=42).

<table>
<thead>
<tr>
<th>Components</th>
<th>Apps, n (%)</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multidisciplinary team involvement in app development</td>
<td>0 (0)</td>
<td>42 (100)</td>
<td></td>
</tr>
<tr>
<td>Health care professional involvement in app health management</td>
<td>26 (62)</td>
<td>16 (38)</td>
<td></td>
</tr>
</tbody>
</table>

**Self-monitoring features**

<table>
<thead>
<tr>
<th></th>
<th>Apps, n (%)</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>35 (83)</td>
<td>7 (17)</td>
<td></td>
</tr>
<tr>
<td>Weight tracker (eg, BMI)</td>
<td>35 (83)</td>
<td>7 (17)</td>
<td></td>
</tr>
<tr>
<td>Diet tracker</td>
<td>29 (69)</td>
<td>13 (31)</td>
<td></td>
</tr>
<tr>
<td>Water intake tracker</td>
<td>16 (38)</td>
<td>26 (62)</td>
<td></td>
</tr>
<tr>
<td>Step count tracker</td>
<td>10 (24)</td>
<td>32 (76)</td>
<td></td>
</tr>
<tr>
<td>Exercise tracker</td>
<td>13 (31)</td>
<td>29 (69)</td>
<td></td>
</tr>
<tr>
<td>Personalised feedback (eg, chat with doctor, nutritionist, health coach)</td>
<td>16 (38)</td>
<td>26 (62)</td>
<td></td>
</tr>
</tbody>
</table>

**Goal setting**

<table>
<thead>
<tr>
<th></th>
<th>Apps, n (%)</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>35 (83)</td>
<td>7 (17)</td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td>35 (83)</td>
<td>7 (17)</td>
<td></td>
</tr>
<tr>
<td>Nutrient intake</td>
<td>16 (38)</td>
<td>26 (62)</td>
<td></td>
</tr>
<tr>
<td>Steps activity</td>
<td>9 (21)</td>
<td>33 (79)</td>
<td></td>
</tr>
<tr>
<td>Exercise</td>
<td>5 (12)</td>
<td>37 (88)</td>
<td></td>
</tr>
</tbody>
</table>

**Medical condition monitoring**

<table>
<thead>
<tr>
<th></th>
<th>Apps, n (%)</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>13 (31)</td>
<td>29 (69)</td>
<td></td>
</tr>
<tr>
<td>Blood pressure</td>
<td>9 (21)</td>
<td>33 (79)</td>
<td></td>
</tr>
<tr>
<td>Glucose</td>
<td>8 (19)</td>
<td>34 (81)</td>
<td></td>
</tr>
<tr>
<td>Heart rate</td>
<td>7 (17)</td>
<td>35 (83)</td>
<td></td>
</tr>
</tbody>
</table>

**Social support**

<table>
<thead>
<tr>
<th></th>
<th>Apps, n (%)</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social support</td>
<td>15 (36)</td>
<td>27 (64)</td>
<td></td>
</tr>
</tbody>
</table>

**Evidence-based testing**

<table>
<thead>
<tr>
<th></th>
<th>Apps, n (%)</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence-based testing</td>
<td>0 (0)</td>
<td>42 (100)</td>
<td></td>
</tr>
</tbody>
</table>

**App Quality Assessment**

The average MARS score among the 42 apps was 3.2 of 5, with a range between 2.0 and 4.1. Of the 4 MARS domains, functionality scored the highest (3.9/5), followed by aesthetics (3.2/5), information (3.1/5), and engagement (2.9/5). There was a large gap in the scores of each subdomain, with the engagement score ranging from 1.6 to 4.0, functionality score ranging from 3.0 to 5.0, aesthetics score ranging from 1.0 to 5.0, and information score ranging from 1.7 to 5.0. The MARS functionality score had the smallest range, and the information score had the largest range. Table 5 shows the MARS subdomain ratings and the total mean score.

Table . App quality rating scores using the Mobile App Rating Scale (n=42).

<table>
<thead>
<tr>
<th>Objective quality rating</th>
<th>Mean score (SD)</th>
<th>Minimum to maximum score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engagement</td>
<td>2.9 (0.6)</td>
<td>1.6-4.0</td>
</tr>
<tr>
<td>Functionality</td>
<td>3.9 (0.5)</td>
<td>3.0-5.0</td>
</tr>
<tr>
<td>Aesthetics</td>
<td>3.2 (0.9)</td>
<td>1.0-5.0</td>
</tr>
<tr>
<td>Information</td>
<td>3.1 (0.7)</td>
<td>1.7-5.0</td>
</tr>
<tr>
<td>Total quality rating</td>
<td>3.2 (0.5)</td>
<td>2.0-4.1</td>
</tr>
</tbody>
</table>
Discussion

Principal Findings

Most NCD management apps in Malaysia lack scientific evidence of efficacy, do not involve multidisciplinary teams, and require significant feature improvements. This study provides a snapshot of current mHealth apps for common NCDs in 2022 that were primarily commercially based. It revealed that the mHealth apps varied in terms of features, functionality, and disease management. The most common category that the mHealth apps currently focus on is self-management of weight, followed by hypertension and type 2 diabetes. Some apps need to be purchased to access extended features, health management information, and medical advice.

Self-monitoring or self-management are crucial components of long-term chronic disease management. Self-management of chronic diseases has been reported to be related to positive health outcomes among patients [29]. Most of the mHealth apps (n=35, 83%) from this study incorporated self-monitoring features that allow the user to monitor their weight, blood pressure, blood glucose level, dietary intake, and fluid intake. The most common self-monitoring features found in the apps were weight trackers, food intake trackers, and step trackers. As technology has advanced, self-management through mHealth apps has been the subject of investigation. Systematic reviews in Korea [14] and the United States [12] examined their impact, shedding light on their potential role in achieving desirable health outcomes through self-management. This suggests that self-monitoring features in mHealth apps may be instrumental in achieving desirable health outcomes.

Some of the apps (n=15, 36%) incorporated social support features to allow users to engage with other users as well as health professionals. These social features could be key for users to continue engaging with the mHealth apps to improve their health. Social support has been shown to improve patients’ health and well-being, and this also applies to online social support networks [30]. Only 38% (n=16) of the apps offered communication with health care professionals via instant messaging or robotic automated message chat functions. These features are beneficial in the self-management of NCDs, as they allow the user to communicate with a health care professional and receive immediate feedback. There is evidence that 2-way communication between patients and health care professionals can improve health outcomes [31]; therefore, app developers should prioritize the inclusion of communication features and health care professionals in app management.

This study reveals a lack of evidence supporting the use of mHealth apps, with none of the apps reporting scientific evidence to indicate the effectiveness of their health management. Based on the star ratings, we cannot deny that mHealth apps could potentially help users improve their health outcomes. Given the fact that most apps on the Apple App Store (iOS) and Google Play Store (Android) did not provide evidence-based testing to prove their effectiveness [32], the app developers were able to make false or misleading claims about their apps. Moreover, health care providers are less likely to feel skeptical of the role of mHealth apps in health care management if the app is supported by research as clinical evidence [33].

The multidisciplinary team approach is a treatment domain that optimizes the health of patients with chronic diseases [34]. Promoting a multidisciplinary team approach is crucial for coordinating the health care system and aiding patients in self-management [35]. However, our research revealed that the majority of mHealth apps did not incorporate a multidisciplinary team to support health management. In this study, only 16% (n=7) of the apps indicated the participation of medical doctors, specialists, nurses, or dietitians. Our findings emphasize the need for multidisciplinary team involvement in health management by using reliable and high-quality mHealth apps. mHealth apps have the potential to function as proactive disease self-management tools [36], meeting diverse needs through collaborative efforts with multidisciplinary teams within the realm of mHealth technology.

The apps in the app stores generally had high star ratings, ranging from 4.3 to 4.6 of 5 stars, which may suggest good user satisfaction. However, it is important to note that user ratings may not always accurately reflect app quality. Therefore, the MARS was used in this study. The overall MARS score of the included apps was 3.2 of 5, which is considered acceptable. Considering the maximum and minimum MARS scores, substantial variability was observed across domains, including aesthetics, which had a range from 1.0 to 5.0. This variance may indicate a significant diversity in app quality, with some evidence suggesting the presence of low-quality apps currently available in the market. However, lower MARS scores could potentially be attributed to reviews of freely available app features.

The review and analysis of mobile app quality for common chronic diseases is crucial for future mHealth app development, as poor app quality can limit their effectiveness in health management [37]. Research on the quality of mHealth apps in Southeast Asia is limited, and the existing studies have primarily focused on COVID-19-related apps rather than those related to chronic disease management [38,39]. Therefore, the authors compared the MARS results with health care app studies from other countries. Our findings align with an assessment of nutrition-related mHealth apps in Korea, where the majority received an average rating, with a mean score of 3.28 of 5 [40]. In contrast, health apps designed for behavioral change in Denmark achieved a slightly higher average quality score, with an average MARS score of 3.48 of 5 [26]. Consistent with previous studies, our study revealed that functionality was best rated (3.9/5), whereas engagement consistently received the lowest score, with the mean score being 2.9 of 5 [41,42]. This indicates that the app developers focused on the functionality of the apps as an essential element in delivering outstanding experiences to the users. On the other hand, some studies discovered that information scored the lowest or fell into the low-to-moderate category [43-45]. This discrepancy highlights that the mHealth apps in the market need special attention to provide more advanced and effective features and capabilities for health management. Importantly, our results indicate that the key areas for improvement in mHealth apps are engagement and information. App developers should prioritize these
domains, focusing on customization, interactive information delivery, and the integration of prompts such as feedback and reminders. Additionally, enhancing sharing functionality and offering more evidence-based content, engaging visuals, and data-driven information for users would improve the overall mHealth app quality.

From a health management perspective, mHealth interventions offer a significant opportunity to facilitate the monitoring of chronic conditions and improve self-management skills [14,24,26]. Evaluating the quality of mHealth apps can help us identify their positive impact on health and behavior outcomes among people with chronic diseases. According to Ryan and Sawin [46], who described an individual and family self-management theory, successful self-management should involve three key components: (1) individual competence, (2) individual motivation, and (3) social factors. mHealth apps have the potential to bridge these factors together to facilitate health management. Additionally, mobile technology promises to enable real-time remote monitoring systems and prompt feedback systems to improve health management [12]. App developers and users should be alert, as our findings show that most apps available in the market are lacking such components.

Strengths and Limitations

The strengths of this study included that the findings were derived and evaluated from the clinical point of view, the study used a validated tool to determine the quality of mHealth apps, and the results can publicly provide review data to users as well as further direction for mHealth app development. This study has several identified limitations. First, we conducted a comprehensive review solely of the apps’ free features, excluding premium features due to budget constraints. Second, at the time the study was completed, it is possible that new mHealth apps or updated features had been released, which we were unable to consider. Third, the review was limited to specific common NCDs, and we could not provide an overview of other diseases. Fourth, our focus was on English-language apps available for download in commonly used mobile app stores, which might limit the generalizability of our findings. According to the world’s largest ranking of countries and regions by English skills, Malaysia ranks among the top 3 in the English Proficiency Index among Asian countries [47]. Nevertheless, we acknowledge that the focus on English-language apps could result in some percentage of Malaysians being left out. Importantly, app developers and researchers are increasingly recognizing the need to cater to diverse linguistic needs within mHealth apps to address the cultural and language diversity of countries like Malaysia [48,49].

Future Direction for mHealth App Development

The findings of this study reveal that the current mHealth app market for managing NCDs in Malaysia is still in its nascent stages and is marked by a shortage of high-quality mHealth apps. In contrast to a previous study conducted in Malaysia in 2017, which predominantly featured informational apps [50], the market is currently undergoing a shift. It is now pivoting toward health management apps. This changing trend underscores the potential for mHealth technology to serve as a cornerstone in the management of chronic diseases in the future. For instance, the results of a prior study that specifically examined hypertension indicated that health care apps could serve as valuable additions to conventional treatment methods [51]. To effectively address NCDs, mHealth apps should incorporate self-monitoring capabilities, such as health tracking, goal setting, and personalized feedback. Additionally, app developers should emphasize elevating the overall quality of their apps by incorporating a variety of perspectives, including input from relevant health professionals and the integration of scientific evidence. It is essential to include this information in the app description to establish trust among users.

Conclusion

A search for mHealth apps for common NCDs available in the Google Play Store and the Apple App Store revealed that most apps focused on weight management, followed by hypertension and type 2 diabetes mellitus. Self-monitoring features such as weight trackers, diet trackers, and step trackers are the core functions of current mHealth apps. This review also highlights the current market’s lack of evidence-based mHealth apps designed specifically for the self-management of chronic diseases. The lack of multidisciplinary teams in app development and health management was observed in the app stores. Evidently, the quality of mHealth apps currently available in the market should undergo ongoing assessment and enhancement to optimize their benefits for users in the realm of health management. Reviews of these apps can offer valuable insights to researchers, health care providers, and app developers, aiding them in delivering high-quality apps for effective health management. App developers and public health authorities should prioritize the development of evidence-based mHealth apps to enhance the mHealth ecosystem for users. Future studies on mHealth content analysis and app evaluation should encompass a broader spectrum of diseases, aiming for a more comprehensive approach that benefits diverse populations.

Authors’ Contributions

KJC and ZAM contributed to conceptualization, methodology, and formal analysis; KJC contributed to writing (original draft preparation); KJC, ZAM, AFML, NHR, SHMA, and NMM contributed to writing (review and editing); ZAM, AFML, and NHR supervised the project. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.
References


Abbreviations

**MARS:** Mobile App Rating Scale
**mHealth:** mobile health
**NCD:** noncommunicable disease
**NHMS:** National Health and Morbidity Survey
**WHO:** World Health Organization

© Khang Jin Cheah, Zahara Abdul Manaf, Arimi Fitri Mat Ludin, Nurul Huda Razalli, Norfilza Mohd Mokhtar, Sawal Hamid Md Ali. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 12.3.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Assessing the Quality and Behavior Change Potential of Vaping Cessation Apps: Systematic Search and Assessment

Fiona McKay1,2, BSc, MPH, PhD; Lilian Chan3, MBBS, MPH, PhD; Rebecca Cerio3, MA; Sandra Rickards3, PhD; Phillipa Hastings3, BHS, MPH; Kate Reakes3, BE, MPH, Grad Cert, MBA, MPA; Tracey O'Brien3, DCH, BSc, LLM, MBA, MBChB; Matthew Dunn1,2, BA, PhD

Corresponding Author:
Fiona McKay, BSc, MPH, PhD

Abstract

Background: An increasing number of people are using vapes (e-cigarettes), and with growing evidence of associated harms, there is a need for acceptable cessation support and interventions. Smartphone apps for health and well-being have increased in popularity and use. Limited published literature assesses the potential of apps to support vaping cessation.

Objective: A systematic search of vaping cessation apps currently available in Australia for iOS and Android platforms was conducted. Apps were assessed against established health app assessment tools for quality and behavior change potential.

Methods: A systematic search through the Australian Apple iTunes and Google Play stores was conducted using the search terms “vape”; “vaping”; “e-cigarette”; and “cessation,” “quit,” or “quitting” in May 2023. Only apps that encouraged the cessation of vaping were included. App descriptions were reviewed to determine if they were relevant for inclusion in this study, and relevant apps were downloaded onto the appropriate mobile device for review. The Mobile App Rating Scale (MARS) was used to rate the quality (engagement, functionality, aesthetics, and information) of the apps using an overall score out of 5. The App Behavior Change Scale (ABACUS) was used to assess the behavior change potential of each app using a score out of 21.

Results: An initial search of the app stores yielded 220 Android apps and 124 iOS apps. Screening against the inclusion criteria left 20 iOS apps and 10 Android apps for review. Six apps were available on both operating systems, and these were downloaded, reviewed, and reported separately for each operating system. The average MARS score for all apps assessed in this review was 3.1 (SD 0.41) out of 5. The reviewed apps overall performed well for the MARS elements relating to functionality, such as ease of use and navigation, but had the lowest scores for information-related elements, such as credibility. The number of ABACUS behavior change features per app ranged from 0 to 19 out of 21, with a mean of 8.9 (SD 4.51). The apps commonly included information-related features, such as requesting baseline information. The least common behavior change features were those relating to goal-setting, such as asking about the user’s willingness for behavior change and providing feedback on current actions in comparison to future goals.

Conclusions: The identified vaping cessation apps had moderate levels of quality and some behavior change components. Future vaping cessation apps could benefit from including more features that are known to support behavior change, such as goal-setting, to improve the potential benefit of these apps to support people to stop vaping. As guidelines for vaping cessation continue to be established, future apps need to reference these in their development.

(JMIR Mhealth Uhealth 2024;12:e55177) doi:10.2196/55177

KEYWORDS

e-cigarettes; quit vaping apps; health apps; behavior change apps; behavior application; behavior; app; application; vaping; smoking; review; vapes; cessation; support; smartphone app; well-being; vapes; mobile device; vaping cessation

Introduction

The use of e-cigarettes, also known as vaping, is a growing health issue, particularly among young people [1,2]. There is increasing concern regarding potential links between vaping and lung, heart, and brain damage as e-cigarettes often contain cancer-causing agents, toxins, heavy metals, and very fine particles that can cause adverse health effects [3-5]. There is also concern that vaping is increasing the prevalence of nicotine addiction, and people who vape are three times as likely as those who do not take up tobacco smoking [3]. Young adults and teenagers disproportionally comprise the vaping population [6]. In Australia, 19.8% of people aged 18-24 years used e-cigarettes, compared with 8.9% in the general adult population [7].
Due to the increasing awareness of the adverse effects of vaping, there is a substantial number of people interested in quitting vaping [8]. In this context, the Australian Federal Government has recently legislated significant reforms to limit the accessibility of e-cigarettes [9]. The addictive nature of nicotine in e-cigarettes means that it can be difficult for people to cease vaping, highlighting the need for developing high-quality, easily accessible, and evidence-based cessation supports. There is currently limited evidence on vaping cessation interventions [10]. Research on tobacco smoking cessation has found that mobile phone–based interventions are effective and acceptable smoking cessation support among young adults [10,11]. As nearly all (99%) people aged 18-24 years in Australia have a smartphone [12], this is a potential medium with which to explore vaping cessation support. Preliminary research demonstrates that smartphone apps would be acceptable or preferred as a vaping cessation tool for people aged 14-25 years [13], and a survey among US high school students found that of those seeking vaping cessation support, 18% had used a mobile app or SMS text messaging support [14].

The use of mobile phones to assist with behavior change (also known as mobile health) has become increasingly common, and apps have been developed for a range of health issues, including smoking cessation [15,16], alcohol reduction or cessation [17,18], and increasing physical activity [19]. There has been some research into the effectiveness of health behavior change apps, and several reviews have identified the need for these apps to have greater integration of behavior change theories and techniques [20-22].

Given the recency of vaping as a health issue, there has been scarce research on the use of apps for vaping cessation. A review in 2020 found that most vaping-related apps available on Google Play promoted vaping (they provided instructions on creating e-liquids, finding stores that sell vaping products, etc), and only 3% supported vaping cessation [23]. A subsequent Canadian review identified only 8 apps that were available on both Android and iOS platforms that had been created for vaping cessation [24]. The review assessed the apps for quality and content, and concluded that there were a limited number of apps available for vaping cessation and highlighted a need for more evidence-based practice in the development of future apps.

While the use of apps for self-management of health issues is increasing, assessment of their ability to support people in changing their health behavior is in its nascent stages, and the rapid pace of change in the app market presents a challenge to thoroughly evaluate their effectiveness in changing health behaviors. Assessment tools that appraise apps based on the inclusion of evidence-based behavior change techniques, such as goal-setting and self-monitoring, are a practical way of rapidly assessing the potential of apps to promote health behavior change [25].

As practitioners are increasingly seeking ways to support people to quit vaping, there is a need for information on both the quality and behavior change potential of vaping cessation support apps. This study aims to assess vaping cessation apps available in Australia by conducting a systematic search and using established health app assessment tools to gauge app quality and behavior change potential.

**Methods**

**Ethical Considerations**

Ethical approval was not required as all data were available in the public domain, and no human participants were involved with this study [26].

**Sample Selection and Inclusion Criteria**

The Australian Apple iTunes and Google Play stores were searched in May 2023 to identify vaping cessation or quitting apps. The Australian app store was searched as the research is based in Australia, and the research team was interested in the apps available in the Australian context. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guideline [27] was modified and adapted to guide this review (Figure 1). All apps that were designed to encourage or promote vaping cessation that were available in Australia were included for analysis. Apps that were designed for smoking cessation but contained some vaping cessation components were also included. Apps that sought to promote behavior change but were not specifically related to vaping or e-cigarette cessation were excluded. Search terms were developed to include any vaping apps: “vape”; “vaping”; “e-cigarette”; and “cessation,” “quit,” or “quitting.”
All apps available for download in Australian app stores containing any of the above keywords in either the title or description were downloaded for analysis. App titles and descriptions were read to determine inclusion in the review. The inclusion criteria included apps that were created for vaping cessation and those in the English language. Any apps promoting vaping or e-cigarettes were excluded. While previous studies of apps have inclusion criteria such as an average user rating or recent update [21], given the small number of apps and the recency of vaping cessation as an issue, broad inclusion criteria were used to include all available apps.

All apps were downloaded for use on an iPhone or Samsung Galaxy Android phone. Where an app was available on both Google Play (Android) and Apple iTunes (iOS), the app was downloaded on each device and rated separately on each operating system because features and functionality may vary across the same app when developed for different operating systems.

The title and description of downloaded apps were recorded in an Excel (Microsoft Corporation) spreadsheet and rated. This study used two scales to rate the apps. The first was the Mobile App Rating Scale (MARS) [28] for quality, and the second was the App Behavior Change Scale (ABACUS) [25] to determine the potential for behavior change. The MARS and the ABACUS show good internal consistency and inter-rater reliability (MARS: $\alpha=.90$, intraclass correlation coefficient 0.79 [28]; ABACUS: $\alpha=.93$, intraclass correlation coefficient 0.91 [25]).
Quality Review and Rating (MARS)
The quality of each app was rated using the MARS [28]. This rating scale examines 19 app elements across four domains: engagement, functionality, aesthetics, and information. Each element was scored using questions on a 5-point ordinal scale. A detailed description of each element and its scoring criteria are provided in the MARS tool [27]. Based on the scores of each element, an aggregated score is calculated for each domain, and an overall app quality mean score (out of 5) is calculated. All apps were reviewed and scored by two authors (FM and MD) who have experience in evaluating health-promoting apps, including those that address addiction and behavior change. Where quality scores differed, the reviewers considered the app together, sought consensus, and determined a final score. The subjective quality rating of the MARS was not included in the analysis due to the subjective nature of this score, in line with commentary from Stoyanov et al [28]. The MARS assessment tool also includes identification of the strategies used by the app (eg, monitoring/tracking, goal-setting, information/education, meditation/mindfulness, and cognitive behavioral therapy) and affiliations of the app (eg, commercial, government, nongovernment organization, or university). These two sections do not contribute to the overall MARS score but are useful to understand the nature and context of the app.

Each app was downloaded and, consistent with other studies [28-30], used for approximately 10 minutes to allow the reviewers to familiarize themselves with the functionality of the app and user experience. Reviewers attempted to use all parts of each app. During the testing, the reviewers noted if the app crashed or if its functions were not accessible. Apps that were not able to be opened or that crashed were removed from the analysis.

Behavior Change Potential Review and Rating (ABACUS)
The ABACUS [25] comprises 21 items and was used to examine the apps’ potential to support behavior change concerning goal-setting, action-planning, barrier identification, self-monitoring, and feedback. Each app was reviewed and scored by two authors (FM and MD). Each app was first explored by the reviewer to gain familiarity with the app and the interface. Reviewers used all app functions including images, cartoons, videos, record-keeping, calendars, and reminders. A total score out of 21 was calculated by summing the item scores.

Where behavior change scores differed, reviewers considered the app together, sought consensus, and determined a final score.

Results
Overview of Search
The initial search of the app stores using the specified search terms yielded 220 Android apps and 124 iOS apps (Figure 1). These apps were screened through a review of the title and description; 18 Android apps and 25 iOS apps met the inclusion criteria and were included for assessment and downloaded onto the appropriate device. Eight Android apps and 5 iOS apps could not be downloaded or were no longer available and were excluded from the sample, leaving 10 Android apps and 20 iOS apps included in this review.

Overview of Included Apps
Most apps that met the inclusion criteria were designed for Apple iOS (n=20), with a smaller number available for the Android operating system (n=10); a minority were available for both operating systems (n=6; see Multimedia Appendix 1 for names of apps). Six apps were created by a commercial organization, 2 by a university, and the remainder (n=22) had an unknown affiliation or developer. Most apps were available for free (n=27); however, of those available for free, most (n=21) had some form of in-app purchase to “enhance” the app experience.

Of the included apps, most used the strategies of allowing the user to monitor/track information (n=25), providing the user with information/education (n=15), providing strategies or tips (n=12), or allowing goal-setting (n=13). A smaller number of apps provided features related to meditation (n=9), cognitive behavioral therapy (n=4), or relaxation (n=2) strategies.

MARS Assessment for Quality
The average overall MARS score of all reviewed apps was 3.1 (SD 0.41). Individual MARS element mean scores ranged from 4 for the elements “ease of use” and “navigation” to 0 for the element “evidence base,” as no app included any information about scientific trialing or testing of the app. iOS apps and Android apps were of similar quality, both with an overall mean MARS score of 3.1. Across all apps, the functionality domain score was the highest, while the information domain score was the lowest (Table 1).
Table. Mean app quality calculated using the Mobile Application Rating Scale.

<table>
<thead>
<tr>
<th>Domain</th>
<th>iOS apps, mean (SD)</th>
<th>Android apps, mean (SD)</th>
<th>Total, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engagement score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entertainment</td>
<td>2.4 (0.8)</td>
<td>2.5 (0.9)</td>
<td>2.4 (0.8)</td>
</tr>
<tr>
<td>Interest</td>
<td>2.4 (0.9)</td>
<td>2.7 (1.0)</td>
<td>2.5 (0.9)</td>
</tr>
<tr>
<td>Customization</td>
<td>2.6 (0.8)</td>
<td>2.4 (0.9)</td>
<td>2.6 (0.8)</td>
</tr>
<tr>
<td>Interactivity</td>
<td>2.8 (0.7)</td>
<td>2.2 (1.0)</td>
<td>2.6 (0.9)</td>
</tr>
<tr>
<td>Target group</td>
<td>3.2 (0.6)</td>
<td>3.1 (0.7)</td>
<td>3.1 (0.6)</td>
</tr>
<tr>
<td>Functionality score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance</td>
<td>3.7 (0.5)</td>
<td>3.7 (0.9)</td>
<td>3.7 (0.6)</td>
</tr>
<tr>
<td>Ease of use</td>
<td>3.9 (0.2)</td>
<td>4.0 (0.5)</td>
<td>4.0 (0.6)</td>
</tr>
<tr>
<td>Navigation</td>
<td>4.0 (0.2)</td>
<td>4.1 (0.4)</td>
<td>4.0 (0.3)</td>
</tr>
<tr>
<td>Gestural design</td>
<td>3.9 (0.2)</td>
<td>4.0 (0.6)</td>
<td>3.9 (0.4)</td>
</tr>
<tr>
<td>Aesthetics score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Layout</td>
<td>3.8 (0.3)</td>
<td>3.8 (0.7)</td>
<td>3.8 (0.5)</td>
</tr>
<tr>
<td>Graphics</td>
<td>3.5 (0.5)</td>
<td>3.8 (0.8)</td>
<td>3.6 (0.6)</td>
</tr>
<tr>
<td>Visual appeal</td>
<td>2.8 (0.9)</td>
<td>3.3 (0.5)</td>
<td>3.0 (0.8)</td>
</tr>
<tr>
<td>Information score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy of app description</td>
<td>3.1 (0.6)</td>
<td>2.9 (0.6)</td>
<td>3.1 (0.7)</td>
</tr>
<tr>
<td>Goals</td>
<td>2.8 (0.5)</td>
<td>2.3 (1.0)</td>
<td>2.6 (0.7)</td>
</tr>
<tr>
<td>Quality of information</td>
<td>3.1 (0.7)</td>
<td>3.0 (1.1)</td>
<td>3.1 (0.8)</td>
</tr>
<tr>
<td>Quantity of information</td>
<td>3.1 (0.8)</td>
<td>3.0 (1.2)</td>
<td>3.1 (0.9)</td>
</tr>
<tr>
<td>Visual information</td>
<td>2.4 (0.7)</td>
<td>3.1 (1.0)</td>
<td>2.6 (0.9)</td>
</tr>
<tr>
<td>Credibility</td>
<td>1.6 (0.7)</td>
<td>1.6 (0.6)</td>
<td>1.6 (0.7)</td>
</tr>
<tr>
<td>Evidence base</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>Total objective score</td>
<td>3.1 (0.5)</td>
<td>3.1 (0.4)</td>
<td>3.1 (0.5)</td>
</tr>
</tbody>
</table>

When considering specific MARS elements across all apps, low mean scores were reported for evidence bases (0, as no apps had this element), credibility (1.6, SD 0.67), and entertainment (2.4, SD 0.84). The highest mean scores were obtained for ease of use (4.0, SD 0.33), navigation (4.0, SD 0.29), and gestural design (3.9, SD 0.41). Over half (n=17, 57%) of all apps were rated with an overall MARS score >3.0.

ABACUS Assessment for Potential to Support Behavior Change

In assessing the apps for behavior change potential, the number of ABACUS behavior change features per app ranged from 0 to 19 out of 21, with a mean of 8.9 (SD 4.51). On average, iOS apps had more behavior change features (9.5, SD 4.6) than Android apps (7.8, SD 4.4). The most common behavior change feature of the 30 apps was the request for baseline information (n=23, 77%), followed by the ability to self-monitor behavior (n=22, 73%) and the ability to personalize or customize the app (n=20, 67%). The least common features were the ability to understand the difference between current action and future goals (n=6, 20%), asking about willingness for behavior change (n=4, 13%), and the ability to export data from the app (n=3, 10%). Table 2 shows the frequencies of the 21 behavior change features evaluated in the apps.
<table>
<thead>
<tr>
<th>Behavior change feature</th>
<th>iOS apps (n=20), n (%)</th>
<th>Android apps (n=10), n (%)</th>
<th>Total apps (N=30), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Knowledge and information</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ability to customize and personalize features</td>
<td>14 (70)</td>
<td>6 (60)</td>
<td>20 (67)</td>
</tr>
<tr>
<td>Consistency with national guidelines or created with expertise</td>
<td>10 (50)</td>
<td>3 (30)</td>
<td>13 (43)</td>
</tr>
<tr>
<td>Request for baseline information</td>
<td>16 (80)</td>
<td>7 (70)</td>
<td>23 (77)</td>
</tr>
<tr>
<td>Instruction on how to perform the behavior</td>
<td>8 (40)</td>
<td>6 (60)</td>
<td>14 (47)</td>
</tr>
<tr>
<td>Information about the consequences of continuing or discontinuing behavior</td>
<td>11 (55)</td>
<td>6 (60)</td>
<td>17 (57)</td>
</tr>
<tr>
<td><strong>Goals and planning</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Request for willingness for behavior change</td>
<td>3 (15)</td>
<td>1 (10)</td>
<td>4 (13)</td>
</tr>
<tr>
<td>Setting of goals</td>
<td>8 (40)</td>
<td>4 (40)</td>
<td>12 (40)</td>
</tr>
<tr>
<td>Ability to review goals, update, and change when necessary</td>
<td>6 (30)</td>
<td>3 (30)</td>
<td>9 (30)</td>
</tr>
<tr>
<td><strong>Feedback and monitoring</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ability to allow the user to easily self-monitor behavior</td>
<td>16 (80)</td>
<td>6 (60)</td>
<td>22 (73)</td>
</tr>
<tr>
<td>Ability to share behaviors with others or allow for social comparison</td>
<td>10 (50)</td>
<td>3 (30)</td>
<td>13 (43)</td>
</tr>
<tr>
<td>Ability to give the user feedback—either from a person or automatically</td>
<td>8 (40)</td>
<td>3 (30)</td>
<td>11 (37)</td>
</tr>
<tr>
<td>Ability to quickly and easily understand the difference between current action and future goals</td>
<td>5 (25)</td>
<td>1 (10)</td>
<td>6 (20)</td>
</tr>
<tr>
<td>Ability to export data from app</td>
<td>1 (5)</td>
<td>2 (20)</td>
<td>3 (10)</td>
</tr>
<tr>
<td>Material or social reward or incentive</td>
<td>8 (40)</td>
<td>3 (30)</td>
<td>11 (37)</td>
</tr>
<tr>
<td>General encouragement</td>
<td>11 (55)</td>
<td>5 (50)</td>
<td>16 (53)</td>
</tr>
<tr>
<td>Reminders or prompts or cues for activity</td>
<td>13 (65)</td>
<td>4 (40)</td>
<td>17 (57)</td>
</tr>
<tr>
<td>App encourages positive habit formation</td>
<td>8 (40)</td>
<td>2 (20)</td>
<td>10 (33)</td>
</tr>
<tr>
<td>App allows or encourages practice or rehearsal in addition to daily activities</td>
<td>10 (50)</td>
<td>5 (50)</td>
<td>15 (50)</td>
</tr>
<tr>
<td>Opportunity to plan for barriers</td>
<td>8 (40)</td>
<td>2 (20)</td>
<td>10 (33)</td>
</tr>
<tr>
<td>Assistance with or suggest restructuring the physical or social environment</td>
<td>8 (40)</td>
<td>2 (20)</td>
<td>10 (33)</td>
</tr>
</tbody>
</table>
Overall Assessment of Individual Apps

The top 5 apps according to their ABACUS and MARS scores are shown in Table 3. While the app with the highest ABACUS score did not have the highest MARS score, the top 5 apps for each operating system were the same, albeit in a different order (see Multimedia Appendix 1 for the ABACUS and MARS scores of all 30 apps included in this review).

Table. Top apps by App Behavior Change Scale (ABACUS) and Mobile App Rating Scale (MARS) scores.

<table>
<thead>
<tr>
<th>App name</th>
<th>Operating system</th>
<th>Developer, affiliation</th>
<th>MARS score (out of 5)</th>
<th>ABACUS score (out of 21)</th>
<th>Subjective quality</th>
<th>Costs (up front, in-app purchases)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quit smoking. Stop vaping app</td>
<td>iOS</td>
<td>Elena Minina, unknown</td>
<td>3.4</td>
<td>19</td>
<td>4</td>
<td>Free up front, no in-app purchases</td>
</tr>
<tr>
<td>Quit vaping for good</td>
<td>iOS</td>
<td>Quit Vaping LLC, unknown</td>
<td>3.6</td>
<td>18</td>
<td>4</td>
<td>Free up front, no in-app purchases</td>
</tr>
<tr>
<td>Quit Tracker: Stop Smoking</td>
<td>Android</td>
<td>despDev, unknown</td>
<td>3.7</td>
<td>15</td>
<td>2.5</td>
<td>Free up front, Aus $4.99 per feature</td>
</tr>
<tr>
<td>QuitSure Quit Smoking Smartly</td>
<td>iOS</td>
<td>Instaquit.org, commercial</td>
<td>3.4</td>
<td>15</td>
<td>3</td>
<td>Free up front, varies Aus $9.99-$39.99</td>
</tr>
<tr>
<td>QuitSure Quit Smoking Smartly</td>
<td>Android</td>
<td>QuitSure, commercial</td>
<td>3.8</td>
<td>14</td>
<td>3</td>
<td>Free up front, varies Aus $0.49-$99.99</td>
</tr>
<tr>
<td>Kwit Quit smoking for good</td>
<td>iOS</td>
<td>KWIT, unknown</td>
<td>3.8</td>
<td>14</td>
<td>4</td>
<td>Free up front, varies Aus $6.49-$119.99</td>
</tr>
</tbody>
</table>

A currency exchange rate of Aus $1=US $0.66 is applicable.

Discussion

Principal Findings

This review identified 30 vaping cessation apps available in app stores and assessed their quality and potential to support behavior change. The 30 apps were created for either iOS (n=20) or Android (n=10), with 6 being available on both operating systems. On average, the vaping cessation apps performed best in functionality features such as navigation and ease of use, and commonly had behavior change features such as allowing users to self-monitor their behaviors. Areas of deficiency were related to specific behavior change strategies such as comprehensive goal-setting, a lack of evidence of trials or testing, and a lack of transparency in the source of the app and its information.

Behavior Change Potential

This review goes beyond the existing research [24] by including an assessment of behavior change potential through the use of the ABACUS tool. The mean ABACUS score identified in this review of vaping cessation apps of 8.9 out of 21 is comparable to the mean ABACUS score of apps focused on other health behaviors, such as 7.6 for physical activity apps, 8.0 for apps to reduce alcohol consumption, and 8.7 for apps to improve mental well-being [21]. Notably, the mean ABACUS score for tobacco cessation apps has been reported as 10.2 [21]. As vaping cessation has some similarities with tobacco smoking cessation [31], app developers could look toward smoking cessation apps to identify strategies and features that could be adapted for vaping cessation apps.

The findings of this assessment indicate that there is potential for developers to create vaping cessation apps that include more features known to support behavior change. Goal-setting is an important feature of behavior change interventions [32], but only 12 of the 30 apps included a goal-setting feature. App developers could consider including more comprehensive goal-setting features, such as adding elements that ask the user about their willingness or readiness for behavior change and providing the user with feedback on how their current actions compare to their future goal, such as in the form of a visual graphic. Future apps could also consider the inclusion of additional features that draw on behavior change strategies, such as meditation, cognitive behavioral therapy, and relaxation, particularly as evidence shows that a significant proportion of young people vape to relieve stress or anxiety [33].

Quality of Apps

The overall quality of vaping cessation apps as assessed by the MARS was moderate, scoring 3.1 out of 5. A review of apps related to other health behaviors identified mean scores ranging from 2.71 for healthy eating apps to 3.26 for apps that aimed to improve mental well-being [21]. One of the areas where vaping cessation apps scored poorly was in credibility, with a mean score of 1.6 compared with a mean of 2.11 that has been identified for health behavior apps overall [21]. It would be...
valuable for future vaping apps to provide more transparency about their information, development, and funding source. One major challenge is that information and guidelines on vaping cessation are in their infancy and may vary across different countries or jurisdictions. As health organizations begin to publish clinical guidelines for vaping cessation [34,35], it would be beneficial for app developers to consider these in the app development process.

The other key area where apps could be strengthened to provide more potential for behavior change is in the establishment of an evidence base. No reviewed app included evidence of being trialed or tested by the developer or an external party. This is a challenge for health apps broadly as the app landscape changes rapidly, and because app development generally outpaces research and knowledge translation, there is often little time for rigorous research. In practice, developing an evidence base for vaping cessation apps may need to occur in parallel to the promotion and general uptake of these apps, as the need for these apps is already present among people who want to stop vaping, particularly among young people for whom digital cessation support is acceptable [14]. Trialing and evaluation of existing and new vaping cessation apps will allow for continual development and improvement over time, and future apps could benefit from partnerships between app developers, researchers, and health behavior change experts.

Considerations for Clinicians and Practitioners

Individuals may seek out assistance and advice from their health care providers when considering vaping cessation. While the treatment approaches for tobacco cessation are long established, there is little guidance relating to vaping cessation [36] nor are there evidence-based clinical guidelines on cessation of dual use (the use of both vapes and tobacco cigarettes) [37]. The Canadian Paediatric Society recently provided preliminary clinical guidance on e-cigarette cessation for young people [38]. These guidelines suggest that behavioral therapy, either in combination with or without pharmacotherapy, should be considered when supporting young people in ceasing vaping or e-cigarette use [38]. This is consistent with Australian guidelines [35] and guidelines from the American Heart Association and American College of Cardiology [39] that suggest that practitioners engage in a range of strategies when advising young people on vaping cessation. A recent randomized controlled trial suggested that tailored mobile interventions could be considered an effective tool to support vaping cessation [40]. As such, in their discussions about vaping cessation, practitioners could encourage people interested in vaping cessation to engage in behavior strategies, such as those apps included in this review.

Considerations for App Developers

Future vaping cessation apps must be developed with an understanding of their target audience and in collaboration with users and clinicians. Users from different age groups may require different features to support their vaping cessation efforts, particularly as they may have different patterns and motivations for vaping [35]. The names and descriptions of some apps included in this review suggest that they were initially developed for tobacco smoking cessation and subsequently adapted for vaping cessation. In Australia, 10.7% of those aged 14 to 17 years are dual users [7], and there are similarities in barriers to and motivations for quitting for both behaviors [31]. Therefore, apps that provide dual support for both health behaviors could be beneficial for some audiences. However, it is important for app developers to create apps with this dual purpose in mind, rather than simply adding vaping cessation to existing smoking cessation app titles.

Finally, it is important to consider the barriers to the uptake of vaping cessation apps. While most of the apps identified in this review were free, many had in-app purchases to access additional features. If a user is unwilling or cannot afford to pay for the additional features, the effectiveness of the app may be compromised. The cost must not be a barrier for people to access health apps [41] and to ensure that people who may not have the ability to pay can access the full level of support for vaping cessation.

Limitations

While there are important findings presented here, this study presents an analysis of apps available at a point in time (May 2023) and only those available in the Australian app stores—app stores in other countries may have different apps. Although this research provides a reference point for further research into the quality of vaping cessation apps, given the fast-moving nature of this field, app developers may modify and update the features of these apps. While the development of the ABACUS assessment tool included reference to tobacco cessation evidence, the tool has only been studied for validity and reliability for physical activity apps [25]. There are perhaps differences in behavior change techniques relevant to vaping cessation in comparison to physical activity. It is also important to note the potential limitation of reviewing apps that were available on both iOS and Android platforms separately. While this may potentially duplicate the identification of certain features in the pool of available apps, this approach has been designed with careful consideration that recognizes that app features and functionality may vary when developed for different operating systems. Unlike other reviews [21], we did not include the app store star rating in our analysis. Many of the apps identified in the app stores were relatively new and as a result had a small number of reviews, making this an unreliable indicator of quality or usability.

Finally, it is important to note that with these data, we are unable to draw firm conclusions relating to long-term behavior change or provide clinical recommendations. It is also known that apps are sometimes used only for a short duration [42]. This is an area that needs more research attention as we strive to create apps that will be able to assist in improving population health at a low cost.

Conclusion

This review of vaping cessation apps found that while apps that are currently available performed reasonably in terms of quality, this review suggests that there is room for improvement, particularly including features that support behavior change. There is a growing interest and need for effective apps to support people to stop vaping, and future vaping cessation apps could
be improved by including specific features known to support behavior change, such as goal-setting, meditation, and relaxation activities, and observing the growing body of clinical guidelines for vaping cessation.

Acknowledgments
This research is funded by the Cancer Institute of New South Wales as part of their vaping control program.

Conflicts of Interest
None declared.

Multimedia Appendix 1
App Behavior Change Scale and Mobile App Rating Scale scores for included apps. [DOCX File, 18 KB - mhealth_v12i1e55177_app1.docx]

Checklist 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist. [DOCX File, 32 KB - mhealth_v12i1e55177_app2.docx]

References


**Abbreviations**

ABACUS: App Behavior Change Scale

MARS: Mobile App Rating Scale

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
“Internet+Nursing Service” Mobile Apps in China App Stores: Functionality and Quality Assessment Study

Shuo Yuan¹,²,*, MSN; Min Liu¹,*, MSN; Yuqi Peng¹, MSN; Jinrui Hu¹, MSN; Bingyan Li¹, MSN; Xia Ding³, BPS; Lunfang Xie¹, PhD

¹these authors contributed equally

Abstract

Background: As the Chinese society ages and the concern for health and quality of life grows, the demand for care services in China is increasing. The widespread use of internet technology has greatly improved the convenience and efficiency of web-based services. As a result, the Chinese government has been implementing “Internet+Nursing Services” since 2019, with mobile apps being the primary tools for users to access these services. The quality of these apps is closely related to user experience and the smooth use of services.

Objective: This study aims to evaluate the functionality, services, and quality of “Internet+Nursing Service” apps; identify weaknesses; and provide suggestions for improving service programs and the research, development, improvement, and maintenance of similar apps.

Methods: In December 2022, two researchers searched for “Internet+Nursing Service” apps by applying the search criteria on the Kuchuan mobile app monitoring platform. After identifying the apps to be included based on ranking criteria, they collected information such as the app developer, app size, version number, number of downloads, user ratings, and number and names of services. Afterward, 5 trained researchers independently evaluated the quality of the apps by using the Chinese version of the user version of the Mobile App Rating Scale (uMARS-C). The total uMARS-C score was based on the average of the five evaluators’ ratings.

Results: A total of 17 “Internet+Nursing Service” apps were included. Among these, 12 (71%) had been downloaded more than 10,000 times, 11 (65%) had user ratings of 4 or higher, the median app size was 62.67 (range 22.71 - 103; IQR 37.51-73.47) MB, 16 (94%) apps provided surgical wound dressing change services, 4 (24%) covered first-tier cities, and only 1 (6%) covered fourth-tier cities. The median total uMARS-C score was 3.88 (range 1.92-4.92; IQR 3.71-4.05), which did not correlate with app store user ratings ($r=0.003; P=.99$). The quality of most apps (11/17, 65%) was average. Most apps (12/17, 71%) were rated as “good” or above ($≥$4 points) in terms of information quality, layout, graphics, performance, and ease of use; however, the vast majority of apps were rated as “fair” or even “poor” (<4 points) in terms of credibility (14/17, 82%) and demand (16/17, 94%).

Conclusions: “Internet+Nursing Service” apps need to broaden their service coverage, increase service variety, and further optimize their service structure. The overall quality of these apps is generally poor. App developers should collaborate with medical professionals and communicate with target users before launching their products to ensure accurate content, complete functionality, and good operation that meets user needs.

JMIR Mhealth Uhealth 2024;12:e52169 doi:10.2196/52169

KEYWORDS mobile phone; mobile applications; home care services; telemedicine; nursing services; China; apps

Introduction

As IT continues to develop, the number of internet users worldwide has reached 4.95 billion in 2022, including 1.067 billion internet users in China, with an internet penetration rate of 75.6%. The proportion of mobile phone users accessing the internet has reached 99.6% [1]. Mobile apps continue to grow steadily, with people using them for communication, shopping, office work, and socializing. The field of internet health care is becoming increasingly standardized, and it was the fastest-growing field in terms of user scale in 2022. As of December 2022, the number of internet health care users in China reached 363 million—an increase of 64.66 million when compared to December 2021—accounting for 34% of all internet users.
users [1]. Mobile health (mHealth) apps are the main media for mHealth and the primary channels for users to obtain medical support and health information. In 2019, 65.9% of internet users chose medical and health apps for medical consultations when feeling unwell [2].

People are increasingly inclined to use health care apps to manage their health due to these apps’ cost-effectiveness, convenience, and speed in accessing health information [3-5] and ability to provide evidence-based health information for better health guidance [6]. There are several popular health care apps in China. According to a report on Statista, as of December 2022, the most popular medical app in China was Ping An Good Doctor—a health care platform owned by the Ping An Insurance Group—with almost 23 million monthly active users [7]. The mobile portal provides real-time medical consultations, web-based appointment booking services, and a health-related discussion forum.

As people become more health conscious, the demand for health care services is increasing. To alleviate the pressure on offline medical institutions and meet the needs of the public, China launched “Internet+Nursing Services” [8]. These services are provided by registered nurses from fixed medical institutions and operate on a web-based app and offline service model. They are designed to serve discharged patients or special populations with medical conditions and limited mobility [9]. Since 2019, “Internet+Nursing Service” organizations in China’s provinces and regions have been actively providing services to countless older adults living at home, pregnant women, infants, and young children who have difficulty with leaving their homes. These services have received unanimous positive reviews [10,11]. The “Internet+Nursing Service” model integrates nursing services with internet technology, using internet IT to break the spatial limitations of traditional medical treatment. It can meet the multilevel nursing care needs of service users, allow for personalized and continuous care, alleviate social problems that result in difficulties with visiting a physician, improve the quality of life of service users, and broaden the channels of communication between nurses and patients [12].

Patients place orders through an app, and managers dispatch orders based on various factors, such as web-based nurses’ qualifications, professionalism, and distance. Platform nurses receive orders within a specified period of time and travel to patients’ homes to provide services, including routine nursing operations (eg, intramuscular injections, intravenous injections, urinary catheterization, gastric tube insertion, and blood sample collection), as well as specialty care (eg, peripherally inserted central venous catheter medication exchange, wound stoma care, and neonatal examinations). By downloading high-quality “Internet+Nursing Service” mobile apps, users can quickly access the care services they need without leaving their homes. However, low-quality apps may not only affect the user experience but also make it impossible for users to obtain the care services they need.

As people become increasingly dependent on smartphones and apps, they are also becoming more concerned about the quality of apps. Users not only expect apps to function properly but also have high expectations for their aesthetics, security, and personalization settings [13]. App quality issues can affect the user experience, determine whether users continue to use the app, and even lead to economic and property losses for both the users and the app development departments. mHealth apps are a special type of app, and studies have shown that factors such as usability, navigability, accuracy of information, and security all affect the user’s experience with and evaluation of an mHealth app. If an mHealth app is of low quality, users may doubt its usefulness and effectiveness, which can even lead to users obtaining low-quality health services and generating erroneous health management concepts that affect their health beliefs and behaviors [6,14,15].

In recent years, various software development organizations have been paying more attention to improving app quality, but there are still deficiencies [16]. The quality of an app is mainly judged by checking user ratings and reviews on app stores. However, the actual quality of an app can be unclear, and it is impossible to know whether an app’s functions are comprehensive or whether its content is scientific based on the data displayed on app store pages. Therefore, it is necessary to objectively evaluate the quality of medical and health apps to identify shortcomings in their development; promote their continuous improvement; improve their applicability, usage experience, and compliance (eg, compliance with data protection laws); and allow users to use apps that are of high quality, are reasonably designed, and are safe to use. There are various methods for evaluating the quality of medical apps, which are also known as mHealth apps. According to a study by Stoyanov et al [17], the quality of mHealth apps is evaluated based on different categories, including engagement, functionality, aesthetics, information quality, and subjective quality [18]. There are various methods for evaluating mHealth services, such as the use of questionnaires, the conduction of interviews, and observation [18]. A systematic review by Nouri et al [19] identified the following seven main classes of assessment criteria for mHealth apps: design, information and content, usability, functionality, ethical issues, security and privacy, and user-perceived value. These criteria can be used to assess the quality of a medical app.

Most studies on “Internet+Nursing Services” focus on service effects [20-22], the establishment of service quality index systems [23], risk management strategies [24,25], and the demand for services from various groups [26-28]. However, there is little attention paid to the quality of “Internet+Nursing Service” apps and a lack of studies that use evaluation tools to objectively evaluate these apps’ quality and functions. The aim of this study was to review the “Internet+Nursing Service” mobile apps that are available on China’s app stores and evaluate their quality.

Methods

Selection of the “Internet+Nursing Service” Apps

We used the Kuchuan mobile app monitoring platform (Beijing Kuchuan Technology Co.) to monitor data from the iOS and Android app stores. This platform provides real-time information about mobile app developers, the latest versions of apps, and the number of app downloads. Two researchers searched for
apps that were available as of December 1, 2022, using the keywords “Internet+Nursing Service,” “Home Nursing,” “Nurses at Home,” “online nurse,” and “shared nurse.” The inclusion criteria for the apps were (1) apps with content that includes home nursing services, (2) apps categorized as health care apps, (3) apps in Chinese, (4) free apps, and (5) functional apps. The exclusion criteria were (1) non–user-side apps, (2) old versions of the same app, (3) apps with different names but the same content, and (4) duplicate apps.

Two researchers independently screened the apps based on their names, profiles, and display images. They then discussed their findings to finalize the list of evaluated apps (Figure 1).

Figure 1. Flowchart of the selection of apps.

Selection of a Standardized Rating Scale for Mobile Apps

We used the Chinese version of the user version of the Mobile App Rating Scale (uMARS-C), and we obtained authorization from the authors of the uMARS-C [29]. Adapted from the Mobile App Rating Scale (MARS), the user version of the MARS (uMARS) has been used for assessing a wide variety of apps, including apps for mental health [30], rheumatology patient management [31], cancer risk assessment [32], and hospital registration [33]. The uMARS-C includes 14 objective items that are rated on a 5-point Likert scale and divided into the following three dimensions: engagement, functionality, and information. Dimension scores are calculated by dividing the
total entry score by the number of entries, and the uMARS-C total score is calculated by dividing the total dimension score by the number of dimensions. According to the rating scale, a uMARS-C total score of 1 indicates poor quality, a score of 2 indicates inadequate quality, a score of 3 indicates fair quality, a score of 4 indicates good quality, and a score of 5 indicates excellent quality.

The uMARS-C has good reliability and validity, with a Cronbach $\alpha$ coefficient of 0.890 and dimension Cronbach $\alpha$ coefficients ranging from 0.853 to 0.895. The test-retest reliability value is 0.967, the item content validity index ranges from 0.78 to 1.00, and the scale content validity index/average is 0.969.

Process of Evaluating “Internet+Nursing Service” Apps

A total of 5 researchers—2 nurses with more than 5 years of experience in their roles, 2 graduate nursing students, and 1 professional internet engineer with more than 7 years of experience in their role—assessed the quality of the apps. Before the evaluation, we made sure that each researcher properly understood and was familiar with the uMARS-C. To ensure their understanding of the scale, they downloaded and assessed 2 mHealth apps that were not included in this study. When there was a difference of more than 2 points in dimension scores or total scores, they discussed until reaching a consensus. The researchers then downloaded the apps that were included in the final analysis onto iOS and Android smartphones. After downloading the apps, they used each app for at least 10 minutes and independently evaluated the ease of use, performance, security, and settings of each app, using the uMARS-C. Basic app information was collected from the app store download page, including the app developer, app size (in MB), version number, number of downloads, user ratings (ranging from 0 to 5), and number and names of services, among others. Some app download data were missing because the iOS app store did not provide these data. The researchers also graded the service coverage cities based on the categorized statistical service items in the Beijing Internet Home Care Service Item Catalog (2022 edition) [34] and the city class divisions in the 2022 China’s City Business Attractiveness Ranking [35].

Statistical Analysis

We used EpiData 3.1 (EpiData Association) for data entry and SPSS 24.0 (IBM Corp) for statistical analysis. Nonnormally distributed measurement data were expressed as medians and quartiles, while count data were expressed as numbers and percentages. The uMARS-C dimension scores for each app were averaged across the five raters, and the final scores were calculated by using the scale’s formula.

Ethical Considerations

This study did not involve human subjects, clinical trials, and vulnerable groups and was therefore exempt from ethical approval.

Results

Characteristics of Selected Apps

Our search found a total of 39,982 apps (iOS: n=209; Android: n=39,773). After initial screening based on the inclusion and exclusion criteria, we downloaded 26 apps. After using them, we excluded 8 apps that were not working properly, leaving a total of 17 apps, which were included in this study (Table 1 and Multimedia Appendix 1). Of these 17 apps, 4 (24%) were released by medical institutions and 13 (76%) were released by corporations (Table 1). Further, 12 (71%) apps had been downloaded more than 10,000 times, with Champion Nurse having the highest number of downloads (n=37,321,776). The median app size was 62.67 (range 22.71-103; IQR 37.51-73.47) MB. App store user ratings ranged from 2.8 to 5.0, with 14 (82%) apps being rated 3.0 or higher and 11 (65%) apps being rated 4.0 or higher. In terms of service coverage, 4 (24%) of the “Internet+Nursing Service” apps covered first-tier cities, including Beijing; 8 (47%) covered new first-tier cities; 4 (24%) covered second-tier cities; and 1 (6%) covered fourth-tier cities.
Table. Characteristics of the “Internet+Nursing Service” apps.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Apps (N=17), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Download count</td>
<td></td>
</tr>
<tr>
<td>0-9999</td>
<td>2 (12)</td>
</tr>
<tr>
<td>10,000-99,999</td>
<td>7 (41)</td>
</tr>
<tr>
<td>100,000-999,999</td>
<td>3 (18)</td>
</tr>
<tr>
<td>1,000,000-9,999,999</td>
<td>1 (6)</td>
</tr>
<tr>
<td>≥10,000,000</td>
<td>1 (6)</td>
</tr>
<tr>
<td>Platform</td>
<td></td>
</tr>
<tr>
<td>iOS and Android</td>
<td>13 (76)</td>
</tr>
<tr>
<td>iOS</td>
<td>3 (18)</td>
</tr>
<tr>
<td>Android</td>
<td>1 (6)</td>
</tr>
<tr>
<td>User rating (number of stars)</td>
<td></td>
</tr>
<tr>
<td>0-2.9</td>
<td>3 (18)</td>
</tr>
<tr>
<td>3.0-3.9</td>
<td>3 (18)</td>
</tr>
<tr>
<td>4.0-4.9</td>
<td>5 (29)</td>
</tr>
<tr>
<td>5.0</td>
<td>6 (35)</td>
</tr>
<tr>
<td>Developer</td>
<td></td>
</tr>
<tr>
<td>Individual developer</td>
<td>8 (47)</td>
</tr>
<tr>
<td>Corporation</td>
<td>9 (53)</td>
</tr>
<tr>
<td>Number of services provided</td>
<td></td>
</tr>
<tr>
<td>1-10</td>
<td>2 (12)</td>
</tr>
<tr>
<td>11-20</td>
<td>9 (53)</td>
</tr>
<tr>
<td>21-30</td>
<td>6 (35)</td>
</tr>
<tr>
<td>Service coverage city class</td>
<td></td>
</tr>
<tr>
<td>First-tier cities</td>
<td>4 (24)</td>
</tr>
<tr>
<td>New first-tier cities</td>
<td>8 (47)</td>
</tr>
<tr>
<td>Second-tier cities</td>
<td>4 (24)</td>
</tr>
<tr>
<td>Fourth-tier cities</td>
<td>1 (6)</td>
</tr>
</tbody>
</table>

*Download count for Android apps only.

Categories of Nursing Services Provided by Apps

The 17 apps provided at-home nursing services, including intravenous injection, intramuscular injection, and nebulized inhalation services, among others. These services were classified as Health Assessment and Guidance, Clinical Nursing, Maternal and Infant Nursing, TCM (traditional Chinese medicine) Nursing, Specialty Nursing, Hospice, and Rehabilitation Nursing services per the categories in the Beijing Internet Home Care Service Item Catalog [34]. Any services that were not in the catalog were classified as uncategorized items.

Of the 17 apps, 16 (94%) provided surgical wound dressing change services (Clinical Nursing category), 15 (88%) provided services for the maintenance of peripherally inserted central catheters (Specialty Nursing category), 14 (82%) provided maternal and infant nursing services (Maternal and Infant Nursing category), and 10 (59%) provided TCM nursing services (TCM Nursing category) and disease rehabilitation guidance (Rehabilitation Nursing category). Further, 4 (24%) apps provided hospice care services (Hospice category), 3 (18%) provided health assessments and guidance (Health Assessment and Guidance category), and only 1 (6%) app provided gastrointestinal decompression, previsit physical examination, conjunctival capsule irrigation, T-tube drainage care, and family room services (uncategorized items).

Quality of the “Internet+Nursing Service” Apps

In our study, the Cronbach α coefficient of the uMARS-C was 0.871, and the dimension Cronbach α coefficients ranged from 0.761 to 0.811. Based on the uMARS-C scores, of the 17 apps, 1 (6%) was rated as “poor,” 11 (65%) were rated as “fair,” 5 (29%) were rated as “good,” and none were rated as “insufficient” or “excellent.” The median total score for the “Internet+Nursing Service” apps was 3.88 (range 1.92-4.92; IQR 3.71-4.05; Figure 2), with Champion Nurse having the
highest score (4.92) and Health WuHan having the lowest score (1.92). There was no significant correlation between app store user ratings and total uMARS-C scores ($r=0.003$; $P=.99$). The median information dimension score was 3.97 (range 1.86-5.00; IQR 3.57-4.29), the median functionality dimension score was 3.95 (range 1.25-5.00; IQR 3.25-4.50), and the median engagement dimension score was 3.80 (range 1.00-5.00; IQR 3.33-4.33; Figure 3). A heat map comparing the scores of each entry in the uMARS-C for the 17 “Internet+Nursing Service” apps showed that Health WuHan scored below 3 points in most entries (11/14, 79%; Multimedia Appendix 2). Further, 14 (82%) apps scored below 4.0 points in the Credibility entry, indicating an average or poor level; 16 (94%) scored below 4.0 points in the Demand entry; 14 (82%) scored at a good or above level in the Quality of Information entry; and 12 (71%) scored at a good or above level in the Layout, Graphics, Performance, and Ease of Use entries.

Figure 2. uMARS-C overall scores of the “Internet+Nursing Service” apps (N=17). The bottom and top edges of the boxes represent the first and third quartiles, respectively; the lines within the boxes represent the medians; the ends of the bottom and top whiskers represent the minimum and maximum values, respectively; and the circles represent outliers. uMARS-C: Chinese version of the user version of the Mobile App Rating Scale.
**Discussion**

**Principal Results**

As internet technology continues to develop, all industries are integrating the “Internet+” model to promote innovation and development [1]. However, unlike services in other industries, “Internet+Nursing Services,” as web-based health care services, are characterized by the high risks and high professionalism of the medical industry, as well as the special risks associated with mHealth [25].

There are several risks associated with “Internet+Nursing Service” apps. One of the risks is that technology barriers can prevent some patients from accessing telehealth services. These barriers can include a lack of access to the internet, a lack of access to the necessary devices, or difficulty with using the technology. Another risk is that there may be issues with insurance coverage for telehealth services, as well as regulatory obstacles that can limit the use of telehealth [36]. Therefore, it is necessary to strengthen the supervision of “Internet+Nursing Service” apps and strictly control all aspects. “Internet+Nursing Services” are human-centered services and aim to improve health. Therefore, “Internet+Nursing Service” app design should...
fully consider user characteristics and needs. In this study, we downloaded and used “Internet+Nursing Service” apps, evaluated them from a user perspective, and rated them objectively based on our usage experience.

Most apps (16/17, 94%) provided services in first-tier cities, new first-tier cities, or second-tier cities, with only 1 app providing “Internet+Nursing Services” in a fourth-tier city. The “Internet+Nursing Service” scope does not yet cover remote areas and townships. Data from China’s seventh national census show that the rural population consists of about 509.79 million people, accounting for 36.11% of the total population [37]. However, 80% of medical resources are concentrated in medium- and large-sized medical institutions in medium- and large-sized cities, leading to an imbalance between the demand for care and the supply of care resources for the grassroots population [38].

Providing high-quality nursing resources and services is key to improving the health and quality of life of people at the grassroots level. Medical institutions at all levels should actively promote the distribution of medical resources to enhance access to basic medical and public health services. Therefore, “Internet+Nursing Service” apps should integrate medical resources and promote a 3-tier “hospital-community-family” linkage to bring professional nursing services into the homes of grassroots people. This would encourage more medical institutions to provide home care, expand the scope of services, make full use of medical resources, and address the imbalance between the supply of and demand for medical resources.

Our study found that “Internet+Nursing Service” apps provide a limited number of service programs—mostly routine care programs—with few special care programs, such as psychological care, hospice care, and child care programs. Only 4 of the 17 apps provided hospice care services, and none provided psychological care services, despite the high demand for these programs. In one study, it was found that 92.3% of the older adult population in urban and rural areas needed psychological comfort [39], and in another study, 10.32% of housebound older adults believed that hospice care should be carried out [40]. Further, as the concept of childbearing changes, people are pursuing more scientific and specialized childcare, and child health care and nursing have become more emphasized. Research has shown that providing nutritional guidance, growth and development guidance, and child psychological care to families of preterm infants through “Internet+Nursing Service” platforms could promote the growth and intellectual development of preterm infants [41]. Medical institutions in each region should have an in-depth understanding of the needs of service users and the characteristics of different groups of people. They should gradually expand the list of “Internet+Nursing Services” by taking into account the actual situations of medical institutions and the local medical resources to optimize the structure of service items and meet the needs of service users.

The total quality scores of the 17 apps ranged from 1.92 to 4.92, with 1 (6%) app rated as “poor” and 11 (65%) rated as “fair,” indicating that the overall quality of “Internet+Nursing Service” apps was not good. This may be related to the fact that app development engineers do not fully understand the medical industry. Most “Internet+Nursing Service” apps (9/17, 53%) were developed by corporations, and the developers may not have taken into account the specificity of medical software before development. They also may not have communicated well with medical staff during app development or understood the content and characteristics of “Internet+Nursing Service” apps.

The process of target user evaluation not only strengthens the interactions between users and the software but also identifies weak points (ie, from user feedback) that developers may have missed [42]. Our researchers found that some apps had problems, such as crashes and the inability to log in during use, which affected the user experience. Therefore, research and development organizations need to conduct premarket research and postmaintenance work to ensure the smooth operation of platforms.

From the heat map (Multimedia Appendix 2), it can be seen that most apps (12/17, 71%) scored at a good or above level in terms of the Layout, Graphics, Performance, and Ease of Use entries. This indicates that most app development teams pay more attention to the visual effects, ease of use, and smoothness of their apps. However, of the 17 apps, 14 (82%) had average scores of less than 4.0 in the Credibility entry, and 16 (94%) had average scores of less than 4.0 in the Demand entry, indicating average or even poor performance in these areas. During the evaluation, our researchers found that some app development teams did not clearly label the source or publisher information when publishing health science articles or videos. This may cause users to doubt the authenticity, authority, and reliability of the articles when reading them, affecting their ability to build health knowledge and manage their own health [43]. It is important to publish health science articles with scientific evidence and to clarify the sources, authors, time of publication, and applicable populations of standardized content to increase the credibility of information. In terms of meeting user needs, we found that most apps provided information from health science literature in text form, with content mostly focused on introducing services, medical institutions, and health care experts. We also noted that after content was published on an app, the content was not updated for a long time. Some apps also only had 2 to 3 articles and could not meet user needs. Therefore, in addition to not meeting user needs in terms of service programs, there are also deficiencies in providing up-to-date and relevant health information.

App developers can improve the credibility of health science articles and videos by taking several steps. First, they should ensure that the information provided is based on scientific evidence and comes from reputable sources. This can be done by clearly labeling the sources, authors, time of publication, and applicable populations of the content. Second, they should regularly update the content to ensure that it is current and relevant. Third, they should provide references or links to the original sources of information, so that users can verify the accuracy of the information. By taking these steps, app developers can increase the credibility of their health science articles and videos and help users build their health knowledge and manage their own health.
Health science popularization should aim to provide basic concepts and knowledge in the field of health, with a focus on healthy lifestyles and behaviors. Health science content should be regularly updated to keep up with social hot spots, seasonal changes, and the occurrence of epidemics to provide users with the most up-to-date and relevant information. To achieve this, app development teams should conduct market research to understand the needs and characteristics of their users, including users’ cultural levels and reading habits. They should also keep track of social hot spots, seasonal changes, and other factors to provide relevant health policies, basic medical knowledge, diet and exercise guidance, psychological guidance, and knowledge regarding disease prevention or first aid in daily life. In addition to providing information in graphic form, app developers can also use video and audio formats to present health science information from multiple angles, dimensions, and levels to meet the needs of users at different levels. By providing accurate and scientific health information in a timely manner, app developers can help users improve their health knowledge and quality of life.

Our study found no correlation between uMARS-C ratings and app store ratings, suggesting that app store ratings do not reflect the quality of apps. This may have been due to the small number of app store ratings, differences in app store rating mechanisms, developer marketing strategies, or users’ preferences (eg, app favoritism among users). If the quality of apps is judged solely based on app store ratings, users may download low-quality apps, thereby affecting their usage experience and even causing them to distrust care services. Additionally, health care professionals may be unable to accurately recommend high-quality apps to patients or their families. This further demonstrates the necessity of evaluating the quality of “Internet+Nursing Service” apps by using objective rating scales.

Limitations
This study has several limitations. First, we only searched for “Internet+Nursing Service” apps that were updated until December 1, 2022, and did not consistently track the uploads and downloads of related apps. Second, we only used iPhone, Huawei, and Xiaomi phones to download and evaluate apps and did not use other systems, such as Meizu, Samsung, and Windows phones. During app development, the development team may modify app functions for different systems due to differences in system algorithms, resulting in differences in app functions. Therefore, future research should take system differences into account and conduct more comprehensive quality evaluations of apps for different systems. Third, there were only 5 researchers in this study; all were under the age of 30 years and had a high level of e-literacy, which may have introduced bias in entries, such as the Ease of Use entry, during the evaluation process. However, older people, who are the main target of “Internet+Nursing Services,” have varying levels of e-literacy and may have different understandings and judgments of an app’s ease of use. Therefore, future studies may consider including evaluators with different backgrounds and health literacy levels.

Conclusions
In this study, we used the uMARS-C to evaluate the quality of “Internet+Nursing Service” apps. We found that the service coverage of these apps was concentrated in first-tier cities (eg, Beijing and Shanghai), new first-tier cities, and some second-tier cities, with a limited number of service items and a need to optimize the structure of service items. The quality evaluation results showed that the quality of apps was not good, especially in terms of information credibility and meeting users’ needs. Further, the scale scores did not correlate with app store scores. Therefore, “Internet+Nursing Service” app development teams need to pay attention to improving the quality of their apps. Before releasing an “Internet+Nursing Service” app, they should fully understand the needs of their target users, as well as the characteristics of this type of app, and communicate with relevant professionals in the field. They should also orient the release of health knowledge in the app toward user needs and improve the credibility and readability of content. After releasing an app, it is necessary to maintain and update it to ensure its normal operation and the timely updating of health education content.

Acknowledgments
We would like to thank three other researchers (Guanghan Zhang, Ying Liu, and Liping Fan) for their participation. This work was supported by the Key Project of Natural Science Research in Colleges and Universities of Anhui Province (grant KJ2021A0255).

Conflicts of Interest
None declared.

Multimedia Appendix 1
Details of the 17 “Internet+Nursing Service” apps.
[DOC File, 23 KB - mhealth_v121e52169_app1.doc ]

Multimedia Appendix 2
Heat map of the average scores for each item and app. The colors range from blue (worst score) to white (best score).
[DOCX File, 1693 KB - mhealth_v121e52169_app2.docx ]
References


TCM: traditional Chinese medicine
uMARS: user version of the Mobile App Rating Scale
uMARS-C: Chinese version of the user version of the Mobile App Rating Scale
Review

Functionality and Quality of Asthma mHealth Apps and Their Consistency With International Guidelines: Structured Search and Evaluation

Billy Robinson¹, BMedSci, BMBS, MMed; Eleni Proimos², BMedSc, MD; Daniel Zou³, BBioMedSci, MD; Enying Gong⁴, BA, MPH, PhD; Brian Oldenburg⁵,⁶, PhD; Katharine See¹, MBBS

¹Department of Respiratory Medicine, Northern Health, Epping, Australia
²Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne, Australia
³Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, Australia
⁴School of Population Medicine and Public Health, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
⁵Baker Heart and Diabetes Institute, Melbourne, Australia
⁶Academic and Research Collaborative in Health, LaTrobe University, Melbourne, Australia

Corresponding Author:
Billy Robinson, BMedSci, BMBS, MMed
Department of Respiratory Medicine
Northern Health
185 Cooper St
Epping, 3076
Australia
Phone: 61 38405 8000
Email: billymed1994@gmail.com

Abstract

Background: Asthma is a chronic respiratory disorder requiring long-term pharmacotherapy and judicious patient self-management. Few studies have systematically evaluated asthma mobile health (mHealth) apps for quality and functionality; however, none have systematically assessed these apps for their content alignment with international best practice guidelines.

Objective: This review aims to conduct a systematic search and evaluation of current mHealth apps in the Australian marketplace for their functionality, quality, and consistency with best practice guidelines.

Methods: The most recent Global Initiative for Asthma (GINA) guidelines were reviewed to identify key recommendations that could be feasibly incorporated into an mHealth app. We developed a checklist based on these recommendations and a modified version of a previously developed framework. App stores were reviewed to identify potential mHealth apps based on predefined criteria. Evaluation of suitable apps included the assessment of technical information, an app quality assessment using the validated Mobile App Rating Scale (MARS) framework, and an app functionality assessment using the Intercontinental Medical Statistics Institute for Health Informatics (IMS) Functionality Scoring System. Finally, the mHealth apps were assessed for their content alignment with the GINA guidelines using the checklist we developed.

Results: Of the 422 apps initially identified, 53 were suitable for further analysis based on inclusion and exclusion criteria. The mean number of behavioral change techniques for a single app was 3.26 (SD 2.27). The mean MARS score for all the reviewed apps was 3.05 (SD 0.54). Of 53 apps, 27 (51%) achieved a total MARS score of ≥3. On average, the reviewed apps achieved 5.1 (SD 2.79) functionalities on the 11-point IMS functionality scale. The median number of functionalities identified was 5 (IQR 2-7). Overall, 10 (22%) of the 45 apps with reviewer consensus in this domain provided general knowledge regarding asthma. Of 53 apps, skill training in peak flow meters, inhaler devices, recognizing or responding to exacerbations, and nonpharmacological asthma management were identified in 8 (17%), 12 (25%), 11 (28%), and 14 (31%) apps, respectively; 19 (37%) apps could track or record “asthma symptoms,” which was the most commonly recorded metric. The most frequently identified prompt was for taking preventive medications, available in 9 (20%) apps. Five (10%) apps provided an area for patients to store or enter their asthma action plan.

Conclusions: This study used a unique checklist developed based on the GINA guidelines to evaluate the content alignment of asthma apps. Good-quality asthma apps aligned with international best practice asthma guidelines are lacking. Future app
development should target the currently lacking key features identified in this study, including the use of asthma action plans and the deployment of behavioral change techniques to engage and re-engage with users. This study has implications for clinicians navigating the ever-expanding mHealth app market for chronic diseases.

**Trial Registration:** PROSPERO CRD42021269894; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=269894

**International Registered Report Identifier (IRRID):** RR2-10.2196/33103

(JMIR Mhealth Uhealth 2024;12:e47295) doi:10.2196/47295

**KEYWORDS**

asthma; mobile health; mHealth; app; mobile; chronic disease; systematic review; smartphone; review methodology; respiratory; compliance; guideline; guidelines; review of apps; evaluation; quality; best practices; apps; mobile phone

**Introduction**

**Background**

Asthma is a chronic respiratory disorder that is clinically defined as a combination of typical episodic respiratory symptoms, such as wheezing, shortness of breath, cough, chest tightness, and significant variable reversible airflow limitation [1]. When the frequency or severity of these symptoms increases compared with the baseline respiratory status, it represents an “asthma exacerbation” or “flare-up” [2]. Judicious self-monitoring and management of regular asthma medications, symptoms, and exacerbations are key to allowing patients with asthma to live with a high quality of life and prevent hospitalizations or death [3]. The Global Initiative for Asthma (GINA) regularly releases updated best practice asthma guidelines based on reviews of scientific literature by an international panel of experts [4]. Many local asthma management guidelines have been derived from these international guidelines. In addition to pharmacotherapy, the guidelines advise that patient education on medication adherence, exacerbation recognition, and management is key to self-management [4].

Asthma is a significant chronic health issue worldwide, affecting 1% to 18% of the global population [4]. Australia is no exception, with asthma affecting millions and accounting for 34% of Australia’s respiratory disease burden and 2.5% of the total disease burden [5]. Asthma leads to numerous emergency department visits and urgent health care visits [5]. Furthermore, those living with asthma report a poorer quality of life and are less likely to rate their health status as excellent or very good [5]. When observing the total cost that asthma has on the Australian health system, it is evident that hospital-related costs outweigh non–hospital-related costs (A$205 million/year [approximately US $150 million] vs A$163 million/year [approximately US $120 million]) [5]. Theoretically, reducing exacerbations would reduce the requirement for hospitalizations; unplanned primary care presentations; and indirect costs, such as work absenteeism, and thus assist in reducing these costs.

With the increasing availability of smartphones, mobile health (mHealth) apps have become accessible to a large percentage of the population and represent a potential medium through which patients can improve their ability to self-manage asthma. Deloitte’s recent review of Australia’s telecommunication status found that 89% of the Australian population uses smartphones [6]. These apps are already available for download and use; however, it is imperative that a review of their quality, functionality, and alignment with evidence-based best practices is conducted to inform both users and health professionals. These apps represent an opportunity for new ways to empower patients to track asthma symptoms, learn about their condition, and undertake practical self-management strategies. The established Mobile App Rating Scale (MARS) is generally used to assess the usability and overall quality of mHealth apps [7,8]. Although systematic evaluations of asthma mobile apps have been conducted in the past, many of these studies did not assess the apps’ functionality or quality using a validated rating scale, such as MARS [9-11]. Furthermore, to our knowledge, none of these prior evaluations assessed all available apps systematically for the presence and quality of information they provide compared with available best practice management guidelines, such as the GINA guidelines [9-11].

This systematic search and evaluation review assessed the functionality and quality of free and paid asthma mHealth apps targeted toward adults with asthma available from the Apple App Store (iOS) and Google Play Store (Android), as well as their consistency with recommendations from the GINA guidelines, making it the first review of its kind.

**Objective**

The objective of this review was to conduct a systematic search of available English-language mHealth apps targeted toward adults with asthma in Australia, to evaluate their overall quality and functionality and to assess the consistency and quality of the content and information they provide in alignment with current best practice guidelines for asthma management.

**Methods**

**Overview**

The GINA guidelines were reviewed by 2 medical professionals to identify and establish a consensus of key recommendations from the guidelines that could feasibly be incorporated into an app for asthma management. The mobile apps in the selected app stores were identified and screened based on the selection criteria. Finally, we assessed the quality, functionality, and alignment of the apps with the guidelines identified in the first step of the screened mHealth apps. An in-depth description of the research protocol was published the previous year [12].

**Study Setting**

Given the primary residences of the researchers involved in this review, this study was conducted by medical practitioners,
medical students, and digital health researchers using apps available in Australia. The mHealth apps presented in English on Australian mobile app stores were assessed. Some mHealth apps identified in this review may not be available in regions outside Australia. Similarly, apps available in other regions may not be available in Australia. However, given that most of the apps identified in this review are also available in other English-speaking regions such as the United Kingdom and the United States, the results are largely generalizable to these regions. Given that the researchers were adult medical practitioners who did not manage pediatric patients, only those mHealth apps targeted toward adults with asthma were evaluated.

Wherever possible, the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines for systematic reviews were followed [13]. Given that this was a review of mobile apps instead of journal articles, some items in the PRISMA checklist were not relevant to this review. The checklist is shown in Multimedia Appendix 1.

**Review of the GINA Guidelines and Checklist Creation**

To assess the usability and overall quality of the app, we used the established MARS [7,8]. A review of the available literature using the CINAHL, MEDLINE, Embase, and PubMed databases revealed that 1 research group had developed an asthma app assessment framework yet to be derived and validated into an instrument [14]. For the reasons outlined in our published research protocol, we decided to combine aspects of the framework by Guan et al [12] with our own checklist derived from the GINA guidelines. Two reviewers, BR and KS, independently assessed the 2020 GINA guidelines for identifiable recommendations that could be incorporated into an mHealth app. Following this, the reviewers examined each other’s identified recommendations to see whether a consensus could be reached on the recommendations from the GINA guidelines that could be incorporated into an mHealth app. The identified recommendations from each author and those where a consensus was reached, which represent the final identified recommendations, are shown in Table 1.

A final checklist modified from the framework by Guan et al [14] (Table 1) was developed to include recommendations we identified from the above process while excluding the information we gathered through the MARS framework. To determine app consistency with the GINA guidelines, participants were assessed for the presence or absence of features identified through this process. This is further discussed in subsequent sections.
Before commencing the initial search for apps, the reviewers ensured that the operating systems on the chosen smartphones were up to date. Each reviewer used different phones to assess these apps followed the approach used in similar studies [9-11]. Our approach for identifying operating systems: the Apple App Store (Apple) and Google Play Store. In the search bar of each store, we input the Android 11 OS (Google) when reviewing apps from the Apple app store.

This review included both free and paid apps from the 2 most popular app stores in Australia across the iOS and Android operating systems: the Apple App Store (Apple) and Google Play Store (Google). Our published protocol outlines the steps taken for quality assurance [12]. Our approach for identifying these apps followed the approach used in similar studies [9-11]. Before commencing the initial search for apps, the reviewers ensured that the operating systems on the chosen smartphones were up to date. Each reviewer used different phones to assess the apps, but all updated the Android operating system (OS) to the Android 11 OS (Google) when reviewing apps from the Google Play Store. In the search bar of each store, we input the term asthma. Two reviewers (BR and DZ) independently searched both app stores on August 10, 2021, from Melbourne, Australia. After obtaining the results for this search term, each reviewer stored the information on an Excel (Microsoft Corp) spreadsheet (Multimedia Appendix 2). The reviewers then compared their results to ensure that they captured all the available apps.

For further evaluation of all the apps identified above, BR and DZ individually reviewed the app title, description, and attached photos and determined whether the app met all inclusion criteria and no exclusion criteria. An identified asthma app was included in the evaluation stage if all the following applied: (1) its primary role was related to asthma, (2) it was targeted to those with asthma, (3) it could be run on mobile phones, and (4) it was in English. Apps were excluded if any of the following

### Table 1. Recommendations identified from the Global Initiative for Asthma guidelines that could be incorporated into a mobile health app.

<table>
<thead>
<tr>
<th>Reviewer 1</th>
<th>Reviewer 2</th>
<th>Consensus reached</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assess symptom control (eg, ACQ[^a])</td>
<td>Support for assessing symptom control for a 4-week period</td>
<td>Support for assessing symptom control for a 4-week period considering the frequency of asthma symptoms, night waking because of asthma, frequency of SABA[^b] use, and any activity limitation because of asthma; uses recognized screening, symptom control or numerical asthma control tools, and tracks peak flow measurement</td>
</tr>
<tr>
<td>Ability to self-track symptoms with or without peak flow</td>
<td>Encourages patients to track symptoms and peak flow measurements</td>
<td></td>
</tr>
<tr>
<td>Risk factors for future exacerbations</td>
<td>Helps users identify the future risk of exacerbations</td>
<td></td>
</tr>
<tr>
<td>Screens for comorbidities and education regarding managing them</td>
<td>Screens for relevant comorbidities and educates patients on the management of these comorbidities</td>
<td></td>
</tr>
<tr>
<td>Inhaler technique with or without video</td>
<td>Provides education on appropriate inhaler techniques</td>
<td></td>
</tr>
<tr>
<td>Ability to record action plan</td>
<td>Provides an area for patients to keep and refer to their written action plan</td>
<td></td>
</tr>
<tr>
<td>Reminder to engage with primary care</td>
<td>Provides reminders to users to see their HCP[^d] for management and review of asthma</td>
<td></td>
</tr>
<tr>
<td>Medication adherence</td>
<td>Specifically provides suggestion to see an HCP if a patient is using only a SABA.</td>
<td></td>
</tr>
<tr>
<td>General asthma education</td>
<td>Provides knowledge on general asthma information, management of asthma, modifiable risk factors and strategies to address them, and when to see an HCP</td>
<td></td>
</tr>
<tr>
<td>Help with activating action plan</td>
<td>Provides advice on when to refer to a patient’s asthma action plan based on self-monitoring of symptoms or PEF[^c]</td>
<td></td>
</tr>
<tr>
<td>—</td>
<td>Prompts patient to see the primary HCP if features of asthma exacerbation (symptoms and SABA use) are identified using the app</td>
<td></td>
</tr>
</tbody>
</table>

[^a]: ACQ: Asthma Control Questionnaire.
[^b]: SABA: short-acting β-agonist.
[^c]: Recommendation identified by one reviewer but not the other.
[^d]: HCP: health care provider.
[^e]: PEF: peak expiratory flow.

**Identification, Screening, and Selection of Mobile Apps for Review**

This review included both free and paid apps from the 2 most popular app stores in Australia across the iOS and Android operating systems: the Apple App Store (Apple) and Google Play Store (Google). Our published protocol outlines the steps taken for quality assurance [12]. Our approach for identifying these apps followed the approach used in similar studies [9-11]. Before commencing the initial search for apps, the reviewers ensured that the operating systems on the chosen smartphones were up to date. Each reviewer used different phones to assess the apps, but all updated the Android operating system (OS) to the Android 11 OS (Google) when reviewing apps from the Google Play Store. In the search bar of each store, we input the term asthma. Two reviewers (BR and DZ) independently searched both app stores on August 10, 2021, from Melbourne, Australia. After obtaining the results for this search term, each reviewer stored the information on an Excel (Microsoft Corp) spreadsheet (Multimedia Appendix 2). The reviewers then compared their results to ensure that they captured all the available apps.

For further evaluation of all the apps identified above, BR and DZ individually reviewed the app title, description, and attached photos and determined whether the app met all inclusion criteria and no exclusion criteria. An identified asthma app was included in the evaluation stage if all the following applied: (1) its primary role was related to asthma, (2) it was targeted to those with asthma, (3) it could be run on mobile phones, and (4) it was in English. Apps were excluded if any of the following
applied: (1) they were not primarily related to asthma, (2) they were primarily targeted toward health care professionals (as stated in the product description), (3) they were not in English, and (4) they were targeted toward pediatric patients. This information was entered into an Excel (Microsoft Corp) spreadsheet for record-keeping (Multimedia Appendix 3). For further evaluation, all apps identified as meeting the above criteria were downloaded by a third reviewer (EP) who identified apps that did not install or function properly after downloading, eliminating them from the review. Finally, the last round of screening was conducted by the reviewer EP. In this round, duplications (ie, apps available on both stores), inaccessible apps, and “lite” version apps, where a pro version was available, were removed from the review. This process was comparable with similar reviews that examined the quality of mobile apps for diabetes self-management [15].

**App Evaluation and Data Extraction**

**Reviewer Training**

A day-long training session was conducted before the initial data extraction. This training session was similar to the one performed by Gong et al [15] for their diabetes app review [15]. A step-by-step reference guide was created by the primary researchers to inform reviewers regarding how to complete the various frameworks and checklists involved in the study. This is provided in Multimedia Appendix 4 [3,4,16,17].

**App Evaluation and Data Extraction Overview**

An internet database was established on Qualtrics (Qualtrics International Inc) for data extraction. A total of 3 reviewers were involved in data collection. A web-based random team generator was used for all apps identified for further evaluation during the screening process so that each app was randomly allocated to 2 assessors (BR, DZ, or EP). The 2 assessors assigned to the app independently reviewed the in-store app description, downloaded the app, and used it for a minimum of 20 minutes to become familiar with all its functions [12]. The reviewers subsequently conducted the evaluations and entered the data into the Qualtrics database. Each reviewer performed this process individually without communicating their results to one another. There are 4 key aspects of the app evaluation and data extraction process, as summarized in the checklist provided in Multimedia Appendix 5 and the step-by-step guide to data collection in Multimedia Appendix 4.

**Technical Information About the App**

The first step in data collection involved gathering basic technical information about the app. The decision of which technical information to include was based on the MARS checklist and previous app review studies [7,9,15]. This was derived from publicly available information in the in-store app descriptions and in-app information sections. If required, the app developer’s website was used. The technical information collected included the app name, date of release, date of update, developer, developer affiliations, price, rating, number of ratings, platform or platforms, size of the app, and number of downloads. A checklist for this section is provided in Multimedia Appendix 5.

**App Quality Assessment**

The app quality assessment was completed using the MARS tool to objectively determine the quality of the apps selected. This scale has 4 separate domains that are assessed to evaluate mobile app quality. These domains are engagement, functionality, esthetics, and information quality [7]. A total of 19 items, each with a 5-point scale regarding quality in the 4 domains mentioned above, make up the MARS score [7]. This framework is presented in Multimedia Appendix 5. Reviewers completed this tool by entering the information into the Qualtrics (Qualtrics International Inc) checklist for each app. Once this was completed, the mean score for that domain and the overall MARS score were calculated for every app. Following the objective MARS section, there are several subjective questions to evaluate user satisfaction and the perceived impact of the app on the user’s knowledge, attitude, motivation to change, likelihood of change, and awareness of the importance of changing their asthma self-management [7]. These questions were answered by reviewers based on their experience using the app and their knowledge gained through the training sessions and clinical practice. Once the data were collected, the mean total MARS value and SD were calculated for each app.

The MARS tool assesses the presence or absence of 19 behavioral change techniques (BCTs). Although 93 types of BCTs are known, only the techniques outlined in MARS were assessed. This approach aligns with previous research, with a scope comparable with our review [7,10,15]. These techniques are outlined in Multimedia Appendix 4. To capture the presence of BCTs, an app was considered to have a BCT present, even if only 1 reviewer identified it. The median number of BCTs and the corresponding IQR of the apps was calculated using Stata (StataCorp) statistical software.

**App Functionality**

App functionality refers to what the app can do for a user and is an important marker of whether an app offers much benefit to users and the overall quality of the app. Although the MARS framework examines the overall quality of a mobile app, it focuses on the performance, ease of use, gestural design navigation, and navigation of the app [7]. Therefore, the Intercontinental Medical Statistics Institute for Health Informatics (IMS) Functionality Scoring System, henceforth known as the IMS functionality score, was used. This score was developed by the above institute and is based on 7 functionality criteria and 4 subcategories in the record functionality criterion. The IMS functionality score focuses on the scope of functions, including the ability of the app to inform, instruct, record, display, guide, remind or alert, and communicate information [18]. Each app was assessed for having or not having these functions and then given a total functionality score between 0 and 11 [18]. To capture the presence of all functionalities, an app was considered to have a functionality present even if one reviewer identified it. The mean, median, and IQR were calculated. The reviewers assessed each app for these functions and entered the data into Qualtrics (Qualtrics International Inc) database. This scoring system is presented in Multimedia Appendix 5.
Presence of App Features Consistent With Asthma Guidelines

As discussed above, key recommendations that could feasibly be incorporated into an asthma mHealth app were identified from the GINA guidelines. These recommendations, summarized in Table 1, were used to develop a more extensive checklist provided in Multimedia Appendix 5. The main functions of the app that we were interested in assessing in our checklist included asthma information, self-management skill training (including peak flow use, inhaler technique, and nonpharmacological strategies), monitoring of asthma symptoms, risk evaluation, and prompting (medication reminders, referring to action plan reminders, and suggestions for seeking health advice). Each of the selected apps was assessed using this checklist, and the data were entered into the Qualtrics (Qualtrics International Inc) database. To ensure consistency, an app was only assessed for the presence of the above function if both reviewers reached a consensus that the said feature was present. For single-reviewed those apps, the sole reviewer’s analysis was used to determine whether the app did or did not have the examined feature.

Quality Assurance, Data Management, and Data Analysis

Training was provided to all the researchers, and a handbook for reviewers was provided to ensure the quality of this research. Selected apps were allocated to reviewers using a web-based random allocation software, and 2 different major app databases were searched to reduce selection bias. The apps were independently reviewed by 2 reviewers to reduce the likelihood of bias affecting the results. A protocol was published to reduce publication bias and enhance the transparency of this study [12]. During app evaluation, all data were entered into either an Excel (Microsoft Corp) spreadsheet during the screening process or into the web-based Qualtrics (Qualtrics International Inc) database. These were stored on a cloud-based system that only the researcher team could access. Once the evaluation was completed, all data were downloaded for subsequent analysis. This analysis comprised a descriptive analysis and calculation of the mean and SD or median and IQR.

All data analyses were performed using Stata statistical software version 14 (StataCorp). Visual figures were generated using Excel (version 16; Office 365; Microsoft Corp).

Results

Identification, Screening, and Selection of Mobile Apps for Review

The process and results of identifying, screening, and selecting mobile apps are shown in Figure 1. A total of 174 unique apps from the Apple Store and 248 unique apps from the Google Play Store were identified. These 422 apps were assessed by 2 reviewers (BR and DZ). In total, 94 apps met all inclusion criteria and no exclusion criteria, although there was a discrepancy between the reviewers’ opinions regarding the eligibility of 39 apps. A third reviewer (EP) identified 17 of these 39 apps as suitable for further review, which resulted in a total of 111 apps suitable for further assessment. A total of 40 apps were removed for reasons outlined in Figure 1. When a more recently updated app was available on 1 platform compared with the other, the older version was excluded from the review. When duplicate apps were available on both platforms and had been updated on the same date, the app from the Apple App Store was retained, whereas the app from the Google Play Store was excluded from the review. This was done to ensure consistency between reviewers and prevent skewing of results by assessing the same app twice. This resulted in 71 apps that were suitable for a complete assessment. Given the delay between the identification of apps and analysis, 18 of the above 71 apps were no longer assessable for the reasons outlined in Figure 1. A total of 4 apps were downloaded by 1 reviewer but not the other before they were removed from the market. This resulted in 49 apps assessed by 2 reviewers and 4 apps assessed by 1 reviewer (53 apps and 102 total reviews).
Figure 1. App screening process and results based on PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines.

Technical Information About the App

The technical information for each reviewed app can be found in Multimedia Appendix 2. Of the 53 apps assessed, 29 (55%) were from the Apple App Store, and 24 (45%) were from the Google Play Store. A total of 19 (36%) apps downloaded from the Apple App Store were also available on Google Play Store. As outlined above, apps available on both marketplaces were only downloaded from the Apple App Store and assessed on the iOS platform. The apps’ last date of update ranged from February 2016 to April 2022. A total of 26 (49%) apps were updated from January 2020. The mean app size was 46.33 MB, and the median app size was 27 MB (IQR 9.2 MB–47.38 MB). App developers were primarily technical companies (28 apps), health care or pharmaceutical companies (4 apps), or a combination of both (4 apps). Six apps were created by private individuals, 3 were created from research or clinical institutions and the remaining 8 were created from developers from a variety of other backgrounds. The number of app downloads ranged from 10 to >10,000. Of 53 apps, 24 (45%) apps had a published user rating, and the median rating was 4 out of 5 (IQR 2.9–4.9). The number of people who provided a rating ranged from 0 to 523, with a median of 2.5 ratings per app (IQR 1–20). Of 53 apps, 42 (79%) apps were completely free, 5 (9%) apps required
users to pay to download, and 6 (11%) of the above free apps had the in-app ability to upgrade for a cost.

**Presence of Behavior Change Techniques**
The total number of each BCT identified across the 53 apps assessed is demonstrated in Table 2. The most frequent BCT observed was *self-monitoring or tracking*, which was identified in 38 (72%) of the 53 apps. The next commonly identified BCTs were *information or education*, seen in 33 (62%) apps, and *advice, tips, strategies, or skills training*, seen in 26 (49%) apps. The average number of BCTs in a single app from those reviewed was 3.26 (SD 2.27). The median number of BCTs for the apps reviewed was 3 (IQR 1-4).

<table>
<thead>
<tr>
<th>Behavioral change technique</th>
<th>Apps with this technique, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information or education</td>
<td>33 (62)</td>
</tr>
<tr>
<td>Self-monitoring or tracking</td>
<td>38 (72)</td>
</tr>
<tr>
<td>Advice or tips or strategies or skills training</td>
<td>26 (49)</td>
</tr>
<tr>
<td>Assessment</td>
<td>17 (32)</td>
</tr>
<tr>
<td>Feedback</td>
<td>12 (23)</td>
</tr>
<tr>
<td>Self-efficacy</td>
<td>9 (17)</td>
</tr>
<tr>
<td>Model or demonstrate behavior</td>
<td>3 (6)</td>
</tr>
<tr>
<td>Rewards and self-rewards</td>
<td>6 (11)</td>
</tr>
<tr>
<td>Goal setting</td>
<td>5 (9)</td>
</tr>
<tr>
<td>Provide social support</td>
<td>3 (6)</td>
</tr>
<tr>
<td>Perceived risks</td>
<td>1 (2)</td>
</tr>
<tr>
<td>Model or demonstrate behavior</td>
<td>3 (6)</td>
</tr>
<tr>
<td>Action planning</td>
<td>11 (21)</td>
</tr>
<tr>
<td>Motivation</td>
<td>2 (4)</td>
</tr>
<tr>
<td>Motivational readiness</td>
<td>2 (4)</td>
</tr>
<tr>
<td>Mindfulness or meditation</td>
<td>1 (2)</td>
</tr>
<tr>
<td>Perceived benefit</td>
<td>1 (2)</td>
</tr>
</tbody>
</table>

**App Quality (MARS)**
The mean MARS score for all reviewed apps was 3.05 (SD 0.54). Of the 53 apps, 27 (51%) achieved a total MARS score of ≥3. A score of 3 on the MARS tool correlates to an “acceptable” quality app, <3 is inadequate or poor quality, and >3 represents a good or exceptional app [7]. Functionality was the highest rated MARS category with a mean score of 3.85 (SD 0.52), followed by esthetics with a mean score of 3.21 (SD 0.6). The information category had an average score of 2.78 (SD 0.83), and engagement had a mean score of 2.77 (SD 0.59). Table 3 shows the mean score for each of the 19 items on the MARS tool. Notably, the apps reviewed had higher scores in the gestural design, app description accuracy, and ease of use domains and lower scores in the evidence base, credibility, and entertainment domains. The mean scores for the quality and quantity of information were 3.21 (SD 1.95) and 2.68 (SD 1.63), respectively.

The final component of MARS allows reviewers to complete a subjective assessment of their opinions on the app. For double-reviewed apps, the mean value for the level of agreement for each domain was first calculated. Therefore, the number of apps will not be a whole number. This is summarized in Table 4 and demonstrates that there were few apps that reviewers would recommend to others, use >2 times in a 12-month period, or pay for. Only 8% (n=4) of the apps were rated >3 stars by the reviewers. For >50% of the apps, reviewers either disagreed or strongly disagreed that the app would impact the user’s knowledge, attitudes, and intentions to change or change the rate of asthma exacerbations. For <20% of the apps, reviewers either strongly agreed or agreed that the app would improve the user’s knowledge, attitudes, awareness, or intention to change behaviors to improve asthma self-management. However, reviewers either strongly agreed or agreed that 29% (n=15) and 27% (n=14) of apps would encourage users to seek further help in asthma management and reduce asthma exacerbations, respectively.
Table 3. Mean score for each category in the Mobile App Rating Scale (MARS) tool for the 53 assessed apps. Each category is assessed on a 5-point scale.

<table>
<thead>
<tr>
<th>MARS category</th>
<th>Mean score for category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entertainment</td>
<td>2</td>
</tr>
<tr>
<td>Interest</td>
<td>2</td>
</tr>
<tr>
<td>Customization</td>
<td>2</td>
</tr>
<tr>
<td>Interactivity</td>
<td>2</td>
</tr>
<tr>
<td>Target group</td>
<td>3</td>
</tr>
<tr>
<td>Performance</td>
<td>3</td>
</tr>
<tr>
<td>Ease of use</td>
<td>3</td>
</tr>
<tr>
<td>Navigation</td>
<td>3</td>
</tr>
<tr>
<td>Gestural design</td>
<td>4</td>
</tr>
<tr>
<td>Layout</td>
<td>3</td>
</tr>
<tr>
<td>Graphics</td>
<td>3</td>
</tr>
<tr>
<td>Visual appeal</td>
<td>2</td>
</tr>
<tr>
<td>Accuracy of app description</td>
<td>4</td>
</tr>
<tr>
<td>Goals</td>
<td>3</td>
</tr>
<tr>
<td>Quality of information</td>
<td>3</td>
</tr>
<tr>
<td>Quantity of information</td>
<td>2</td>
</tr>
<tr>
<td>Visual information</td>
<td>2</td>
</tr>
<tr>
<td>Credibility</td>
<td>1</td>
</tr>
<tr>
<td>Evidence base</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 4. Results from the subjective assessment section of the Mobile App Rating Scale framework (n=53).

<table>
<thead>
<tr>
<th>Would you recommend this app to people who might benefit from it?</th>
<th>Apps, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not at all, I would not recommend this app to anyone.</td>
<td>12 (23)</td>
</tr>
<tr>
<td>There are very few people I would recommend this app to.</td>
<td>18 (35)</td>
</tr>
<tr>
<td>Maybe, there are several people whom I would recommend it to.</td>
<td>18 (35)</td>
</tr>
<tr>
<td>There are many people I would recommend this app to.</td>
<td>4 (8)</td>
</tr>
<tr>
<td>Definitely, I would recommend this app to everyone.</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>How many times do you think you would use this app in the next 12 months?</th>
<th>Apps, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4 (8)</td>
</tr>
<tr>
<td>1-2</td>
<td>28 (52)</td>
</tr>
<tr>
<td>3-10</td>
<td>7 (13)</td>
</tr>
<tr>
<td>11-50</td>
<td>14 (27)</td>
</tr>
<tr>
<td>&gt;50</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Would you pay for this app?</th>
<th>Apps, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>34 (64)</td>
</tr>
<tr>
<td>Yes</td>
<td>18 (35)</td>
</tr>
<tr>
<td>Maybe</td>
<td>1 (2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What is your overall star rating for the app?</th>
<th>Apps, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (one of the worst apps I have used)</td>
<td>10 (19)</td>
</tr>
<tr>
<td>2</td>
<td>20 (38)</td>
</tr>
<tr>
<td>3 (average)</td>
<td>18 (35)</td>
</tr>
<tr>
<td>4</td>
<td>3 (6)</td>
</tr>
<tr>
<td>5 (one of the best apps I have used)</td>
<td>1 (2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strongly disagree or disagree that the app will improve</th>
<th>Apps, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Awareness</td>
<td>40 (75)</td>
</tr>
<tr>
<td>Knowledge</td>
<td>37 (69)</td>
</tr>
<tr>
<td>Attitudes</td>
<td>32 (60)</td>
</tr>
<tr>
<td>Intention to change</td>
<td>32 (60)</td>
</tr>
<tr>
<td>Help seeking</td>
<td>27 (50)</td>
</tr>
<tr>
<td>Behavior change (reduce asthma exacerbations)</td>
<td>29 (54)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strongly agree or agree that the app will improve</th>
<th>Apps, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Awareness</td>
<td>8 (15)</td>
</tr>
<tr>
<td>Knowledge</td>
<td>10 (19)</td>
</tr>
<tr>
<td>Attitudes</td>
<td>7 (13)</td>
</tr>
<tr>
<td>Intention to change</td>
<td>6 (12)</td>
</tr>
<tr>
<td>Help seeking</td>
<td>15 (29)</td>
</tr>
<tr>
<td>Behavior change (reduce asthma exacerbations)</td>
<td>14 (27)</td>
</tr>
</tbody>
</table>

IMS Functionality Score

Out of a potential 11 functionalities, an average IMS functionality score of 5.1 (SD 2.79) was achieved by the reviewed apps. The median number of functionalities identified was 5 (IQR 2-7). A total of 3 apps had 11 functionalities, although most apps had the ability to capture user-entered data (n=37, 70%), provide information in a variety of formats (n=35, 66%), and provide instructions to the user (n=33, 62%). Out of 53 apps, 7 (13%) apps had the ability to send alerts or propose behavioral interventions based on the collected data. The total number of apps that met these IMS functionality criteria is summarized in Table 5.
Table 5. Total number of apps meeting each of the Intercontinental Medical Statistics Institute for Health Informatics (IMS) functionality criteria (n=53).

<table>
<thead>
<tr>
<th>IMS functionality</th>
<th>Apps containing this functionality, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inform</td>
<td>35 (66)</td>
</tr>
<tr>
<td>Instruct</td>
<td>33 (62)</td>
</tr>
<tr>
<td>Record</td>
<td>37 (70)</td>
</tr>
<tr>
<td>Display</td>
<td>30 (57)</td>
</tr>
<tr>
<td>Guide</td>
<td>11 (21)</td>
</tr>
<tr>
<td>Remind or alert</td>
<td>27 (51)</td>
</tr>
<tr>
<td>Communicate</td>
<td>17 (32)</td>
</tr>
<tr>
<td>Collect data</td>
<td>37 (70)</td>
</tr>
<tr>
<td>Share data</td>
<td>26 (49)</td>
</tr>
<tr>
<td>Evaluate data</td>
<td>12 (23)</td>
</tr>
<tr>
<td>Intervene</td>
<td>7 (13)</td>
</tr>
</tbody>
</table>

Presence of App Features Consistent With Asthma Guidelines

Knowledge

Table 6 outlines the results of evaluating the presence of information or knowledge identified as important based on the GINA guidelines. This details the number of apps where reviewer consensus was achieved, the number of apps that provided knowledge or did not provide knowledge in the subcategories, whether knowledge was individualized, and whether knowledge was based on evidence.

Table 6. Number of apps providing knowledge in asthma in the various relevant domains identified as important based on the Global Initiative for Asthma guidelines (n=53).

<table>
<thead>
<tr>
<th></th>
<th>Reviewer consensus achieved, n (%)</th>
<th>Apps that provided knowledge in this domain, n (%)</th>
<th>Apps that did not provide knowledge in this domain, n (%)</th>
<th>Apps that provided individualized knowledge, n (%)</th>
<th>Apps that provided evidence-based knowledge, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General asthma knowledge</td>
<td>45 (85)</td>
<td>10 (22)</td>
<td>35 (78)</td>
<td>0 (0)</td>
<td>7 (16)</td>
</tr>
<tr>
<td>Asthma medications</td>
<td>46 (87)</td>
<td>13 (28)</td>
<td>33 (72)</td>
<td>0 (0)</td>
<td>12 (26)</td>
</tr>
<tr>
<td>Exacerbation management</td>
<td>44 (83)</td>
<td>14 (32)</td>
<td>30 (68)</td>
<td>0 (0)</td>
<td>12 (27)</td>
</tr>
<tr>
<td>Asthma risk factors and triggers</td>
<td>46 (87)</td>
<td>11 (24)</td>
<td>35 (76)</td>
<td>0 (0)</td>
<td>11 (24)</td>
</tr>
</tbody>
</table>

Skill Training

The number of apps that provide skills training in peak flowmeter use, inhaler device use, recognizing and responding to asthma exacerbations, and nonpharmacological management to reduce asthma exacerbations are summarized in Table 7.
Table 7. Number of apps which provide the specific skill training in the areas identified as important in the Global Initiative for Asthma guidelines (n=53).

<table>
<thead>
<tr>
<th>技能训练</th>
<th>Apps with reviewer consensus, n (%)</th>
<th>Apps that provide skill training, n (%)</th>
<th>Apps which provide personalized skill training, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>提供一般技能训练</td>
<td>46 (87)</td>
<td>8 (17)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>提供技能训练</td>
<td>0 (0)</td>
<td>5 (11)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>具有审查者共识</td>
<td>0 (0)</td>
<td>2 (4)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>提供技能训练</td>
<td>0 (0)</td>
<td>5 (11)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>峰流量计使用教育</td>
<td>49 (93)</td>
<td>12 (25)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>提供使用指导</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>通过视频或照片进行操作示范</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>提供使用指导</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>通过视频或照片进行操作示范</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>哮喘发作识别与应对</td>
<td>39 (74)</td>
<td>11 (28)</td>
<td>3 (8)</td>
</tr>
<tr>
<td>鼓励监测症状</td>
<td>0 (0)</td>
<td>9 (23)</td>
<td>1 (3)</td>
</tr>
<tr>
<td>提供哮喘行动计划区域</td>
<td>0 (0)</td>
<td>5 (13)</td>
<td>2 (5)</td>
</tr>
<tr>
<td>引导使用哮喘行动计划</td>
<td>0 (0)</td>
<td>3 (8)</td>
<td>2 (5)</td>
</tr>
<tr>
<td>提供哮喘行动计划的指导</td>
<td>0 (0)</td>
<td>7 (18)</td>
<td>1 (3)</td>
</tr>
<tr>
<td>激励患者寻求健康护理服务</td>
<td>0 (0)</td>
<td>7 (18)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>非药物性管理策略以减少哮喘发作</td>
<td>45 (85)</td>
<td>14 (31)</td>
<td>1 (2)</td>
</tr>
<tr>
<td>帮助识别可能引起症状加剧的触发因素</td>
<td>0 (0)</td>
<td>9 (20)</td>
<td>1 (2)</td>
</tr>
<tr>
<td>避免环境烟雾暴露</td>
<td>0 (0)</td>
<td>13 (29)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>避免可能使哮喘恶化药物</td>
<td>0 (0)</td>
<td>2 (4)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>避免职业暴露</td>
<td>0 (0)</td>
<td>1 (2)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>避免过敏暴露</td>
<td>0 (0)</td>
<td>13 (29)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>避免室内或室外污染</td>
<td>0 (0)</td>
<td>12 (27)</td>
<td>1 (2)</td>
</tr>
<tr>
<td>避免情感压力</td>
<td>0 (0)</td>
<td>4 (9)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>定期进行中等强度运动</td>
<td>0 (0)</td>
<td>6 (13)</td>
<td>1 (2)</td>
</tr>
</tbody>
</table>

App’s Ability to Track and Display Health Information

Table 8 demonstrates the results of the assessment of whether apps had the ability to track and display different aspects of a user’s key asthma information. All information is only related to apps where reviewer consensus was achieved or those that were reviewed by a single researcher. This demonstrates that most apps did not support tracking of all relevant asthma data, and for those that did, manual data input was the predominant entry method.
Table 8. Number of apps which allowed tracking and displaying of the specified asthma information identified as important from the Global Initiative for Asthma guidelines and by what means this information could be input into the app (n=53).

<table>
<thead>
<tr>
<th>App’s ability to track and display users’ asthma information</th>
<th>Apps with reviewer consensus, n (%)</th>
<th>Apps that allow tracking and recording of data, n (%)</th>
<th>Apps with manual data input, n (%)</th>
<th>Apps that allow data entry through external sensors or devices, n (%)</th>
<th>Apps with the ability to create tables or graphs demonstrating trends or analysis of entered data, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma symptoms</td>
<td>51 (96)</td>
<td>19 (37)</td>
<td>19 (37)</td>
<td>1 (2)</td>
<td>14 (27)</td>
</tr>
<tr>
<td>Night waking because of asthma</td>
<td>49 (92)</td>
<td>4 (8)</td>
<td>4 (8)</td>
<td>0 (0)</td>
<td>2 (4)</td>
</tr>
<tr>
<td>Activity limitation because of asthma</td>
<td>47 (89)</td>
<td>3 (6)</td>
<td>3 (6)</td>
<td>0 (0)</td>
<td>3 (6)</td>
</tr>
<tr>
<td>Peak flow meter values</td>
<td>50 (94)</td>
<td>20 (40)</td>
<td>17 (34)</td>
<td>4 (8)</td>
<td>20 (40)</td>
</tr>
<tr>
<td>SABA&lt;sup&gt;a&lt;/sup&gt; use</td>
<td>46 (87)</td>
<td>14 (30)</td>
<td>14 (30)</td>
<td>3 (7)</td>
<td>10 (21)</td>
</tr>
<tr>
<td>Preventive medication adherence</td>
<td>48 (91)</td>
<td>11 (23)</td>
<td>11 (23)</td>
<td>3 (6)</td>
<td>8 (17)</td>
</tr>
</tbody>
</table>

<sup>a</sup>SABA: short-acting β-agonist.

**App’s Ability to Provide Prompts or Reminders**

Table 9 demonstrates the results of the assessment of whether the apps could provide reminders or prompts on the areas deemed relevant from the review of the GINA guidelines, as outlined in the Methods section. Overall, there were few apps that provided reminders or prompts to users, with only 9 (20%) out of 53 apps providing a reminder to use a person’s daily preventive medication. Few apps prompted users to assess the severity and frequency of their asthma symptoms or to seek health advice based on the data they entered.

Table 9. Number of apps that provided reminders and prompts on the specified asthma features chosen based on Global Initiative for Asthma guidelines (n=53).

<table>
<thead>
<tr>
<th>App provides reminders or prompts on asthma features</th>
<th>Apps with reviewer consensus, n (%)</th>
<th>Apps that provide reminders or prompts, n (%)</th>
<th>Apps where reminders or prompts can be individualized, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessing asthma symptoms for the last month</td>
<td>50 (94)</td>
<td>2 (4)</td>
<td>1 (2)</td>
</tr>
<tr>
<td>Appointment with physicians</td>
<td>47 (89)</td>
<td>1 (2)</td>
<td>1 (2)</td>
</tr>
<tr>
<td>Performing peak flow test</td>
<td>50 (94)</td>
<td>5 (10)</td>
<td>4 (8)</td>
</tr>
<tr>
<td>Preventive medication adherence</td>
<td>45 (85)</td>
<td>9 (20)</td>
<td>8 (18)</td>
</tr>
<tr>
<td>Checking the date of expiry and dosage of inhalers</td>
<td>51 (96)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Warning of changing health data (ie, very frequent exacerbations)</td>
<td>53 (100)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Seeking urgent health advice based on the health data the user inputs into the app</td>
<td>52 (98)</td>
<td>2 (4)</td>
<td>2 (4)</td>
</tr>
</tbody>
</table>

**Other App Information**

Table 10 summarizes the other important features assessed during this review. Only 1 app allowed the user to make an appointment with a physician. Of import, only 5 (10%) of the apps were identified as having an area where users could keep a record of their asthma action plan. Most of these (4 apps) allowed the user to type in their action plan, and 1 app allowed users to upload a photo of their action plan. Only 1 app was identified using recognized asthma screening tools to assess patient’s current asthma symptoms and severity.
Table 10. Summary of number of apps containing further features and content identified as important (n=53).

<table>
<thead>
<tr>
<th>Feature Description</th>
<th>Reviewer consensus, n (%)</th>
<th>Apps with this feature, n (%)</th>
<th>Apps without this feature, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Make appointment with physicians</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>For web-based consultation</td>
<td>53 (100)</td>
<td>1 (2)</td>
<td>52 (98)</td>
</tr>
<tr>
<td>For face-to-face consultation</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>53 (100)</td>
</tr>
<tr>
<td><strong>App provides an area for patients to keep record of their asthma action plan</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Can type in action plan</td>
<td>52 (98)</td>
<td>5 (10)</td>
<td>47 (90)</td>
</tr>
<tr>
<td>Can upload a photo of action plan</td>
<td>0 (0)</td>
<td>4 (8)</td>
<td>48 (92)</td>
</tr>
<tr>
<td><strong>The app includes social forums or blogs that promote peer-support and communication among asthma patients</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50 (94)</td>
<td>1 (2)</td>
<td>49 (98)</td>
</tr>
<tr>
<td><strong>The users could send recorded data to others</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>To physicians</td>
<td>44 (83)</td>
<td>11 (25)</td>
<td>33 (75)</td>
</tr>
<tr>
<td>To each other</td>
<td>0 (0)</td>
<td>9 (21)</td>
<td>33 (75)</td>
</tr>
<tr>
<td><strong>The app could help the users to evaluate the risk of having future asthma exacerbations</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using a validated scoring system</td>
<td>44 (83)</td>
<td>2 (5)</td>
<td>42 (96)</td>
</tr>
<tr>
<td>Yes, but not using a validated scoring system</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>44 (100)</td>
</tr>
<tr>
<td>The app could guide the users to find out the closest pharmacy, hospital, or clinic</td>
<td>51 (96)</td>
<td>0 (0)</td>
<td>51 (96)</td>
</tr>
<tr>
<td><strong>The app uses recognized screening, symptom control and numerical asthma control tools</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes, all 3</td>
<td>46 (87)</td>
<td>1 (2)</td>
<td>45 (98)</td>
</tr>
<tr>
<td>Yes, screening tool</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>46 (100)</td>
</tr>
<tr>
<td>Yes, symptom control tool</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>46 (100)</td>
</tr>
<tr>
<td>Yes, numerical asthma control tools</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>46 (100)</td>
</tr>
<tr>
<td><strong>The app allows users to connect to wearable technology</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smartwatch</td>
<td>53 (100)</td>
<td>8 (15)</td>
<td>45 (85)</td>
</tr>
<tr>
<td>Activity sensor (eg, Fitbit)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>53 (100)</td>
</tr>
<tr>
<td>Smart peak flowmeter</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>53 (100)</td>
</tr>
<tr>
<td>Handheld spirometer</td>
<td>0 (0)</td>
<td>3 (6)</td>
<td>50 (94)</td>
</tr>
<tr>
<td>Oximeter</td>
<td>0 (0)</td>
<td>2 (4)</td>
<td>51 (96)</td>
</tr>
<tr>
<td>Other</td>
<td>0 (0)</td>
<td>6 (11)</td>
<td>47 (89)</td>
</tr>
</tbody>
</table>

Discussion

Principal Findings

This review aimed to evaluate the quality and functionality of asthma apps and their consistency with international best practice guidelines. We conducted a comprehensive review of 422 asthma apps available on the Australian App Store and Google Play Store, of which we selected 53 apps that met our inclusion and exclusion criteria for detailed analysis. The most common reason for app exclusion was that they were not primarily related to asthma. Most apps were developed by technical companies rather than health care facilities and clinical or research institutes. This lack of involvement of practicing experts in the field of health care is concerning, as these apps are primarily related to the management of chronic health conditions. Apps that involved a clinician during the designing phase demonstrated a greater ability to facilitate behavioral change than those that did not [19]. We believe that in the future, asthma mHealth apps should be developed in consultation with health care professionals and organizations to ensure that they meet an appropriate standard.

Our review revealed that most asthma apps do not use key BCTs that can promote behavioral changes through feedback, goal setting, and rewards. Instead, the most commonly used BCTs were self-monitoring or tracking, education, and advice. This is consistent with other studies and demonstrates a potential area where asthma apps could be improved in the future [19,20]. Functionalities describe what an app can do for its user and are an important marker of whether an app offers much benefit to users. Our review demonstrated that although basic functionalities, such as informing, reading, collecting, instructing, and displaying, were prominent, more complex functionalities were lacking. The ability to evaluate and intervene based on the app information entered is not present
in most apps. Creating apps with this functionality could, for example, guide patients to see their health care provider based on the data they are entering, such as excessive short-acting β-agonist use or asthma symptoms. This feature was demonstrated in the Asthmahub app from NHS Wales, one of the higher-rated apps on assessment.

Our evaluation using the MARS tool showed that the asthma mHealth apps performed poorly in the information and engagement domains compared with the functionality and esthetics categories. These findings are consistent with previous asthma app reviews that showed poorer results in these categories [9]. Only half of the apps achieved an acceptable standard, and even fewer achieved a dual rating average >4, indicating a “good-quality” app. These apps were Kiss myAsthama, Asthmahub, and AioCare Patient. Acceptable or good ratings for subcategories such as app credibility, evidence base, and entertainment were particularly lacking in the apps we assessed. In contrast to this lack of evidence-based content identified through app assessment with the MARS framework, assessment with our checklist found that the knowledge presented in apps was largely based on evidence. “Evidence-based” in the MARS checklist refers to whether an app has undergone a clinical trial. By contrast, in our checklist we refer to “evidence-based” as that the information presented in the app was factual and in alignment with our clinical knowledge of asthma and the GINA guidelines. This explains the contrasting results between the MARS framework and our checklist for “evidence-based” knowledge. Although information was often factual and based on guidelines, the apps had not undergone clinical trials. The subjective star rating provided by reviewers was rarely >3, whereas the mean user rating was 3.56 out of 5, indicating a discrepancy between the perspectives of the reviewers and the user. This discrepancy in user ratings and reviewers’ perspectives has been demonstrated in similar app reviews in the past [9]. We propose that this could be because of our health care background bias when assessing the apps, even though we were attempting to assess apps from the perspective of a patient with asthma, or because we were approaching the review with a critical lens following an objective app assessment, a different mindset from the usual user.

Our review demonstrated that asthma mHealth apps do not contain key features consistent with international best practice guidelines for asthma management. Few apps contained important information regarding asthma and asthma management, and even when provided, they were not individualized to the user. Personalization is a key part in the management of asthma, with an individual’s triggers, baseline respiratory function, and inhaler device being common things we assess for, educate on, and consider when managing asthma in the community. Furthermore, personalized therapeutic management is a key aspect of asthma management, as outlined in the GINA guidelines [4]. There is also an expectation from previous research that personalized technological interventions may lead to better health outcomes, although this has not yet been specifically demonstrated for asthma mHealth apps [21]. The NHS Wales AsthmaHub app provides one solution to this problem by starting with the creation of a profile in which the user answers questions regarding asthma triggers, control, and medications. This allows the app to provide information targeted toward the user. Skill training in a field can be achieved by apps through written and visual information to teach users how to use peak flow meters, inhalers, and nonpharmacological strategies to manage asthma and asthma exacerbation. These are skills that the GINA guidelines advocate patients to learn and become proficient in, with the support of their health care provider [4]. Training in the use of inhaler devices, spacers, and peak flow meters, all of which in our experience, patients can inadvertently misuse, is lacking in most apps and is something that we feel is key to be included in an asthma mHealth app. Although not essential for all inhaler types, such as dry-powder inhalers, spacers are considered an essential part of using metered-dose inhaler preventer and reliever medications. Therefore, we believe it is reasonable that a high-quality asthma mHealth app should include advice to the users regarding the use of spacers, particularly given the prevalence of salbutamol or albuterol metered-dose inhaler use among those with asthma. No apps could individualize this training to the user, and given the magnitude of devices on the market and the difference in how they work, this represents a deficiency. In total, 40% (n=20) of the apps could record and display peak flow meter values and asthma symptoms. Peak flow is representative of worsening airway obstruction or asthma. If this value is reducing, it can indicate that a person’s asthma may be on the precipice of or in the middle of an exacerbation, and as such, it can be a valuable metric for patients and clinicians to monitor. Other key metrics that could be tracked were lacking, such as the amount of short-acting β-agonists and adherence to preventive medications. It is important for these metrics to be recorded, as excessive SABA use can indicate an asthma exacerbation or poorly controlled asthma and should prompt a review by a health care professional. Subtler signs of poor asthma control, such as activity limitation or night waking because of asthma, were rarely recorded and represented missed opportunities. Prompting and reminding users to do something is a basic functionality that is largely underutilized by apps. Chronic asthma management involves remembering to consume daily inhaler medications, assessing symptoms, and regular interactions with a health care provider [4]. All these tasks lend themselves to a reminder from an app to assist with asthma management, and their absence from most apps is a missed opportunity.

Asthma action plans are a key aspect of contemporary patient self-management of asthma [4]. The ability to quickly refer to this plan on a user’s digital device should be easy to include; however, most apps lack this feature. We see this as a significant deficiency and missed opportunity. Few apps use validated asthma symptom scoring systems for users to assess their symptoms and risk of exacerbation. Wearable technology and external sensors are a growing medium through which data related to a patient’s asthma status can be captured, yet only few apps use this technology. When external sensors are used, they are often smart spirometers (such as in AioCare patient) or peak flow meters, both of which carry extra costs and may not be palatable to all users. No apps were found to be linked to smartwatches or activity sensors, the use of which is becoming more prominent [22]. With the availability of pulse
oximetry in smartwatches, this may be a new method for asthma apps to collect crucial data in the future.

Overall, we determined the quality of apps to be average at best, with many lacking features consistent with international best practice guidelines. Three apps achieved a MARS rating of >4. These were Kissmyasthma, Asthmahub, and AioCare patient. Their alignment of these apps with international best practice guidelines was mixed, with AioCare patients not having many of the functions deemed important by the guidelines. The Kissmyasthma and Asthmahub apps had great amounts of information related to asthma, consistent with the guidelines. Asthmahub stood alone in having these features and many features to support skill training and track and record information; however, its prompting or reminding functions were minimal. Notably, both these apps that ranked higher in MARS and our checklist were developed by health services or medical research centers (NHS Wales, the University of Sydney, The University of Melbourne, and the Woolcock Institute of Medical Research). All 11 functionalities from the IMS scale were identified in Asthmahub and AioCare patients, and Kissmyasthma had 6 functionalities, which were above the mean identified. A total of 6 to 8 BCTs were identified in these apps, which, although greater than the median number of BCTs, still did not include several potential BCTs.

Comparison With Previous Work
Prior studies that have conducted these assessments have primarily focused on evaluating the quality and functionality of apps using the MARS framework [9-11]. From a review of the literature for the past 5 years, only 2 prior studies were found to have conducted some sort of assessment of the alignment of apps with asthma self-management principles. Both these studies only looked at free apps, eliminating a number of apps from the review [11,23]. Data collection for both reviews occurred years ago [11,23]. In the rapidly developing marketplace of mobile apps, a number of new apps have been released during this time. Our review examined both free and paid apps, and provided an updated assessment, given that our data collection took place in 2022. Furthermore, Househ et al [23] did not assess apps from the Apple App Store, focusing only on the Google Play Store, and therefore, did not fully assess the breadth of available English-language apps in the marketplace. These authors evaluated whether apps included or did not include information consistent with GINA guidelines as per a checklist created by 1 author [23]. This was limited to asthma information and education and did not include further features, such as the ability of an app to track information, provide asthma skill training, or personalize information. This review also did not examine the overall app quality using the validated MARS framework [23].

Our review benefits from having 2 independent clinicians review the guidelines to establish all GINA self-management recommendations that can be feasibly incorporated into an mHealth app and review app quality using the MARS framework. Furthermore, we examined not only the presence of information, but also the presence or absence of the ability to track asthma symptoms and provide reminders and skill training, as well as all features derived from the GINA guidelines provided in Multimedia Appendix 5. As part of their app review, Tan et al [11] established a framework for assessing the alignment of mHealth apps with the theoretical principles of self-management of allergic rhinitis or asthma [11]. A total of 6 asthma self-management principles were identified based on a literature review and author consensus [11]. Our review has taken this a step further, specifically deriving self-management principles from the international best practice GINA guidelines and creating a more extensive checklist based on these principles. The inclusion of paid apps, the creation of an asthma self-management principle checklist derived from international best practice guidelines, and the up-to-date nature makes our study unique.

Limitations
This study has several limitations. First, we limited it to English-language apps available from 2 app stores in Australia. Although these stores make up a significant portion of the market, other stores do exist, such as the Blackberry store, which were not included in the review. It is also expected that some apps available on the Australian marketplace may not be available on international marketplaces, whereas some apps available internationally may not be available on the Australian marketplace, and thus not included in the review. Although the operating systems for phones used by reviewers were updated on the same day to ensure that the same OS was on each smartphone, the phones themselves were different models. This may have affected the user or reviewer experience to an unknown degree, reducing standardization in this study. A total of 4 apps were reviewed by 1 researcher only as they were removed from the market or were not available to the second reviewer by the time they tried to assess it. This may skew some results, although previous studies have only single-reviewed apps; therefore, the fact that the most apps in this study are double-reviewed is a strength of this research.

Furthermore, we only assessed the presence or absence of those BCTs embedded in the MARS checklist. This limited the number of BCTs that we assessed; however, reviewing the presence or absence of all 93 BCTs was outside the scope of this review and our protocol and requires further research.

This evaluation was researcher-based, and thus does not reflect the real assessment from a patient’s perspective. Future research should also include people with asthma to determine their responses to the quality and functionality of these apps.

Conclusions
Currently, there is a lack of high-quality asthma apps aligned with international best practice asthma guidelines. Most apps do not provide patients with important asthma information, skills training in key aspects of asthma management, recording and tracking relevant data, or reminders regarding asthma control. The lack of interaction with smart technology or use of asthma action plans are significant flaws that merit attention in future apps. Few BCTs or in-depth functionalities have been deployed to engage and re-engage with users or generate meaningful behavioral modifications. Again, we see that app designers are typically skilled in providing an aesthetically pleasing functional app but lack skills in providing engagement and information, as assessed in the MARS tool. Future app...
development should target the key features identified in this study as currently lacking. Furthermore, not only are high-quality asthma mHealth apps lacking, there are minimal robust clinical trials examining the use of these apps and their effect on patient outcomes. Further research in this area will be valuable to determine the true clinical utility of these apps. A wide spectrum of technological quality, accuracy, and breadth of health information was seen among available apps. Given this spectrum of poor-to high-quality apps in an unregulated market, it is unlikely that future guidelines or health professionals will be able to make generic recommendations to patients regarding asthma mHealth app use and instead will need to make targeted recommendations about specific apps. Guidelines that incorporate reviews such as this review that identify apps known to be of high quality, such as Asthmahub, AioCare patient, and Kissmyasthma, will be an important future resource for general clinicians navigating the ever-expanding mHealth app market for chronic diseases.

Acknowledgments
Funding was received from the Northern Health Respiratory Department (Victoria, Australia) for the costs associated with submitting the protocol and manuscript and purchasing mobile health apps. No funding or sponsorship was received from the mobile app developers or third parties.

Authors' Contributions
BO and KS conceived this study. BR developed the study protocol by adapting a protocol previously created by EG for a diabetes mobile health app review project. KS and BR reviewed the Global Initiative for Asthma guidelines to establish key recommendations that could be feasibly incorporated into a mobile health app. BO and KS provided valuable guidance and advice for this research. BR, DZ, and EP reviewed the apps. BR was the primary author of this manuscript. EG and BR also provided training to all the reviewers of the project. All authors contributed to the refinement of the study protocol and approval of the final manuscript. The Respiratory Research Team at the Northern Hospital assisted in providing general advice to the authors.

Conflicts of Interest
The Northern Health Respiratory Department is providing funding for this study. The head of this department, KS, is a key author of this paper. The department does not have competing or fiduciary interests that could be affected by the results of this study. KS is not responsible for directly reviewing the apps included in the study or for the data analysis.

Multimedia Appendix 1
Completed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 769 KB - mhealth_v12i1e47295_app1.pdf ]

Multimedia Appendix 2
App technical information.
[XLSX File (Microsoft Excel File), 23 KB - mhealth_v12i1e47295_app2.xlsx ]

Multimedia Appendix 3
Data from initial app review process to identify apps that met inclusion but not exclusion criteria.
[XLSX File (Microsoft Excel File), 265 KB - mhealth_v12i1e47295_app3.xlsx ]

Multimedia Appendix 4
Stepwise guide on the key steps of data extraction and evaluation for reviewers to follow.
[DOCX File , 161 KB - mhealth_v12i1e47295_app4.docx ]

Multimedia Appendix 5
Data extraction forms (technical information, Mobile App Rating Scale, Intercontinental Medical Statistics functionality score, and asthma assessment checklist).
[DOCX File , 47 KB - mhealth_v12i1e47295_app5.docx ]

References


Review

Developing a Comprehensive List of Criteria to Evaluate the Characteristics and Quality of eHealth Smartphone Apps: Systematic Review

Janette Ribaut¹,², PhD, RN; Annette DeVito Dabbs³,⁴, PhD, RN; Fabienne Dobbels¹,⁵, PhD; Alexandra Teynor⁶, PhD, Dipl-Ing; Elisabeth Veronica Mess⁶, MA; Theresa Hoffmann¹,⁷, MSc, RN; Sabina De Geest¹,⁵, PhD, RN

¹Institute of Nursing Science, Department Public Health, University of Basel, Basel, Switzerland
²Department of Hematology, University Hospital Basel, Basel, Switzerland
³School of Nursing, Department of Acute & Tertiary Care, University of Pittsburgh, Pittsburgh, PA, United States
⁴Clinical Translational Science Institute, University of Pittsburgh, Pittsburgh, PA, United States
⁵Academic Center for Nursing and Midwifery, Department of Public Health and Primary Care, KU Leuven, Leuven, Belgium
⁶Department of Computer Science, University of Applied Sciences, Augsburg, Germany
⁷Department Pflege und Betreuung, Genossenschaft Alterszentrum Kreuzlingen, Kreuzlingen, Switzerland

Corresponding Author:
Sabina De Geest, PhD, RN
Institute of Nursing Science
Department Public Health
University of Basel
Bernoullistrasse 28
Basel, 4056
Switzerland
Phone: 41 61 207 09 16
Fax: 41 61 207 09 55
Email: sabina.degeest@unibas.ch

Abstract

Background: The field of eHealth is growing rapidly and chaotically. Health care professionals need guidance on reviewing and assessing health-related smartphone apps to propose appropriate ones to their patients. However, to date, no framework or evaluation tool fulfills this purpose.

Objective: Before developing a tool to help health care professionals assess and recommend apps to their patients, we aimed to create an overview of published criteria to describe and evaluate health apps.

Methods: We conducted a systematic review to identify existing criteria for eHealth smartphone app evaluation. Relevant databases and trial registers were queried for articles. Articles were included that (1) described tools, guidelines, dimensions, or criteria to evaluate apps, (2) were available in full text, and (3) were written in English, French, German, Italian, Portuguese, or Spanish. We proposed a conceptual framework for app evaluation based on the dimensions reported in the selected articles. This was revised iteratively in discussion rounds with international stakeholders. The conceptual framework was used to synthesize the reported evaluation criteria. The list of criteria was discussed and refined by the research team.

Results: Screening of 1258 articles yielded 128 (10.17%) that met the inclusion criteria. Of these 128 articles, 30 (23.4%) reported the use of self-developed criteria and described their development processes incompletely. Although 43 evaluation instruments were used only once, 6 were used in multiple studies. Most articles (83/128, 64.8%) did not report following theoretical guidelines; those that did noted 37 theoretical frameworks. On the basis of the selected articles, we proposed a conceptual framework to explore 6 app evaluation dimensions: context, stakeholder involvement, features and requirements, development processes, implementation, and evaluation. After standardizing the definitions, we identified 205 distinct criteria. No criteria had to be moved between dimensions.

Conclusions: This study provides a comprehensive overview of criteria currently used in clinical practice to describe and evaluate apps. This is necessary as no reviewed criteria sets were inclusive, and none included consistent definitions and terminology.
Although the resulting overview is impractical for use in clinical practice in its current form, it confirms the need to craft it into a purpose-built, theory-driven tool. Therefore, in a subsequent step, based on our current criteria set, we plan to construct an app evaluation tool with 2 parts: a short section (including 1-3 questions/dimension) to quickly disqualify clearly unsuitable apps and a longer one to investigate more likely candidates in closer detail. We will use a Delphi consensus-building process and develop a user manual to prepare for this undertaking.

**Trial Registration:** PROSPERO International Prospective Register of Systematic Reviews CRD42021227064; https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42021227064

**KEYWORDS**
telemedicine; smartphone; mobile apps; program evaluation; decision-making; systematic review; mobile phone

**Introduction**

**Background**

eHealth, that is, “the use of information and communication technology for health” [1], can support the delivery of interventions for self-management support and behavior change in patients with acute and chronic illnesses [2,3]. According to the World Health Organization (WHO) [4], self-care health interventions can be classified into individual agency (eg, promoting awareness about self-care), health information seeking (eg, education for informed health decision-making), social and community support (eg, peer mentorship and counseling), personal health tracking (eg, home-based records for health and diagnostic data), self-diagnosis of health conditions (eg, self-testing), self-management of health (eg, self-medication or treatment), individuals’ links to their health systems (eg, individuals sharing data with health care professionals [HCPs]), and individuals’ financial outlays for health (eg, expenses for prescription medicines). However, a recent evaluation of self-care interventions delivered via eHealth apps noted that only 20% of the 100 included apps used evidence-based information, whereas experienced HCPs considered only 32% to be useful and deemed 52% to be misleading and 11% to be dangerous [5].

Searching for a common characteristic linking the most effective apps, several systematic reviews and meta-analyses have found that those developed on firm theoretical foundations are more likely to be effective [6,7]. However, a systematic review of health-promoting smartphone apps found that only 55.6% of the included 27 studies described a theoretical basis for their smartphone app development [8]. A 2018 review found that only 8 (1.2%) of 681 smartphone apps to support medication adherence had documented evidence of their effectiveness. Such evidence is health care systems’ main consideration regarding certification and reimbursement [9]. Furthermore, although a user-centered design (sometimes also called human-centered design) and the involvement of patients and HCPs in the development of smartphone apps is known to provide insights into end users’ needs and helps ensure both relevant, reliable content and high quality [9,10], only 84 (12.3%) of the apps in this review had been developed in collaboration with HCPs. None reported patient involvement in their development processes [9].

Owing to the increasing availability of eHealth smartphone apps, it is vitally important and increasingly challenging for HCPs to identify, evaluate, and recommend relevant, trustworthy, and high-quality eHealth smartphone apps [11-14].

One tempting way for HCPs to form a first idea of an eHealth smartphone app’s quality is the star ratings and written reviews it receives on an app store [15]. However, this information is often subjective and distorted by individuals, comes from unverified or fraudulent sources, or provides no insights on an eHealth app’s quality [15-18]. HCPs also face a lack of reliable guidance on evaluating eHealth smartphone apps’ applicability to clinical practice [13,19,20]. Therefore, many are now struggling to describe and evaluate eHealth smartphone apps. A guideline regarding their characteristics and quality using standardized methods that will allow HCPs to propose reliable apps to their patients is needed [21,22].

Previous efforts to evaluate apps have generally focused on guidance for researchers [23-25]. Although the criteria were often overly complex or tailored to specific health areas, they also tended to be incomplete. Their underlying theories, scientific rationales, and development processes have rarely been described [20,25-27]. Furthermore, their unsuitable foci, nontransparent development processes, complexity, and often excessive time demands make them a poor fit for clinical practice. Finally, the existing instruments used a variety of criteria that only partially overlapped [20,23-27]. A clear description and evaluation of an app is important as, in rapidly evolving fields, even small changes or improvements to an app can have significant impacts on its use and usefulness [28]. To date, evaluation tools to help clinicians describe and evaluate eHealth apps, allowing them to recommend high-quality apps to their patients and share their thoughts using common terminology, are lacking [29,30].

**Objectives**

Therefore, the aim of this study was to obtain an overview of the evaluation criteria used in the literature. This process was conducted in three steps: (1) conducting a systematic review to identify existing criteria for evaluating eHealth apps, (2) developing a conceptual framework for the evaluation of eHealth smartphone apps, and (3) developing a comprehensive list of criteria for describing and evaluating eHealth smartphone apps. This was the foundational phase 1 of an overarching project to develop and pilot-test a theory-based tool to help HCPs evaluate the characteristics and quality of eHealth smartphone apps in a practical and standardized way (Figure 1). Phase 2 will involve...
narrowing down, refining, and testing the evaluation criteria via 3 further steps: conducting a Delphi survey to narrow down the criteria (step 4), developing a user guide for the processes of description and evaluation (step 5), and pilot-testing the user guide and processes with HCPs (step 6). This paper focuses on reporting on the foundational phase 1 and outlining the proposed steps for phase 2.

Figure 1. Overview of the 2 phases and steps in the development of the eHealth smartphone app evaluation tool (the focus of this paper is framed on the left side).
Information Sources
We queried the MEDLINE (OvidSP), CENTRAL (via Cochrane), CINAHL (EBSCOhost), and Web of Science databases. Supplementary searches were conducted on trial registries (ClinicalTrials.gov and WHO trial registry) and the reference lists of the included papers. The search was conducted on December 5, 2022. No time restrictions were imposed.

Search Strategy
We developed our MEDLINE search string based on the terms used in articles on (partially) similar topics [34-41] combined with key Medical Subject Heading and free-text terms (see the search strategy in Multimedia Appendix 1). For the other databases, we adapted the search string accordingly. We combined thematic blocks with various keywords related to eHealth, smartphone, application, evaluation, and tool. No filters were applied.

Selection Process
All identified titles and abstracts were independently screened for relevance by 2 reviewers (JR and TH). The full texts were assessed by the same reviewers using the criteria described previously. The reasons for full-text exclusion were reported. In one case of disagreement, an independent third researcher (SDG) contributed to help reach a consensus.

Data Collection Process and Data Items
In total, 2 reviewers (JR and TH) independently extracted the data and cross-checked their results. We extracted information on the author, year, country, research question or study aim, design, operating system, population or specific condition, main intended intervention purpose, name of the tool, and framework or theoretical guidance. The intended purpose of each eHealth app–delivered intervention was categorized according to the WHO classification for self-care health interventions [4]: individual agency, health information seeking, social and community support, personal health tracking, self-diagnosis of health conditions, self-management of health, individuals’ links to their health systems, and individuals’ financial outlays for health. Specific eHealth app quality evaluation dimensions or criteria were extracted and tabulated in a separate table.

Study Assessment
The included studies were assessed using the Appraisal of Guidelines for Research and Evaluation–II (AGREE-II) instrument [42], which is widely used to evaluate guideline development processes. As many of the included studies did not a priori intend to develop an evaluation guideline, this instrument might not have been the best option for all studies. However, as we were mainly interested in the justification and development of the dimensions or criteria used in the studies, this instrument provided us with the best support for evaluating these aspects. The AGREE-II instrument consists of 23 items classified into 6 domains (3 items on scope and purpose; 3 on stakeholder involvement; 8 on rigor of development; 3 on clarity of presentation; 4 on applicability; and 2 on editorial independence, ie, whether funding body and competing interests were reported). We concluded our AGREE-II assessment by rating the degree to which each included study described each domain. For this, we used a 4-descriptor scale: accurately (all AGREE-II items fulfilled), partially (two-thirds of all AGREE-II items fulfilled), hardly (one-third of all AGREE-II items fulfilled), and not at all (0 AGREE-II items fulfilled).

Step 2: Development of a Conceptual Framework
The original dimensions of the frameworks reported in the selected articles were listed in a table. Similar descriptions of dimensions were merged. The first draft of the proposed conceptual framework and graphical representations was reviewed and discussed with various stakeholders (researchers, clinicians, designers, and software developers with backgrounds in nursing, medicine, ethics, and informatics). During these discussions, the participants recommended that we distinguish between technical dimensions (eg, design, usability, security, safety, and privacy) and those that focused on content (eg, evidence base and scientific evaluation). It was also recommended that the dimensions be presented as a linear, step-by-step process. During these meetings, the first author (JR) took notes and recorded the proposed changes until consensus was reached on the next version.

The second draft of the framework was discussed with 18 international volunteers (patient representatives, researchers, clinicians, and technology developers) from diverse backgrounds in health care (eg, psychology, nursing, and pharmacy) who were participating in a public webinar on quality evaluation of eHealth technology. As it was a public webinar, only limited data on participant demographics were collected (Table S1 in Multimedia Appendix 2). A survey via AhaSlides (AhaSlides Pte Ltd) to rate the importance and clarity of the dimensions and subgroups (1=not important; 5=very important) and open-ended questions to add missing dimensions or subgroups were used to engage with the participants. In addition, the participants were engaged to add comments orally that the first author (JR) put in writing. The quantitative data were analyzed descriptively (eg, frequency and mean), whereas the qualitative data were analyzed using the mind-mapping technique. The participants found the technology dimension too large (ie, covering too many subtopics) and partially unclear. Therefore, it was recommended to split this dimension into technological concerns (eg, technical requirements, security, safety, and privacy) and functional requirements (eg, the user-centeredness and usability of the design). In addition, they understood eHealth evaluation as a cyclical process and recommended presenting the conceptual framework as a continuous circuit as opposed to the initially linear process recommended previously.

Their feedback was used to draft a third version of the framework, which was presented and discussed with 34 researchers, clinicians, and technology developers (Table S2 in Multimedia Appendix 2; only 1 person overlaps with the volunteers from the webinar) mainly with a background in pharmacy who participated in the Next Chapter in Patient Care Conference in April 2022 in Pärnu, Estonia. This time, a survey via Mentimeter (Mentimeter AB) was used to rate the importance and clarity of the dimensions and subgroups (1=not important; 5=very important), and open-ended questions were used to add missing dimensions or subgroups. In addition, the participants were engaged to add comments orally while the
first author (JR) took notes. Quantitative data were analyzed descriptively, whereas qualitative data were analyzed using the mind-mapping technique. The Next Chapter in Patient Care participants recommended highlighting the overarching nature of the ethical, legal, and social aspects, which must be considered in all eHealth smartphone app evaluation dimensions. This resulted in a fourth version reflecting the general character of relevant ethical, legal, and social considerations.

Subsequent rounds of discussion and feedback with the research team focused on the scope and relationships between the dimensions. The participants highlighted the legal, ethical, and social aspects to be treated as part of the context. In addition, they agreed that stakeholder involvement should be seen as another overarching aspect that is important in all dimensions of eHealth app evaluation. This discussion resulted in the final version of our new conceptual framework for evaluating eHealth apps: the eHealth Smartphone App Evaluation (eHAPPI) framework.

**Step 3: Development of a Comprehensive Criteria List**

The eHAPPI framework was then used to synthesize all the eHealth smartphone app evaluation criteria identified in the selected studies. The redundant criteria were combined. The classification of the criteria according to the eHAPPI dimensions and suggestions for changes and regarding additional or irrelevant criteria were discussed and refined by the research team according to consensus.

**Results**

**Step 1: Systematic Review**

**Study Selection**

The results of our study selection process are presented in the PRISMA flow diagram (Figure 2). The search strategy described previously yielded 1021 nonduplicate titles. After screening of the titles and abstracts as well as full-text assessment for eligibility, our final analysis included 128 articles that met all the inclusion criteria (Multimedia Appendix 3 [43]).

**Figure 2.** PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

**Study Characteristics**

**General Description**

A detailed description of the characteristics of the included studies can be found in Multimedia Appendix 4. The years of publication ranged from 2013 to 2022. All but 1.6% (2/128) of the articles were written in English—1 was in French [44] and the other in Spanish [45]. Four-fifths of the studies (101/128, 78.9%) were conducted in Western Europe, North America, or Australia. The studies used a variety of designs, mainly cross-sectional reviews of existing apps (42/128, 32.8%); reports on various app development and evaluation approaches (29/128, 22.7%); quantitative, qualitative cross-sectional, or longitudinal user testing of a single app (27/128, 21.1%); or different forms of reviews (15/128, 11.7%). The 128 included studies involved apps covering 30 topics, such as mental health, health promotion, or support for specific physical conditions (eg, heart disease and diabetes). However, only 50.8% (65/128) of the studies provided enough detail to categorize them according to
the WHO classification for self-care interventions for health and well-being [4].

**Evaluation Tools**

In total, the included studies used 142 distinct tools to evaluate eHealth apps. Although 76.6% (98/128) of the included studies used a single evaluation tool, 23.4% (30/128) used multiple tools and scales. Almost one-quarter of the studies (30/128, 23.4%) used an evaluation tool with investigator-developed criteria and then provided only scantily described development processes or theoretical backgrounds for those criteria. The most frequently used tool was the Mobile App Rating Scale (33/128, 25.8%) followed by its adapted versions (8/128, 6.3%) and the System Usability Scale (22/128, 17.2%). Less frequently used tools were the Post-Study System Usability Questionnaire (2/128, 1.6%), the Questionnaire for User Interaction Satisfaction (2/128, 1.6%), and the quality of experience survey (2/128, 1.6%). A total of 43 other tools were used in only 0.8% (1/128) of the studies each. In total, 10.2% (13/128) of the studies used qualitative methods (eg, interviews and focus groups) to generate the app evaluation criteria. In some cases (4/128, 3.1%), the origin of the criteria was unclear, or similar names were used for different tools.

**Theoretical Frameworks**

Most studies (83/128, 64.8%) did not clearly report a theoretical underpinning. The 32% (41/128) that did used 59 different frameworks, including various non–eHealth-specific behavioral, social, or implementation theories (10/59, 17%), the technology acceptance model (7/59, 12%), heuristic evaluation (5/59, 8%), models of the International Organization for Standardization (3/59, 5%), the (extended) Unified Theory of Acceptance and Use of Technology (3/59, 5%), or user-centered design (2/59, 3%). In total, 29 frameworks were used in only 2% (1/59) of the studies each. Of the 59 frameworks used in the included studies, 16 (27%) guided the development and 43 (73%) guided the evaluation of eHealth smartphone apps. In 1.6% (2/128) of the studies, different frameworks were used for development and evaluation.

**Study Assessment**

On average, the studies described 8.6 (SD 2.4; range 3–15) of the AGREE-II instrument’s 23 items [42]. Few studies described the scientific or theoretical basis and development processes of the app evaluation criteria that they applied. The most completely described or justified domains were related to scope and purpose (116/128, 90.6% described it accurately; 11/128, 8.6% described it partially; and 1/128, 0.8% hardly described it), editorial independence (91/128, 71.1% described it accurately; 30/128, 23.4% described it partially; and 7/128, 5.5% did not describe it at all), and stakeholder involvement (13/128, 10.2% described it accurately; 42/128, 32.8% described it partially; 70/128, 54.7% hardly described it; and 3/128, 2.3% did not describe it at all). The least fulfilled domains were applicability (1/128, 0.8% described it accurately; 2/128, 1.6% described it partially; 22/128, 17.2% hardly described it; and 103/128, 80.5% did not describe it at all), rigor of development (13/128, 10.2% described it partially; 53/128, 41.4% hardly described it; and 62/128, 48.4% did not describe it at all), and clarity of presentation (14/128, 10.9% described it accurately; 32/128, 25% described it partially; 52/128, 40.6% hardly described it; and 30/128, 23.4% did not describe it at all).

**Step 2: Development of a Conceptual Framework**

The full list of frameworks and original eHealth evaluation dimensions identified in the selected studies can be found in Multimedia Appendix 5 [43,45-47]. Some dimensions were included in only a few frameworks, and no framework included all possible dimensions.

The condensed dimensions were presented graphically and refined iteratively with the stakeholders until consensus was reached and no further adaptions were needed. The final eHAPPI conceptual framework (Figure 3) consists of six interrelated dimensions: (1) context, (2) stakeholder involvement, (3) development processes, (4) implementation, (5) evaluation, and (6) features and requirements.

A detailed definition of each dimension, including the subgroups, is presented in Textbox 1. Context describes a set of unique factors and conditions in which the app will be implemented [48]. This is an overarching dimension that depends on and, in turn, influences the other domains. Stakeholder involvement is essential in all aspects of eHealth. It involves the active engagement of relevant partners in all processes of the app life cycle, from conceptualization to sustainable implementation (eg, with end users, HCPs, researchers, and health insurers) [49,50].

Several subgroups were defined to further outline and structure the framework: basic information (concerning the app) and ethical, legal, and social aspects were seen as relevant subgroups of the context dimension. The featuresandrequirements dimension was assigned 4 subgroups: evidence-based content; functionality; usability, privacy, and security; and performance. Similarly, the development process dimension was divided into cocreation/user-centered design and characteristics of the development team, and adoption (integration into daily life) and maintenance were seen as relevant subgroups of the implementation dimension. Finally, the evaluation dimension included only 1 subgroup: scientific evaluation.
Figure 3. The final eHealth Smartphone App Evaluation framework.
Textbox 1. Overview of framework dimensions, subgroups, and definitions.

Context
- **Context** describes the set of unique factors and conditions in which the app is implemented. During the implementation process, the app, implementation, and context interact, influencing, modifying, facilitating, or hindering one another. As an overarching concept, context encompasses a physical location and roles, relationships, and interactions at various levels (ethical, legal, social, geographical, epidemiological, and political) [48,51].
- **Basic information** describes general information concerning the app (eg, name, URL, or available languages).
- **Ethical, legal, and social aspects** describe activities to help evaluators consider ethical, legislative, and social implications of an app’s development and implementation. These aspects cannot be completely separated from each other as they overlap somewhat [28].

**Ethics** describes activities to understand and explore the moral life, wherein morality includes beliefs, norms of behaviors, principles, and rules that guide individual and institutional behavior. Morality is a widely shared set of norms that result from a certain social consensus. Ethical aspects include the prevailing moral norms and values during the development and implementation of apps. Although moral values are influenced by cultural, sociopolitical, religious, and economic differences, many ethical reflections are shared by all countries and societies. Regarding apps, important ethical topics include benefit-harm balance, autonomy, respect for persons, and justice and equity and legislation (when using the app, ethical challenges may arise that were not considered in existing legislation and regulations) [28].

**Legal aspects** include rules and regulations that must be considered when developing and implementing apps. Rules and regulations are designed to protect the rights and interests of the patients and other members of society (eg, legislation on patients’ rights; data protection laws; or the regulations, rights, and obligations of health care professionals [HCPs] in general). Important legal topics are, for example, autonomy of the patient (eg, legal requirements), privacy of the patient (eg, no use of “unnecessary” patient information), equality in health care, ethical aspects (eg, impact on achievement of human rights), authorization and safety (eg, laws and rules regarding safety), ownership and liability, and regulation of the market [28].

**Social aspects** describe human-centered activities that approach end users and their social networks as reference points in an app’s development and implementation. These include groups of patients or individuals who might require special consideration (eg, vulnerable populations, people living in remote communities, people with learning difficulties, older people, ethnic minority individuals, and immigrants). Patients’, individuals’, and relatives’ perspectives should be considered when developing and implementing apps. Some social groups may be important for a particular app (eg, ethnic minority individuals and individuals with disabilities), which should be specified. Important social topics include patients’ perspectives (eg, expectations and wishes), social group aspects (accessibility), and communication aspects (eg, explanation of treatment choices) [28].

Stakeholder involvement
- **Stakeholder involvement** describes the active engagement of relevant partners in all processes of the smartphone app life cycle, from conceptualization to sustainable implementation (eg, end users, HCPs, researchers, and health insurers [49,50]).

Note: for end-user involvement in the development process, see the Cocreation/user-centered design section under the Development process subheading in this textbox.

Development process
- **Development process** describes all activities performed regarding the app’s creation, such as activities related to its aim, target population, guiding principles, applied approaches, stakeholder contributions, required changes for subgroups, or continuing uncertainties [52-54].
- **Cocreation of technologies or user-centered design:** cocreation describes a stepwise process in which partners (eg, patients, caregivers, and HCPs) are actively involved in the strategic design and planning of the app’s development and implementation at every stage of the process [55].

**User-centered design** is a multidisciplinary design approach using cyclic iteration and evaluation. As it actively involves users to improve the developers’ understanding of their requirements and wishes, this approach is seen as essential regarding product benefits and usability [56].

- **Characteristics of the development team** describes information about the multidisciplinary group responsible for the creation of the working, validated app. Important characteristics of the development team include the name of the app provider (developer and manufacturer of technology and content), contact details, organization attributes, and funding, as well as any conflicts of interest [57].

Evaluation
- **Evaluation** describes the assessment of the app’s efficacy, effectiveness, cost-effectiveness, safety, implementation, and impact [58].
- **Scientific evaluation** is the systematic assessment of the app’s efficacy, cost-effectiveness, safety, implementation, and impact through observation, measurement, and experimentation in a scientific study. Such an evaluation is essential to reliably measure an app’s effects and outcomes as a basis for decision-making [58-60].

Implementation
- **Implementation** describes the uptake and sustainable integration of evidence-based innovations such as apps into routine use [61].
- **Adoption** (ie, integration into daily life) deals with the app’s uptake (ie, activities focused on using the app in everyday life). This includes the user’s reaction to the app with respect to desired activities and interactions, such as downloads, clicks, and data entries. User engagement, that is, the user’s investment in learning about and participating in the app (eg, time and memory load), is an important prerequisite for the app’s long-term integration into daily life [62-64].
•  **Maintenance** covers activities that ensure that an app is sustained in an acceptable and safe condition to perform its specified functions. Maintenance also includes repair as a stand-alone action to restore deteriorated or damaged parts. The activities needed to ensure the required performance of the app during its lifetime include updates, performance evaluations, and the planning and execution of necessary repairs [65].

**Features and requirements**

•  **Features and requirements** describe the app and the characteristics or attributes it must provide to meet the users’ needs (ie, its information content, functionality, usability, security, privacy, and performance). The indications, performance features, and app options may differ for different generations or versions of an app. Important features and requirements topics include app description, claimed benefits, care level of use, reference values, or cutoff points used [28, 66]. A clear description of the app is important as, in rapidly evolving fields, even small changes or improvements to an app can have large impacts on its use and benefits. The description should enable those who are not familiar with the app to understand what it does, how it works, and how it can be used.

•  **Evidence-based content** describes information, subject matter, and data content made available by the app based on the best available scientific knowledge and clinical expertise [67].

•  **Functionality** describes the properties or functional requirements that affect the app’s use (ie, its features, components, and usefulness). In the context of self-monitoring, there are specific requirements related to design issues, particularly those concerning wearable technologies, as well as requirements regarding the behavioral aspects that are part of the design. These requirements need to be considered when evaluating a product or system for self-monitoring. The functionality information is intended to summarize the app’s overall suitability for use in a particular situation [38, 88, 69].

•  **Usability, privacy, and security** are nonfunctional requirements (ie, they deal with how the system should operate in terms of usability, security, privacy, design, modularity, modifiability, reliability, availability, portability, and operability). **Usability** describes the extent to which an app can be used by its target users to achieve its intended goals effectively, efficiently, and satisfactorily. Ideally, an app should be easy to use, easy to learn, and easy to understand [70]. **Privacy** describes the protection of and control over personal data during the app’s processing operations [71]. **Security** describes the provision of safeguards that serve the security of the individual or community. For example, this helps prevent or avoid poverty, hardship, theft, or espionage. Security is a major component of a stable, relatively predictable environment in which people can pursue their goals without interference, harm, or the fear of them. This includes protecting the app and associated data from events that could cause loss or severe damage, such as fire, burglary, theft, or vandalism [72, 73].

•  **Performance** describes whether an app works quickly and without errors and does not cause problems. Important performance topics include reliability and scalability (ie, whether an app still works properly when the number of users increases [74, 75]).

**Step 3: Development of a Comprehensive Criteria List**

In total, 205 criteria for describing and evaluating eHealth smartphone apps were reported in the selected articles. None of the articles included all the possible criteria. The use of terminology differed within the publications or was attributed different meanings. For example, several were named usability but referred to different aspects of that criterion, such as ease of use, usefulness, or speed. Others, such as research-backed, scientific references, information accuracy, and information quality, were named differently but clearly referred to a single criterion—in this case, evidence-based information.

Using the new eHAPPI framework, it is clear that most studies reported criteria that focused on usability, evidence-based content, functionality, or scientific evaluation. However, only 1 reported criterion dealt with context, and only 5 dealt with development processes.

The research team’s discussions emphasized that some criteria are objective (eg, average rating in the app store and purpose of the app). In contrast, others are more subjective (eg, matching the needs of the target population and intention to use). In addition, several were dependent on the tested app’s purpose or content (eg, whether an app community exists and features to support behavior change).

The research team agreed on how all the criteria were classified into dimensions and how most were formulated. However, to improve understanding or align with known formulations, they suggested revisions to the wording of 12 criteria [28]. On the basis of the research team’s recommendations (Multimedia Appendix 6 [25, 45-47, 76-85]), 11 new criteria were added (Multimedia Appendix 7 [28]). These additions were mainly to the ethical, legal, and social aspects section. No criteria were deemed completely irrelevant, so none were removed. Finally, the research team agreed to classify these added criteria into the existing dimensions.

The final list contained 216 criteria for describing and evaluating eHealth smartphone apps (Multimedia Appendix 8). Although this list was comprehensive, our discussions highlighted that it was not practical for use by HCPs in clinical practice. The research team agreed that, in the next phase, a short version with only 1 to 3 items per dimension would be useful to make a quick initial decision (ie, acting as an algorithm to gauge whether an app should be given further consideration). Only if an app passed this pretest would it undergo a more thorough evaluation using detailed criteria and offering nuanced results. In addition, the research team recommended a specific algorithm with thresholds that could be adjusted depending on each tested app’s purpose and context.

The research team recommended that the next step for the project’s second phase be a Delphi process to condense the list and develop the proposed decision support tool. This process will have two aims: (1) to provide a means to reach a consensus and (2) to develop a useful and feasible (ie, practical for use in...
clinical practice) tool to describe and evaluate eHealth smartphone apps.

Moreover, participants expressed concerns that the necessary information to complete the tool may be difficult to find. Therefore, the development of a user guide for HCPs on how to apply the tool and where to typically find the required information was also proposed for the next phase. Finally, the eHealth smartphone app evaluation tool and user guide will need to be pilot-tested with HCPs.

**Discussion**

**Principal Findings**

Although the eHealth field is rapidly expanding and evolving, there is no consensus on how the quality of eHealth apps should be defined and evaluated by HCPs [13,19,20,86]. In this paper, we described how we developed a comprehensive list of criteria to evaluate eHealth apps. We used a meticulous methodological approach to derive the list, consisting of a systematic literature review and iterative rounds with stakeholders from various backgrounds to compose a comprehensive framework—the eHAPPI—and use it to synthesize all the criteria identified in the selected studies. We found 6 overarching dimensions (ie, context, stakeholder involvement, features and requirements, development processes, implementation, and evaluation) of eHealth app evaluation and 205 criteria in the literature. A research team discussion resulted in 11 additional criteria, bringing the new total to 216. Using this comprehensive list, HCPs would be able to evaluate eHealth apps designed for diverse health care needs. No original studies included all the dimensions or all the criteria. Most of the selected studies (83/128, 64.8%) did not describe an underlying framework or theoretical guidance or how their criteria were developed. In addition, a general lack of common terminology among the included studies further complicated efforts toward comparison and direct application in clinical practice.

Although the proposed complete list of criteria was comprehensive, applicable to diverse health care needs, and applied consistent terminology, its length and depth limited the feasibility of using it in clinical practice. As a stand-alone reference, this list would be better suited for comprehensive evaluations of a broad range of eHealth apps by a regulatory body. Therefore, we proposed a 2-step approach to developing these criteria for use by HCPs. The first step would be defining a few critical initial decision criteria. A more detailed description and assessment would be useful only for positive evaluations using these first criteria. Using such an algorithm would allow the most obviously inferior apps to be quickly excluded from the process.

However, our progress in this direction was slowed by a lack of consensus regarding which criteria were essential and which would simply be nice to have [23]. As a compromise, we agreed that developing the proposed algorithm would require more expertise than we had and further expert discussion. Therefore, for the second phase of this project, we proposed a Delphi process [87] to guide the further development and fine-tuning of the final evaluation tool.

Although we did not reach a consensus on which criteria qualify as essential, our group discussions provided insights into which qualities to consider in describing and evaluating apps as well as how an optimal tool might be structured. For example, given the dynamic and rapidly evolving use of apps in clinical practice [29,88], flexibility is a significant concern [23,25]. Therefore, the research team recommended an algorithm not only whose cutoff criteria can be modified to fit each evaluated app’s purpose and context but also whose overall functionalities can evolve alongside the surrounding technology [89].

Another point of discussion concerned the difficulty we encountered in finding the information to complete this evaluation tool. HCPs who are less familiar with eHealth apps may have difficulty gathering even basic data, such as the name of an app’s developer or its latest update [13,21,22]. More technologically savvy individuals may have trouble finding information on that app’s scientific basis, how or whether its development processes included stakeholder involvement, or which strategies were used in its implementation. Therefore, we suggest that the proposed tool include a user guide describing why such criteria are important and where to find and how to judge the required information. This echoes a recommendation by the European Network to Advance Best Practices and Technology on Medication Adherence in their Cooperation in Science and Technology Action (CA19132), which facilitates the use of a web-based repository of information on medication adherence technology [90,91]. Although eHealth app developers clearly need to provide relevant details in a clear and easily accessible way, health educators also need to include eHealth evaluation in HCP education and training curricula.

The previous discussion provides the foundation for conducting phase 2 of this study. This phase has three goals: (1) to conduct a Delphi survey to narrow down the number of criteria and develop an algorithm for initial decision-making, (2) to develop a user guide, and (3) to pilot-test the resulting iteration.

**Limitations**

This study has several limitations. Most notably, at this point in the project, although our list of eHealth app description and evaluation criteria is comprehensive, it remains a preliminary version. That is, despite discussions with various interdisciplinary experts, phase 1 did not produce an in-depth consensus on the essential criteria for a richly detailed but broadly feasible means of evaluating eHealth apps. This drove the decision to design a 2-phase project. In phase 2, which will be a Delphi study [87], we aim to develop a criteria-sorting algorithm. With this in place, the phase will culminate in a version of an eHealth app evaluation tool for pilot-testing.

In addition, all the included studies were assessed using the AGREE-II instrument [42], which was specifically designed to assess clinical practice guideline development reports. Considering the high level of heterogeneity across many of the study characteristics, direct comparability using a single tool was limited. However, although other instruments would have been more suitable in many cases, using various instruments would have yielded equally varied results. As we were primarily interested in the rationales and development processes that...
supported the dimensions and criteria, the AGREE-II scales provided a consistent assessment of these aspects.

Finally, the participants in the discussion rounds focused on the development of the conceptual framework for their critical feedback in developing the study methods and manuscript, as well as to Chris Shultis for his contribution to the development of the conceptual framework and criteria. Special thanks also go to the participants of the Research Round Table paper for publication. The authors thank all the participants in the various discussion rounds for their reflections and input for their participation in the development of the study methods and manuscript, as well as to Chris Shultis for his critical feedback in developing the study methods and manuscript. The authors thank all the participants in the various discussion rounds for their reflections and input for their participation in the development of the study methods and manuscript.

**Comparison With Prior Work**

Although 23 of the existing tools were explicitly intended for HCPs [14,43,46,47,76-86,93-100], none of these were complete; rather, they were too focused on specific conditions, or their theoretical justifications or development processes were not described. Such omissions make it difficult for HCPs in clinical practice to comprehensively but feasibly describe and evaluate eHealth apps in a standardized way to guide the recommendation of relevant, reliable, and high-quality apps to their patients [11,13,14]. Other studies supplemented the dimensions and criteria for describing and evaluating apps. However, it remains unclear which criteria are essential and how detailed they need to be. Recently, there has been much discussion about how to define and evaluate eHealth quality and what criteria are needed for an app to be used in the health care system [23,89,101]. Future findings from the planned phase 2 will likely provide a basis for further discussion on this topic among app developers or providers, HCPs, patients, researchers, and policy makers.

Our first comprehensive list of criteria as a result of phase 1 provides an excellent basis for the next steps in phase 2 to develop a new eHealth app evaluation tool. The need for a tool to describe and evaluate eHealth apps and help HCPs and their patients navigate the digital health ecosystem is pressing [23]. Our path to a proposed resolution has been quite complex as this field is also complex. After listing the criteria identified via a literature review, we developed them through expert discussions, revealing important improvement areas. In particular, compared with recommendations from the Health Technology Assessment Core Model [28], the criteria concerning the ethical, legal, and social aspects of eHealth apps were deemed incomplete. Therefore, in addition to adapting many criteria, we added several.

**Contribution of This Study**

This study focuses on addressing the rapidly growing and somewhat chaotic field of eHealth, particularly the challenges faced by HCPs when it comes to evaluating and recommending health-related smartphone apps to their patients. This study’s contribution lies in its comprehensive methodology for gathering and categorizing existing criteria for evaluating health apps, which is essential as no single framework or evaluation tool effectively serves this purpose. The methodology involved a systematic review of the literature, which resulted in the identification of 216 distinct evaluation criteria organized within a conceptual framework comprising 6 app evaluation dimensions. These dimensions encompass various aspects, including the app’s context, stakeholder involvement in its development, features, development processes, implementation, and evaluation. This study highlights the need for a more purpose-built, theory-driven tool to help HCPs assess and recommend apps effectively and outlines plans to create a 2-part app evaluation tool based on the gathered criteria, which will expedite the process of qualifying unsuitable apps and scrutinizing potential candidates more closely. This study serves as a crucial foundational step toward developing a practical tool that can guide HCPs in evaluating and recommending health-related apps.

**Conclusions**

Developing a tool comprehensive enough for HCPs to reliably describe and evaluate the full range of eHealth apps yet short enough to be feasible for daily clinical practice is a daunting challenge. After our literature review yielded a list of criteria too bulky for routine use, there was a lack of consensus either on terminology or on relevance to define and evaluate app quality. In this report of phase 1, we provided our initial comprehensive overview of 216 relevant criteria used in the selected studies to describe, evaluate, and recommend eHealth apps. To condense this list to a more manageable size, in phase 2, we will formulate and apply a robust consensus-building process to generate a list of criteria ranked by importance, followed by the creation of an algorithm to produce short- and long-form evaluations to match the characteristics of the apps to be evaluated. In addition, the development of a user guide and pilot-testing of the tool are planned. As a basis for informed guidance and decision-making, such a tool will help HCPs reliably describe and evaluate eHealth apps for their patients.

**Acknowledgments**

This work was conducted mainly by a PhD student as part of her PhD project funded by the University of Basel. The funders had no role in the study design; collection, analysis, and interpretation of the data; writing of the report; or decision to submit the paper for publication. The authors thank all the participants in the various discussion rounds for their reflections and input for the development of the conceptual framework and criteria. Special thanks also go to the participants of the Research Round Table Meeting seminar for their critical feedback in developing the study methods and manuscript, as well as to Chris Shultis for his language editing.
Authors' Contributions
The study was conceptualized by JR, SDG, and ADD. The literature search strategy was developed by JR, SDG, and FD. Articles were located, identified, and evaluated by JR and TH. Data were extracted and checked by these same authors. The initial dimension and criteria list was composed and discussed by all the authors. The manuscript was drafted and edited by JR with close revision and feedback from SDG as well as review and feedback from all other coauthors.

Conflicts of Interest
None declared.

Multimedia Appendix 1
MEDLINE search strategy.
[PDF File (Adobe PDF File), 153 KB - mhealth_v12i1e48625_app1.pdf ]

Multimedia Appendix 2
Demographics of participants.
[PDF File (Adobe PDF File), 184 KB - mhealth_v12i1e48625_app2.pdf ]

Multimedia Appendix 3
Included articles.
[PDF File (Adobe PDF File), 470 KB - mhealth_v12i1e48625_app3.pdf ]

Multimedia Appendix 4
Characteristics of the included studies.
[PDF File (Adobe PDF File), 289 KB - mhealth_v12i1e48625_app4.pdf ]

Multimedia Appendix 5
Original frameworks and dimensions.
[PDF File (Adobe PDF File), 282 KB - mhealth_v12i1e48625_app5.pdf ]

Multimedia Appendix 6
Revised criteria.
[DOCX File, 129 KB - mhealth_v12i1e48625_app6.docx ]

Multimedia Appendix 7
Additional criteria.
[DOCX File, 45 KB - mhealth_v12i1e48625_app7.docx ]

Multimedia Appendix 8
Complete criteria list.
[PDF File (Adobe PDF File), 376 KB - mhealth_v12i1e48625_app8.pdf ]

References


73. Cobb M. Physical security. Physical Security TechTarget. URL: https://searchsecurity.techtarget.com/definition/physical-security [accessed 2023-11-29]


77. Collins R. Nurses’ perceived usefulness of secure texting applications for the purpose of patient car. Online J Nurs Inform 2019;23(1) [FREE Full text]


Abbreviations

AGREE-II: Appraisal of Guidelines for Research and Evaluation–II

eHAPPI: eHealth Smartphone App Evaluation

HCP: health care professional

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

WHO: World Health Organization
Evaluation of Chinese HIV Mobile Apps by Researchers and Patients With HIV: Quality Evaluation Study

Peng Liu1,2, MS; Lingmeng Wang1, BS; Fuzhi Wang1,2, PhD

1School of Health Management, Bengbu Medical College, Bengbu, China
2Innovation Team of Health Information Management and Application Research, Bengbu Medical College, Bengbu, China

Corresponding Author:
Fuzhi Wang, PhD
School of Health Management
Bengbu Medical College
2600 Donghai Road
Bengbu, 233030
China
Phone: 86 0552 3173195
Email: wfz.bbmc@foxmail.com

Abstract

Background: Against the backdrop of globalization, China remains one of the most heavily burdened countries in Asia with regard to AIDS. However, many high-risk groups and patients affected by AIDS may be less likely to actively seek care from medical institutions because of fear of experiencing shame or discrimination. Mobile apps provide a promising avenue for supporting the prevention, diagnosis, and treatment of AIDS. However, a comprehensive systematic evaluation of these mobile apps’ functionality and quality has not been conducted yet.

Objective: This study aims to identify the available mobile apps for AIDS in China, assess and discuss the functional features and quality of these Chinese AIDS mobile apps, and offer decision support for patients and clinical practitioners in accessing high-quality AIDS mobile apps. Furthermore, based on the evaluation results, recommendations for improvement will be provided.

Methods: A systematic search was conducted on the Qimai app data platform, the Aladdin WeChat applet data platform, and WeChat to identify mobile apps related to AIDS. A snowball sampling method was used to supplement the potentially overlooked apps. The selected mobile apps underwent a rigorous screening process based on unified criteria. Subsequently, assessments were independently undertaken by 3 separate researchers and 2 patients with HIV, using both the Mobile App Rating Scale (MARS) and the User Mobile App Rating Scale (uMARS). Quantitative interpretations of the data were facilitated by the MedCalc statistical software (version 20.217, MedCalc Software).

Results: A total of 2901 AIDS mobile apps were included in the study, with 2897 identified through information retrieval and an additional 4 added via snowball sampling. After a rigorous selection process, 21 apps were determined to be usable. Among them, the Hong Feng Wan app achieved the highest combined average score, calculated based on the MARS (3.96, SD 0.33) and uMARS (4.47, SD 0.26). Overall, there was no significant correlation between MARS and uMARS ($r_{app quality total score}=0.41; P=.07; r_{subjective quality}=0.39; P=.08$). A notable issue was the widespread lack of user privacy protection, with only 24% (5/21) of the apps offering this feature.

Conclusions: The number of available Chinese AIDS mobile apps is limited, with WeChat applets dominating the market. Nonetheless, the performance of WeChat mini-apps is generally inferior to that of independent apps, and there may be significant discrepancies between assessments conducted by researchers and those provided by genuine end users, emphasizing the necessity of involving real users in the development and evaluation of HIV mobile apps. In addition, developers of these Chinese HIV mobile apps need to devote attention to improving privacy protection mechanisms, in addition to considering the evaluations of researchers and real users. This will help attract more users and increase user loyalty.

(JMIR Mhealth Uhealth 2024;12:e52573) doi:10.2196/52573

KEYWORDS
HIV; mobile app; evaluation; mobile phone
Introduction

Background

AIDS is a chronic infectious disease caused by HIV infection, resulting in high mortality rates. Since the first reported case of AIDS, it has rapidly spread worldwide, becoming a major public health concern. China, influenced by globalization, is no exception. According to the 2021 National Notifiable Infectious Diseases Summary released by the Chinese Center for Disease Control and Prevention, AIDS has the highest mortality rate and death toll, with 19,623 reported deaths. Available data indicate that as of 2022, the number of confirmed AIDS cases in China has exceeded 1.05 million and is projected to surpass 1.6 million infections by 2023. AIDS remains a significant infectious disease affecting public health and socioeconomic development in China [1].

For individuals who are at a high risk of HIV infection but have not yet been infected, pre-exposure prophylaxis (PrEP) is a highly effective method of preventing HIV through daily medication. However, the current use of and adherence to PrEP remain low. However, mobile apps have shown effectiveness in promoting PrEP use and adherence among high-risk populations [2]. Muessig et al [3] highlighted the advantages of internet and mobile-based interventions, which facilitate wider dissemination of PrEP at a lower cost compared with conventional methods. The MyChoices app developed by Biello KB demonstrated feasibility and potential in improving HIV PrEP use rates among gay and bisexual men in the United States [4]. The prevalence of HIV infections remains disproportionately high among gay, bisexual, and other men who have sex with men (GBMSM) [5]. With the widespread use and convenience of smartphones, sexual networking apps have replaced traditional dating websites as the primary online social platform for GBMSM. As such, apps have become crucial venues for sexual health research [6]. Targeted video and text-based sexually transmitted infection and HIV information provided through mobile apps has proven effective in reducing new infection rates among GBMSM [5].

For people living with HIV, the highly active antiretroviral therapy developed by scientist David Ho represents the most effective treatment method. Highly active antiretroviral therapy can suppress HIV replication and prevent the emergence of drug-resistant viruses. However, strict adherence is required, necessitating patients to adhere to prescribed medication schedules [7]. Despite simplified treatment regimens, adherence remains challenging for some people living with HIV [8]. Evidence suggests that HIV case management can improve treatment adherence and quality of life and reduce risky sexual behavior [9,10]. However, case managers face significant workloads and limited policy and funding support [11,12].

Mobile health (mHealth) apps have demonstrated potential in assisting people living with HIV with effective self-management and delivering personalized interventions. Schnall et al [13,14] identified ideal features for an HIV app, including reminders, health information delivery, medication logs, communication, settings, and search functions. However, comprehensive apps meeting these criteria are scarce in the current market, and there is a lack of rigorously evaluated mHealth apps specifically designed for people living with HIV [13,14]. Yang et al [15] found that most Chinese GBMSM apps, selected based on relaxed inclusion criteria, primarily focused on dating and lacked HIV prevention and health information.

For both high-risk populations and individuals infected with HIV, HIV testing plays a vital role in combating HIV. Early HIV testing allows individuals to learn about their infection status promptly, facilitating timely access to antiretroviral therapy and significantly reducing mortality rates. However, fear of shame or discrimination may discourage many high-risk populations and patients from actively seeking care at health care facilities [16-19]. With the increasing prevalence of mobile phones, various mHealth interventions have been developed to diversify HIV self-testing (HIVST) approaches, including telephone hotlines, SMS text messaging–based interventions, and internet-based interventions. These interventions have shown potential in improving testing rates, particularly among hard-to-reach populations [20-26]. Although these platforms have achieved varying levels of success, the use of mobile apps has emerged as a highly popular trend because of their flexibility and scalability. For instance, the mLab app serves to enhance users’ understanding of their HIV test results while facilitating their access to pertinent HIV information and services [27]. Another noteworthy app is Aspect HIVST, which offers an acceptable means of uploading mobile HIVST results and demographic information to a centralized database [28]. In addition, ApiDé serves as a multilingual electronic tool (app) that assists health care providers in offering and explaining HIV screening to immigrants facing language barriers [29].

In conclusion, the use of mobile apps is becoming increasingly prevalent in HIV/AIDS prevention and control. These apps provide a convenient and accessible means for high-risk individuals and patients to access information, consultations, and support related to HIV/AIDS. They serve as effective adjunct measures in improving antiretroviral therapy adherence and reducing AIDS incidence rates, thereby mitigating the currently imbalanced resource allocation between patients and health care providers involved in antiviral treatment. However, research on Chinese AIDS mobile medical apps is relatively scarce and late, and empirical studies evaluating the effectiveness of HIV mobile medical apps remain sparse. Therefore, the effectiveness of Chinese HIV mobile apps in meeting user needs requires further assessment. To evaluate the efficacy of AIDS mobile apps in China, the Mobile App Rating Scale (MARS) and the User Mobile App Rating Scale (uMARS) scales, which are widely accepted and applied for uses such as assessing chronic disease management [30], COVID-19 tracking [31], psychiatric interventions [32], physical exercise among older people [33], and menstrual monitoring [34], were incorporated into this study as standardized assessment tools for the quality of mobile apps.

Objectives

This study aims to achieve the following objectives:

1. To identify a comprehensive list of available Chinese mobile apps for HIV/AIDS.
2. To evaluate the functional features and quality of Chinese HIV/AIDS mobile apps from 2 distinct standpoints: those of researchers and those of individuals diagnosed with HIV, and to provide decision support for AIDS-related groups and health care professionals in accessing high-quality AIDS mobile apps.

3. To conduct a thorough analysis of the evaluation results and to provide improvement recommendations based on the findings, with the ultimate goal of enhancing the quality of Chinese HIV/AIDS mobile apps.

Methods

Search Strategy
Considering the use status of Chinese AIDS apps, a systematic search was conducted on the Qimai app data platform, Aladdin WeChat applet data platform, and WeChat from February 18 to 19, 2023, adhering to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The specific search strategies are as follows.

Data Collection of Apps
The Qimai platform, a well-known domestic mobile app data analytics platform, was used to retrieve AIDS-related apps. Qimai provides comprehensive data on iOS and Android app markets (including Huawei, Baidu, Xiaomi, Vivo, etc), along with app store optimization and app store search marketing optimization service tools and SearchAds data reports. It also offers professional data analysis and optimization strategies. The platform features an intelligent keyword expansion and association tool, which proved valuable for obtaining the required data for this study. In this research, the Qimai Keyword Expansion Assistant was used to search for AIDS-related terms. We selected keywords with a relevance score exceeding 50% and a search index surpassing 4605 (typically indicating a higher search frequency) to identify relevant apps that met the criteria. The search was conducted on February 18, 2023.

WeChat Applet Data Collection
WeChat applets, launched in China in 2017, are app programs developed on the WeChat platform. These applets offer users the convenience of accessing and using various functionalities and services directly within the WeChat app without requiring any separate downloads or installations. Known for their lightweight, fast, and user-friendly nature, they have gained popularity. To collect WeChat applet data, searches were performed on both the Aladdin Index and WeChat. The Aladdin Index, developed by Beijing Aladdin Future Technology Co, Ltd, serves as a ranking platform for the WeChat applet, providing a comprehensive reference for applet developers’ operations nationwide. For this study, keywords such as HIV, AIDS, Human Immunodeficiency Virus, Acquired Immunodeficiency Syndrome, as well as their corresponding Chinese keywords with similar meanings, such as huo-de-xing-mian-yi-que-xian-zong-he-zheng, ai-zi, ai-zi-bing, and ai-si-bing, were used to search for the WeChat applet related to AIDS in both the Aladdin applet list and WeChat. The search was conducted on February 19, 2023.

Supplementary Method
In collaboration with the HIV prevention and control team at the local Centers for Disease Control and Prevention (CDC), a snowball sampling method was used to distribute electronic questionnaires to individuals living with AIDS during their follow-up visits conducted by the CDC staff. The questionnaires (without personal information and providing adequate privacy protection) aimed to inquire about the HIV/AIDS apps they had used. Participants were encouraged to provide feedback using the Wenjuanxing platform and to share the electronic questionnaire with other individuals living with AIDS they knew. This supplementary method aimed to identify AIDS-related apps that were widely used within the AIDS community but were not captured by the aforementioned search strategies. The feedback collection period was extended from February 20 to May 19.

The exclusion criteria for apps were as follows: (1) nonsimplified Chinese language, (2) games, (3) apps with excessively limited functionality, (4) mobile apps that did not primarily focus on HIV or AIDS-related content, (5) >50% of the content is inaccessible, (6) apps that had terminated their services, (7) apps restricted to internal use by specific personnel (apps had to be accessible and functional for evaluation purposes), and (8) apps predominantly intended for advertising and product sales.

Evaluation Tool
In this study, we selected the MARS and uMARS scales as the evaluation tools for assessing the quality of the apps. These scales were chosen for 2 main reasons. First, these scales have been extensively validated and have demonstrated good reliability and validity across different contexts. Second, in our study, the included apps were a mix of independent apps and WeChat applets, with the majority being WeChat applets rather than independent apps. However, the MARS and uMARS scales offer broader applicability in this regard.

Developed by Zelmer et al [35], the MARS scale provides a standardized set of criteria for evaluating mobile apps in terms of engagement, functionality, esthetics, and information quality. The MARS scale has been widely used in user research, mobile app evaluations, and related fields. Previous studies have indicated that the MARS scale demonstrates good validity and reliability, making it a reliable measurement tool [30-34]. The uMARS scale, developed by Manning et al [36], serves as an extension of the original MARS scale, specifically designed to prioritize the assessments of average users toward mobile apps.

Evaluation Process
The evaluation was conducted by 3 researchers. Rater 1 (PL) holds a master’s degree in medical informatics and has extensive experience in medical information analysis. Rater 2 (Bin Li) holds a master’s degree in computer science and has abundant software development experience. Rater 3 (FW), an esteemed scholar in medical informatics in China, possesses a master’s degree in computer science and a PhD in social medicine. Being an expert in app quality evaluation, rater 3 boasts rich experience in software development experience. Being an expert in mobile information analysis, rater 3 boasts rich experience in mobile app evaluation tools. These raters, with an average rating of 3.67 (SD 0.65), offer broader applicability in this regard.

The evaluation was conducted by 3 researchers. Rater 1 (PL) holds a master’s degree in medical informatics and has extensive experience in medical information analysis. Rater 2 (Bin Li) holds a master’s degree in computer science and has abundant software development experience. Rater 3 (FW), an esteemed scholar in medical informatics in China, possesses a master’s degree in computer science and a PhD in social medicine. Being an expert in app quality evaluation, rater 3 boasts rich experience in mobile app evaluation tools. These raters, with an average rating of 3.67 (SD 0.65), offer broader applicability in this regard.

https://mhealth.jmir.org/2024/1/e52573

JMIR Mhealth Uhealth 2024 | vol. 12 | e52573 | p.638 (page number not for citation purposes)
The evaluations by the researchers and patients with HIV were conducted separately to avoid any interference. Each group followed a 3-step evaluation process. First, basic information for 21 apps, including their names, developers, platforms, and core functionalities, was collected from the SevenMa platform, Aladdin WeChat platform, and WeChat. Second, 2 WeChat applets and 2 independent apps were randomly selected for pilot evaluation. Before conducting the pilot evaluation, all 3 raters watched training videos on the MARS scale to better understand the purpose and significance of each item on the scale. Third, the 21 collected apps underwent a formal evaluation. To ensure consistent evaluation results, rater 1 and rater 2 independently assessed the same samples, completing the evaluation of the nth subdimension for all samples before moving on to evaluate the n+1th dimension. In cases where rater 1 and rater 2 had conflicting evaluations, we applied the Delphi method, which involved multiple rounds of consultation and feedback with experts, to achieve agreement regarding the discrepancies. Rater 3, an expert with relevant qualifications, reassessed the dimension that exhibited disagreement. The evaluation results from rater 3 were then shared with rater 1 and rater 2, along with appropriate explanations. This gave rater 1 and rater 2 the opportunity to revise their ratings and evaluations. If substantial differences persisted even after re-evaluation, the entire process was repeated until their evaluations converged or reached a satisfactory level of agreement.

Statistical Analysis

All analyses were performed using MedCalc software (version 20.217, MedCalc Software). Descriptive scores were derived from the MARS and uMARS scales. To evaluate the reliability of the raters’ assessments, Bland-Altman analysis was used to assess both interrater agreement and the range of variability in their scores. Visual representations were also used to visualize the differences in ratings between the 2 evaluators, providing a more concrete quantification of the discrepancies.

Ethical Considerations

This study focused solely on mobile apps and did not involve the collection of any personal information or data. Therefore, there were no ethical concerns or disputes associated with this research.

Results

Overview

A total of 21 apps that met the criteria were obtained through screening, with 18 being WeChat applets, dominating in terms of quantity, whereas only 3 were independent apps (Figure 1). Representative AIDS mobile apps were also displayed, as detailed in Figure 2. The characteristics of these apps are shown in Tables 1 and 2.

The developers of these apps can be categorized into 3 groups: companies (6/21, 29%), government organizations (7/21, 33%), and nonprofit social organizations (8/21, 38%). In addition, of the 21 included apps, 18 (86%) provided HIV/AIDS knowledge dissemination, 12 (57%) offered appointment booking for testing, 11 (52%) had counseling services, 8 (38%) had features related to PrEP and postexposure prophylaxis (PEP), 2 (10%) had community networking functionalities, and only 1 (5%) had live streaming capabilities. Of the 21 apps included in the study, 16 (76%) were able to update their content within a year, whereas 5 (24%) had not been updated for >1 year. Only 6 apps had privacy protection features, all of which provided privacy policies or agreements, with 2 of the apps offering privacy password functionality.
Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the selection process for inclusion of the apps.

Figure 2. Example of a typical Chinese AIDS mobile app.
<table>
<thead>
<tr>
<th>Features</th>
<th>Total apps, n (%)</th>
<th>Independent apps (n=3), n (%)</th>
<th>WeChat applets (n=18), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Affiliation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Company</td>
<td>6 (29)</td>
<td>3 (14)</td>
<td>3 (14)</td>
</tr>
<tr>
<td>Government organization</td>
<td>7 (33)</td>
<td>0 (0)</td>
<td>7 (33)</td>
</tr>
<tr>
<td>Nonprofit social organization</td>
<td>8 (38)</td>
<td>0 (0)</td>
<td>8 (38)</td>
</tr>
<tr>
<td><strong>Content type</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Live streaming</td>
<td>1 (5)</td>
<td>0 (0)</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Appointment for testing</td>
<td>15 (71)</td>
<td>3 (14)</td>
<td>12 (57)</td>
</tr>
<tr>
<td>Consultation</td>
<td>10 (48)</td>
<td>2 (10)</td>
<td>8 (38)</td>
</tr>
<tr>
<td>AIDS knowledge</td>
<td>18 (86)</td>
<td>2 (10)</td>
<td>16 (76)</td>
</tr>
<tr>
<td>PrEP\textsuperscript{a} or PEP\textsuperscript{b}</td>
<td>8 (38)</td>
<td>3 (14)</td>
<td>5 (24)</td>
</tr>
<tr>
<td>Web community</td>
<td>2 (10)</td>
<td>1 (5)</td>
<td>1 (5)</td>
</tr>
<tr>
<td><strong>Last updated</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within 1 month</td>
<td>5 (24)</td>
<td>1 (5)</td>
<td>4 (19)</td>
</tr>
<tr>
<td>Within 2 to 6 months</td>
<td>8 (38)</td>
<td>2 (10)</td>
<td>6 (29)</td>
</tr>
<tr>
<td>Within 7 to 12 months</td>
<td>3 (14)</td>
<td>0 (0)</td>
<td>3 (14)</td>
</tr>
<tr>
<td>&gt;1 year</td>
<td>5 (24)</td>
<td>0 (0)</td>
<td>5 (24)</td>
</tr>
<tr>
<td><strong>Privacy</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Privacy policy or agreement</td>
<td>6 (29)</td>
<td>3 (14)</td>
<td>3 (14)</td>
</tr>
<tr>
<td>Privacy password</td>
<td>2 (10)</td>
<td>2 (10)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

\textsuperscript{a}PrEP: pre-exposure prophylaxis.
\textsuperscript{b}PEP: postexposure prophylaxis.
Table 2. Basic information about the AIDS mobile apps.

<table>
<thead>
<tr>
<th>App name</th>
<th>App OR WC</th>
<th>Knowledge sharing</th>
<th>Appointment testing</th>
<th>Live streaming</th>
<th>Consultation</th>
<th>Risk assessments</th>
<th>PrEP\textsuperscript{b} and PEP\textsuperscript{c}</th>
<th>Web-based community</th>
<th>Privacy protection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chabei</td>
<td>WC</td>
<td>✓✓✓✓✓✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓✓</td>
<td>✓✓</td>
</tr>
<tr>
<td>Danlan Happy Test</td>
<td>WC</td>
<td>✓✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td>Suzhou Red Ribbon Hiv test</td>
<td>WC</td>
<td>✓✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baiyin HIV test</td>
<td>WC</td>
<td>✓✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linqu County Cdc\textsuperscript{d}</td>
<td>WC</td>
<td>✓✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wu Ai Fang Hua</td>
<td>WC</td>
<td>✓✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓✓</td>
<td></td>
</tr>
<tr>
<td>E Ai Jian</td>
<td>WC</td>
<td>✓✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ai Yi Jian</td>
<td>WC</td>
<td>✓✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rong Ai Jian</td>
<td>WC</td>
<td>✓✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liaoehong Dongchangfu District Anti-AIDS Service Platform</td>
<td>WC</td>
<td>✓✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhecheng County AIDS consulting and test</td>
<td>WC</td>
<td>✓✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Qingai Health Services</td>
<td>WC</td>
<td>✓✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ai Zhiku</td>
<td>WC</td>
<td>✓✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ai Cheng Wang Shi</td>
<td>WC</td>
<td>✓✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beijing AIDS Association</td>
<td>WC</td>
<td>✓✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Douai Check</td>
<td>WC</td>
<td>✓✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nanyue Gaozhibao</td>
<td>WC</td>
<td>✓✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Red Ribbon Volunteer House</td>
<td>WC</td>
<td>✓✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hong Feng Wan</td>
<td>App</td>
<td>✓✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓✓</td>
<td></td>
</tr>
<tr>
<td>Life4me+</td>
<td>App</td>
<td>✓✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓✓</td>
<td></td>
</tr>
<tr>
<td>Xiao Ai</td>
<td>App</td>
<td>✓✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{a}WC: WeChat applet.  
\textsuperscript{b}PrEP: pre-exposure prophylaxis.  
\textsuperscript{c}PEP: postexposure prophylaxis.  
\textsuperscript{d}CDC: Centers for Disease Control and Prevention.

**Functionality of the App**

Most mobile apps (18/21, 86%) provided HIV/AIDS knowledge dissemination, and 71% (15/21) of the apps offered appointment booking for testing. Nearly half of the apps (10/21, 48%) provided counseling services, whereas 38% (8/21) of the apps offered features related to PrEP or PEP. Only 10% (2/21) of the apps had functionalities for live streaming and web-based community.

In addition to these functions related to function evaluation, we also included additional statistics on privacy protection settings. Among the 21 apps included in the study, only 6 (29%) provided privacy protection settings and 2 (10%) designed privacy passwords.
Quality of the App

Overview of App Composite Scores

The composite scores for the app quality total score in MARS and uMARS were obtained by averaging the scores for each app. The overall composite score for the 21 included apps was 3.43 (mean app quality total score of MARS=3.47, SD 0.37; mean app quality total score of uMARS=3.38, SD 0.53). Hong Feng Wan achieved the highest composite score, with a score of 4.22 (mean app quality total score of MARS=3.96; mean app quality total score of uMARS=4.47), followed by Wu Ai Fang Hua with a composite score of 3.9 (mean app quality total score of MARS=4.06; mean app quality total score of uMARS=3.74), and Chabei secured the third rank with a composite score of 3.76 (mean app quality total score of MARS=3.66; mean app quality total score of uMARS=3.86). Suzhou Red Ribbon obtained the lowest composite score, with a score of 2.68 (mean app quality total score of MARS=3.67; mean app quality total score of uMARS=1.68).

Comparative Analysis of MARS Score and uMARS Score

Figure 3 presents a correlation analysis between the scores of MARS and uMARS. No significant relationship was observed between the app total quality scores of MARS and uMARS (0.41; P=.07) or between their subjective quality scores (0.39; P=.08). The app most reflective of the disparity between MARS and uMARS is Suzhou Red Ribbon, which had the greatest discrepancy between its app total quality scores on MARS (3.67) and uMARS (1.68). Its score on uMARS (1.68) was the lowest in the ranking, with respective dimensional scores of engagement (1.8), functionality (1.5), esthetics (1.67), information (1.75), and subjective quality (1). Conversely, for MARS, it ranked seventh for the app quality total score, with respective dimensional scores of engagement (2.9), functionality (4.38), esthetics (3.5), information (3.92), and subjective quality (2.5). There was considerable variation disclosed in the rankings of the most highly rated apps between MARS and uMARS. The top 5 app total quality scores on MARS were Wu Ai Fang Hua (4.06), Hong Feng Wan (3.96), Ai Yi Jian (3.95), Qing Ai Health Services (3.77), and Ai Zhiku (3.76), whereas the top 5 on uMARS were Hong Feng Wan (4.47), Chabei (3.86), Wu Ai Fang Hua (3.74), Linqu County CDC (3.66), and Zhecheng County AIDS consulting and test (3.62).

Interpreting composite data from Figures 4 and 5 and Table 3 reveal that within the 4 MARS scale dimensions of the evaluated 21 apps, the functionality dimension achieved the highest score (4.28, SD 0.39).

The information (mean 3.82, SD 0.33), aesthetics (mean 3.4, SD 0.49), and engagement (mean 2.4, SD 0.53) dimensions sequentially trailed behind. In contrast, under the uMARS scale’s evaluation, the information dimension prevailed with the top score (mean 3.89, SD 0.74), followed by the functionality (mean 3.84, SD 0.59), aesthetics (mean 3.21, SD 0.48), and engagement (mean 2.56, SD 0.63) dimensions.
Figure 4. Box plot of the Mobile App Rating Scale (MARS) score.

Figure 5. Box plot of the User Mobile App Rating Scale (uMARS) score.
Table 3. The MARS\textsuperscript{a} and uMARS\textsuperscript{b} scales’ scores for apps. The top 5 apps with app quality total score for each dimension are italicized.

<table>
<thead>
<tr>
<th>App name</th>
<th>Section A: engagement</th>
<th>Section B: functionality</th>
<th>Section C: aesthetics</th>
<th>Section D: information</th>
<th>App quality total score</th>
<th>Section E: subjective quality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MARS</td>
<td>uMARS</td>
<td>MARS</td>
<td>uMARS</td>
<td>MARS</td>
<td>uMARS</td>
</tr>
<tr>
<td>Ai Cheng Wang Shi</td>
<td>2.00</td>
<td>2.10</td>
<td>4.25</td>
<td>4.00</td>
<td>2.33</td>
<td>3.33</td>
</tr>
<tr>
<td>Ai Yi Jian\textsuperscript{c}</td>
<td>3.00</td>
<td>2.80</td>
<td>4.63</td>
<td>4.00</td>
<td>4.00</td>
<td>3.17</td>
</tr>
<tr>
<td>Ai Zhiku\textsuperscript{c}</td>
<td>2.50</td>
<td>2.30</td>
<td>4.38</td>
<td>4.00</td>
<td>4.00</td>
<td>3.17</td>
</tr>
<tr>
<td>Baiyin HIV test</td>
<td>2.30</td>
<td>2.60</td>
<td>4.63</td>
<td>4.00</td>
<td>3.50</td>
<td>3.33</td>
</tr>
<tr>
<td>Beijing AIDS Association</td>
<td>2.40</td>
<td>2.20</td>
<td>3.50</td>
<td>4.00</td>
<td>3.33</td>
<td>3.17</td>
</tr>
<tr>
<td>Chabei\textsuperscript{d}</td>
<td>2.20</td>
<td>2.70</td>
<td>4.50</td>
<td>4.50</td>
<td>3.67</td>
<td>3.50</td>
</tr>
<tr>
<td>Danlan Happy Test</td>
<td>1.60</td>
<td>2.30</td>
<td>3.25</td>
<td>3.75</td>
<td>2.84</td>
<td>3.00</td>
</tr>
<tr>
<td>Dou Ai Jian</td>
<td>1.90</td>
<td>1.80</td>
<td>4.13</td>
<td>3.25</td>
<td>3.00</td>
<td>2.67</td>
</tr>
<tr>
<td>E Ai Jian</td>
<td>2.30</td>
<td>2.70</td>
<td>4.75</td>
<td>4.00</td>
<td>3.84</td>
<td>3.33</td>
</tr>
<tr>
<td>Red Ribbon Volunteer House</td>
<td>2.00</td>
<td>2.50</td>
<td>3.88</td>
<td>4.00</td>
<td>3.00</td>
<td>3.17</td>
</tr>
<tr>
<td>Liaocheng Dongchangfu District Anti-AIDS Service Platform</td>
<td>2.00</td>
<td>2.20</td>
<td>4.50</td>
<td>4.00</td>
<td>3.50</td>
<td>3.33</td>
</tr>
<tr>
<td>Linqu County CDC\textsuperscript{d}</td>
<td>2.10</td>
<td>2.60</td>
<td>4.63</td>
<td>4.00</td>
<td>3.17</td>
<td>3.67</td>
</tr>
<tr>
<td>Nanyue Gaozhibao</td>
<td>1.90</td>
<td>2.20</td>
<td>4.00</td>
<td>4.00</td>
<td>2.84</td>
<td>3.17</td>
</tr>
<tr>
<td>Qing Ai Health Services\textsuperscript{c}</td>
<td>2.90</td>
<td>3.30</td>
<td>4.50</td>
<td>4.00</td>
<td>3.84</td>
<td>3.17</td>
</tr>
<tr>
<td>Rong Ai Jian</td>
<td>2.30</td>
<td>2.40</td>
<td>4.63</td>
<td>3.88</td>
<td>3.84</td>
<td>3.33</td>
</tr>
<tr>
<td>Suzhou Red ribbon</td>
<td>2.90</td>
<td>1.80</td>
<td>4.38</td>
<td>1.50</td>
<td>3.50</td>
<td>1.67</td>
</tr>
<tr>
<td>Wu Ai Fang Hua\textsuperscript{c,d}</td>
<td>3.20</td>
<td>3.30</td>
<td>4.63</td>
<td>4.00</td>
<td>4.17</td>
<td>3.67</td>
</tr>
<tr>
<td>Zhecheng County AIDS consulting and test\textsuperscript{d}</td>
<td>2.40</td>
<td>2.60</td>
<td>4.00</td>
<td>4.00</td>
<td>3.17</td>
<td>3.50</td>
</tr>
<tr>
<td>Life4me+</td>
<td>1.80</td>
<td>2.00</td>
<td>4.13</td>
<td>3.50</td>
<td>2.67</td>
<td>2.67</td>
</tr>
<tr>
<td>Hong Feng Wan\textsuperscript{c,d}</td>
<td>3.60</td>
<td>4.70</td>
<td>4.25</td>
<td>4.25</td>
<td>3.67</td>
<td>4.17</td>
</tr>
<tr>
<td>Xiao Ai</td>
<td>3.10</td>
<td>2.70</td>
<td>4.38</td>
<td>4.00</td>
<td>3.50</td>
<td>3.33</td>
</tr>
</tbody>
</table>

\textsuperscript{a}MARS: Mobile App Rating Scale.  
\textsuperscript{b}uMARS: User Mobile App Rating Scale.  
\textsuperscript{c}The top 5 apps in terms of app quality total score in MARS.  
\textsuperscript{d}The top 5 apps in terms of app quality total score in uMARS.

**Comparative Analysis of Quality Between WeChat Mini-Apps and Independent Apps**

As shown in Table 4, insignificant disparities are discernible between WeChat mini-apps and stand-alone apps regarding 2 metric domains: app quality ($P_{\text{MARS}}=.70$; $P_{\text{uMARS}}=.54$) and subjective app quality ($P_{\text{MARS}}=.48$; $P_{\text{uMARS}}=.80$).
Table 4. Independent samples $t$ test for unequal variances.

<table>
<thead>
<tr>
<th>MARS&lt;sup&gt;a&lt;/sup&gt; quality</th>
<th>MARS, mean (SD)</th>
<th>uMARS&lt;sup&gt;b&lt;/sup&gt;, mean (SD)</th>
<th>MARS, $t$ value ($df$)</th>
<th>uMARS, $t$ value ($df$)</th>
<th>MARS, $P$ value</th>
<th>uMARS, $P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>App quality</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>WeChat applets</td>
<td>3.45 (0.37)</td>
<td>3.33 (0.50)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Independent apps</td>
<td>3.58 (0.47)</td>
<td>3.66 (0.76)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>App subjective quality</td>
<td>—</td>
<td>—</td>
<td>—0.852 (2.2)</td>
<td>0.289 (2.3)</td>
<td>.48</td>
<td>.80</td>
</tr>
<tr>
<td>WeChat applets</td>
<td>2.39 (0.57)</td>
<td>2.72 (0.58)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Independent apps</td>
<td>2.92 (1.05)</td>
<td>2.88 (0.90)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

<sup>a</sup>MARS: Mobile App Rating Scale.
<sup>b</sup>uMARS: User Mobile App Rating Scale.
<sup>c</sup>Not available.

Internal Consistency and Reliability Testing for MARS and uMARS Scores

Similarly, we formed 4 Bland-Altman plots using the disparities and mean values between the scorings by reviewers on app quality (both MARS and uMARS) and app subjective quality (MARS and uMARS). The relatively limited range of the 95% limits of agreements suggest that the evaluators’ judgment outputs contain minor dissimilarities. A significant proportion of dots in all 4 Bland-Altman plots lie within the concordance interval (Figures 6-9), with their arithmetic means impressionably approaching 0 (Table 5).

This provides evidence of the high degree of internal uniformity and dependability in both MARS and uMARS scores.

Figure 6. Bland-Altman plot of the app quality (Mobile App Rating Scale).
Figure 7. Bland-Altman plot of the app subjective quality (Mobile App Rating Scale).

Figure 8. Bland-Altman plot of the app quality (User Mobile App Rating Scale).
**Figure 9.** Bland-Altman plot of the app subjective quality (User Mobile App Rating Scale).

**Table 5.** Summary of results of Bland-Altman analysis.

<table>
<thead>
<tr>
<th>Index</th>
<th>95% LoA(^a)</th>
<th>95% CI (arithmetic mean)</th>
<th>95% CI (LoA: upper limit)</th>
<th>95% CI (LoA: lower limit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>App quality (MARS(^b))</td>
<td>-0.19 to 0.21</td>
<td>-0.04 to 0.06</td>
<td>0.13 to 0.30</td>
<td>-0.28 to -0.11</td>
</tr>
<tr>
<td>App subjective quality (MARS)</td>
<td>-0.89 to 1.03</td>
<td>-0.15 to 0.29</td>
<td>0.64 to 1.41</td>
<td>-1.27 to -0.50</td>
</tr>
<tr>
<td>App quality (uMARS(^c))</td>
<td>-0.38 to 0.64</td>
<td>0.01 to 0.25</td>
<td>0.44 to 0.85</td>
<td>-0.58 to -0.17</td>
</tr>
<tr>
<td>App subjective quality (uMARS)</td>
<td>-0.66 to 1.12</td>
<td>0.02 to 0.43</td>
<td>0.76 to 1.47</td>
<td>-1.02 to -0.30</td>
</tr>
</tbody>
</table>

\(^a\)LoA: limits of agreement.
\(^b\)MARS: Mobile App Rating Scale.
\(^c\)uMARS: User Mobile App Rating Scale.

**Discussion**

**Principal Findings**

This study provides a comprehensive statistical summary of the functionalities and characteristics of 21 Chinese HIV-related mobile apps, along with their quality assessment using the MARS scale. The primary objectives behind the development of these mobile apps were the dissemination of HIV prevention knowledge (18/21, 86%), appointment booking for testing (15/21, 71%), and counseling services (10/21, 48%). These mobile apps aim to assist users in HIV prevention and treatment while improving the quality of communication and interaction between patients and health care providers.

**Overview of App Functions**

In recent years, the Chinese government has implemented a series of regulations and policies for HIV/AIDS prevention and has established an effective mechanism for HIV/AIDS control. These efforts have achieved certain progress in HIV/AIDS prevention. However, the burden of AIDS disease continues to increase, and the prevention and control situation remains severe. With the widespread adoption of smartphones and the rapid development of mobile internet, mobile apps have played an increasingly important role in disseminating HIV/AIDS knowledge and raising public awareness of prevention [1,37].

Among the 21 apps included in this study, 18 (86%) had HIV/AIDS knowledge dissemination functionality. This indicates that HIV/AIDS knowledge dissemination is an almost indispensable feature of Chinese HIV/AIDS mobile apps, which have become important tools for promoting HIV/AIDS knowledge in China.

In this study, 17 (81%) of the 21 mobile apps were identified, including 14 (67%) WeChat applets (such as Chabei, Danlan Happy Test, Baiyin HIV test, Suzhou Red Ribbon, Linqu County CDC, E Ai Jian, Ai Yi Jian, Rong Ai Jian, Liaocheng Dongchangfu District Anti-AIDS Service Platform, Zhecheng County AIDS consulting and test, Qing Ai Health Service, Ai Cheng Wang Shi, Beijing AIDS Association, and Dou Ai Jian) and 3 (14%) independent apps (Life4me+, Hong Feng Wan, and Xiao Ai). These apps provided various forms of health services, including appointment booking, counseling, risk assessment, and PrEP or PEP. A total of 2 mobile apps (Wu Ai Fang Hua and Hong Feng Wan) offered web-based community functionalities and scored excellently in terms of app quality and subjective quality, ranking among the top 2 positions. This finding is consistent with that of a previous study by Nour et al [38]. Therefore, we recommend considering the incorporation of web-based community functions when developing and designing mobile apps related to AIDS. This can encourage users to share their experiences, ask questions, provide...
suggestions, and offer mutual assistance and support. Organizing regular web-based or offline activities can also promote face-to-face interactions among users, thereby enhancing their sense of belonging and engagement.

Privacy protection is another crucial topic that should be given sufficient attention in AIDS-related mobile apps. As social stigmatization and discrimination are still associated with AIDS, many users are concerned about their personal information being disclosed or misused, posing a threat to their privacy and security [39]. Therefore, unlike general mobile apps, ensuring user privacy protection in AIDS mobile apps is of utmost importance, aligning with previous research findings [40]. However, in this study, of the 21 apps, only 5 (24%) mobile apps, including 2 (10%) WeChat applets (Chabei and Su Zhou Red Ribbon) and 3 (14%) apps (My Life+, Hong Feng Wan, and Xiao Ai), provided privacy policies or agreements. The average subjective quality score (mean 2.85, SD 0.77) was significantly higher than that of the mobile apps without privacy protection features (mean 2.34, SD 0.58). This observed phenomenon may be attributed to the fact that inadequate privacy protection design may lead to user wariness and lower intentions and frequencies of use regarding AIDS mobile apps [41]. Therefore, we strongly suggest that all AIDS-related mobile apps incorporate privacy policies or agreements to alleviate user concerns about privacy protection, enhance user trust, and promote willingness to use these apps. In addition to transparent privacy policies and user agreements, user education is also critical. Developers of AIDS mobile apps should provide users with relevant information and educational resources on privacy protection. This can be achieved through in-app prompts, tutorials, and frequently asked questions. Users should receive clear instructions on configuring their privacy settings, handling their personal data, and the steps to take in reporting privacy concerns or data breaches. These instructions should be easily accessible, comprehensible, and provide specific examples to guide users through each process.

Rating of App Quality
According to the results from MARS scores, among the 4 dimensions of app quality, both researchers and patients with HIV rated 21 mobile apps lowest on engagement (meanMARS 2.4, SD 0.53; meanuMARS 2.56, SD 0.63), especially the WeChat mini-apps, which had the lowest average scores (meanMARS 2.33, SD 0.43; meanuMARS 2.47, SD 0.42). The average engagement scores for the 7 mobile apps developed by the Chinese government were even lower (meanMARS 2.27, SD 0.33; meanuMARS 2.34, SD 0.30). This underlined the concern that the level of engagement would remain low, even for officially developed apps. An analysis of the respective scores for each item within the MARS and uMARS engagement portions revealed a predominant lack of interest and amusement within these mobile apps, thus leading to decreased user engagement. Many AIDS-related mobile apps merely provide basic information and functions without interactive and stimulating designs to attract user participation, lacking a strong appeal to users. In addition, the absence of personalized and customized features limits user engagement. A study has shown a significant correlation between user engagement and an increase in the adoption rate of mobile apps [42]. Therefore, it is recommended that developers not only provide high-quality HIV/AIDS prevention and treatment information but also focus on meeting users’ needs in terms of multidimensionality, functionality, and depth. In addition, attention should be paid to design in terms of amusement, entertainment, customization, interactivity, and other participatory aspects. This will help attract more user participation and enhance user stickiness.

This investigation revealed a lack of correlation between the MARS and uMARS scores. Specifically, (1) the app with the lowest overall uMARS scores across all dimensions surprisingly ranked seventh in terms of MARS scores. (2) A notable discrepancy was found in the functionality dimension ratings among patients with HIV and researchers, exposing the highest demographic variance in this attribute. (3) Moreover, the apps predominantly preferred by patients with HIV exhibited robust performance in functionality and information dimensions, with the latter appearing particularly predominant. However, the apps gravitating toward researchers demonstrated high competence in functionality, esthetics, and information, with functionality being the most superior. These data imply possible significant divergences distinct between the evaluations of researchers and genuine users of the apps. To maintain superior app quality and consumer satisfaction, rigorous surveillance of app quality should be sustained from both research and real-user perspectives. This bifocal assessment permits the accurate identification of genuine user requirements and researcher appraisal, providing valuable insights for the pinpoint and scientific enhancement of both app quality and user experience.

In this study, we observed that most Chinese AIDS-related mobile apps are WeChat applets. A reason for this is that the WeChat applet has advantages such as not requiring downloading or installation, having minimal resource consumption, and high user retention rates, which independent apps do not possess [43]. Another reason is that the HIV/AIDS population is relatively niche [44], which means that the market for AIDS mobile apps represents a low-frequency demand niche market. Independent apps often overlook this niche market owing to their high development costs, whereas WeChat applets, with their low development costs and the ability to cater to low-frequency demand niches through segmented scene construction, can effectively meet the needs of this market. Although WeChat mini-apps dominate the landscape of mobile apps for HIV/AIDS in China, their comprehensive performance lags behind that of stand-alone apps. The app quality score (meanMARS 3.45, SD 0.37 vs 3.58, SD 0.47; meanuMARS 3.33, SD 0.50 vs 3.66, SD 0.76) and subjective quality score (meanMARS 2.39, SD 0.57 vs 2.92, SD 1.05; meanuMARS 2.72, SD 0.58 vs 2.88, SD 0.90) for WeChat mini-apps are both lower compared with stand-alone apps. A granular analysis of the scores in different dimensions reveals the greatest discrepancy in the engagement dimension, with WeChat mini-apps scoring markedly lower (meanMARS 2.33, SD 0.43 vs 2.83, SD 0.93; meanuMARS 2.47, SD 0.42 vs 3.13, SD 1.40). This can be attributed to the constraints and limitations imposed by the WeChat platform, which prevent WeChat applets from providing the same user experience as independent apps in terms of

https://mhealth.jmir.org/2024/1/e52573
interface design, interaction methods, and customization settings. Improvement in these aspects should be a key focus for the future development of AIDS-related WeChat applets.

**Limitations**
The mobile apps evaluated in this study represent a snapshot of the current status of Chinese HIV-related mobile apps during the research period. Over time, mobile apps may be removed or updated, so the list of included mobile apps meeting the criteria may change in the future. In addition, some mobile apps intended solely for internal organizational use or no longer available in the market were excluded.

**Conclusions**
This study presents a systematic introduction to the functionality and quality of the currently available Chinese mobile apps for AIDS, providing valuable decision-making support for AIDS-related groups and health care professionals in accessing high-quality AIDS mobile apps. Through our systematic search and evaluation of existing Chinese mobile apps for AIDS, it has been observed that because of the lower demand frequency of AIDS mobile apps in China, less costly WeChat mini-apps have become the primary mode of app, and the overall quality attains a merely average level. A significant implication of our research is identifying the potentially significant discrepancy between the assessments made by researchers and the authentic users of the apps. Consequently, the inclusion of genuine users during the assessment and refinement stages of HIV apps is crucial. The main purpose of developing these mobile apps is to spread HIV prevention knowledge and facilitate booking appointments for testing and counseling services. However, most of these apps lack privacy protection features. Unlike general mobile apps, privacy protection is especially crucial in AIDS-related mobile apps because it directly affects users’ willingness to use them [40]. Therefore, the introduction of legal and ethical frameworks for privacy protection as well as privacy protection technologies is essential. In addition, enhancing user education on privacy protection and ensuring informed consent is of utmost importance. Research related to privacy protection in Chinese AIDS mobile apps may be a vital and urgent topic to address in the future. It is our intent that the findings of our research may function as a road map and reference for the future development of HIV apps in China. Furthermore, we aim to provide crucial decision-making support for individuals living with HIV in their quest for superior HIV apps.

**Acknowledgments**
The authors sincerely acknowledge the editors and anonymous reviewers for their insights and comments to further improve the quality of the manuscript. The authors thank Wang Xiang, chief of the AIDS Prevention Section of Bengbu Centers for Disease Control and Prevention, for his great help in the investigation. Finally, the first author (PL) would like to personally express his gratitude to his son (Liu Daixin) and his family for their support all through this work. This work was supported by the Anhui Provincial University Philosophy and Social Science Major Project (grant 2023AH040283).

**Data Availability**
The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

**Authors’ Contributions**
PL and FW initiated this study. PL and FW contributed to the study design. PL drafted the manuscript. LW conducted data sorting and screening. All authors revised the manuscript and approved the final version of the manuscript.

**Conflicts of Interest**
None declared.

**References**


25. Chan PS, Chidgey A, Lau J, Ip M, Lau JT, Wang Z. Effectiveness of a Novel HIV self-testing service with online real-time counseling support (HIVST-online) in increasing HIV testing rate and repeated HIV testing among men who have sex with HIV.


Abbreviations

CDC: Centers for Disease Control and Prevention
GBMSM: gay, bisexual, and other men who have sex with men
HIVST: HIV self-testing
MARS: Mobile App Rating Scale
mHealth: mobile health
PEP: postexposure prophylaxis
PrEP: pre-exposure prophylaxis
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
uMARS: User Mobile App Rating Scale
Introduction

Low back pain (LBP) is a complex multifactorial disorder, often considered a combination of physical, psychological, and social dysfunction [1]. A multidisciplinary self-management intervention based on a biopsychosocial model holds significant potential to manage LBP [2,3] and has been demonstrated to be more effective than unimodal exercise therapy [4]. The
National Institute for Health and Care Excellence (NICE) guidelines recommend using self-management for LBP, which can be described as the patient’s proactive adoption of strategies to manage their symptoms and monitor their health and well-being [5]. Despite being a promising approach to managing LBP, it can be challenging for an individual to self-manage any long-term condition [6]. The adherence to self-management strategies is commonly poor, especially without support, feedback, and positive reinforcement [7,8]. A qualitative study also noted poor adherence to advice and exercises as a limiting factor to recovery from LBP [9].

With the increasing popularity of electronic products, digital health solutions such as smartphone apps can be used as an innovative way to support self-management for many conditions, including LBP and may provide a solution to some of the problems outlined above [6,10]. Mobile health (mHealth) apps for pain management may be beneficial to patients [11,12], helping monitor those with acute or chronic pain and providing them with information and support for pain management. However, while many mHealth apps have been introduced into the commercial marketplace to manage pain, most have not been regulated in a uniform or standardized way before being released to the market [12,13]. The involvement of health care professionals in their development and content has been lacking [14]. This has raised concerns about the quality of these mHealth apps and whether their content information is evidence-based [15,16]. It is therefore desirable to assess the quality of current apps in the UK market and whether their content aligns with guideline recommendations. Another concerning issue is the paucity of evidence on specific intervention approaches in developed mHealth apps that have been developed, including underpinning evidence and theory [17] and relative risk management [18]. This potentially impacts on safety and efficacy of health-related smartphone apps [19].

Since there is no unified framework for assessing the theoretical and risk-based approaches associated with LBP self-management applications, we developed a theoretical framework that considers the theoretical care model of the intervention, the personalization of care, and the rate of intervention progression, as well as a risk-related framework that includes the targeted age group and the provision of appropriate safety checks. Interventions based on specific theoretical frameworks are known to be more effective in health care [20]. The Medical Research Council (MRC) guidelines strongly recommend using theoretical approaches in designing complex interventions [21], thus its inclusion in the evaluation. A critical factor in mHealth apps is personalization, where the management or treatment of the disease is tailored to the patient’s situation and individual needs, which will make the user feel that it is relevant and meaningful to them [22,23], thus the inclusion of criteria in relation to personalization. Treatment methods such as exercise need to be both adaptable and progressive. This means starting with basic exercises and gradually advancing to more complex levels at a pace that patients with LBP can comfortably handle. Such a progression helps enhance their functional abilities and quality of life, while simultaneously equipping them with strategies to effectively manage their pain [24]. Since the treatment paths for LBP vary according to different age groups, including children, adults, and the elderly [25], the target user age group must be defined. Finally, patient safety is an essential component of health care provision and is critical to primary care management, which will effectively reduce the clinical risks associated with LBP management [18,26].

Therefore, this app review, aimed to first, identify apps for the self-management of LBP currently on the market and second, to assess their quality (eg, functionality and design), intervention content (compliance with best practice guidelines), underlying theory (eg, theoretical care model), and risk-related approaches (eg, the age group targeted).

Methods

Search Strategy

Apps currently on the market for the self-management of LBP were identified, reviewed, and analyzed using a systematic approach. The UK official app stores for both Apple’s iOS and Google’s Android OS were used to search for mobile apps. These 2 operating systems currently dominate the marketplace of mobile medical apps [27]. We logged into Apple iTunes and Android Google Play stores in May 2022. A total of 3 keywords recommended by the Cochrane Back and Neck Group [28] were used to search apps “low back pain,” “back pain,” and “lumbago.” A subsequent search in June 2023 was repeated to ensure that any relevant new apps in the last year were incorporated into the review.

Study Selection

Criteria for inclusion in the review were (1) apps were a self-contained product (ie, did not depend on an external device or add-ons), (2) apps offered at least 1 active treatment option for LBP (eg, unsupervised exercise program or patient education), (3) apps only designed for people with LBP, (4) apps created or updated in the last 5 years to ensure software functionality and ongoing technical support, and (5) apps developed in English. Exclusion criteria were (1) apps targeted at managing general chronic pain, (2) apps only focused on specific factors and diagnostic tests for LBP, (3) apps only focused on specific LBP pathologies (eg, lumbar disk herniation), (4) apps designed for clinicians, (5) general back fitness apps with no mention of physiotherapy or physical therapy or musculoskeletal (MSK) conditions, and (6) apps were not downloadable or had restrictions (eg, requiring an activate access password). Apps that incurred a cost were also included; however, when both paid and free versions of an app were available, we reviewed only the paid version to ensure access to the full content. If the same app was available on iOS and Android, the iOS version was kept for inclusion and analysis.

An independent reviewer initially screened the eligible apps based on the apps’ names and descriptions and the screenshots provided. After the preliminary screening phase, the same independent reviewer downloaded apps that met the eligibility criteria for a second screening. Concerns regarding inclusion were discussed and resolved within the research group (AM and DS) until a final decision was reached.
Data Extraction

The selected apps were downloaded onto either a Samsung SM-N975F (Android version 7.1.2) or an iPhone 12 (iOS version 16.5) for a complete assessment of eligibility and characteristics. Relevant background information offered in the included apps was extracted, including the name, version, developer, update date, cost, and presence of in-app purchases. We extracted the age or content rating and consumer rating (5-star rating system) when available. In addition, we extracted if the apps contained advertisements and whether these adverts were relevant to their back pain. It was also noted if the apps were asynchronous or synchronous (available with support from a provider), and whether the apps tracked user engagement. The collection of personally identifiable information by apps and whether consent is stated were also checked. The category of management content, specific intervention component, theoretical care model, personalization of care, the rate of intervention progression, the age group targeted, and safety checks were also extracted.

App Content Assessment

Main categories and specific components of LBP app management content were identified and classified. Frequency analysis was performed to determine the number of apps providing these intervention contents. The recently updated 2020 NICE guideline for LBP was used to assess whether the included apps provided evidence-based interventions (categorized as “yes/no”) [29]. For this, we mapped app interventions to recommendations listed in the NICE guidelines. This guideline provides the most recent best practice recommendations for assessing and managing LBP and sciatica in people aged 16 years or older [5].

App Quality Assessment

The Mobile App Rating Scale (MARS) was used to assess the quality of included apps in this review. MARS is a brief tool with a 23-item questionnaire to classify and assess the quality of mHealth apps for researchers, professionals, and clinicians [30]. It assesses app quality across 4 domains: engagement, functionality, aesthetics, and information quality. All items are assessed on a 5-point scale (1= inadequate, 2= poor, 3= acceptable, 4= good, and 5= excellent). The MARS has demonstrated excellent internal consistency and interrater reliability for evaluating the quality of mHealth apps [30]. To standardize the quality ratings, the assessor completed a MARS video training recommended by the developers of MARS [30]. A total of 10 apps were randomly selected for training until a consensus on the scores was reached.

App Assessment of the Theoretical and Risk-Related Approaches

A total of 5 aspects considered in our theoretical and risk-related approaches were assessed. These included the underpinning LBP management theory, tailoring of content, the intervention of progression approach, the age group the app targeted, and appropriate safety checks.

Ethical Considerations

This study does not involve human participants.

Results

App Selection

The searches performed in May 2022 yielded 392 apps from 2 platforms. Of these, 156 apps were identified from iTunes stores, and 236 apps were identified from Google Play stores. A total of 5 duplicates were removed, resulting in 387 apps identified for screening based on the titles and app descriptions. After initial screening, 319 apps were excluded. The eligible 68 apps were downloaded for a full evaluation and further 7 apps were excluded. The subsequent search in June 2023 found a total of 33 newly developed apps based on the initial search, of which 8 apps were newly developed. Finally, 69 apps were included in this review, of which 25 were iOS apps and 44 were Android apps. Figure 1 illustrates the selection procedure of smartphone apps for LBP.
Figure 1. The flowchart of the app selection process. LBP: low back pain.

Characteristics of Included Apps

Of the 69 apps included in this review, 25 (36%) were found on iTunes exclusively, 44 (64%) on Google Play exclusively, and 5 (7%) were found on both app stores. There are 5 (7%) of 69 apps that required payment, ranging in price from US $1.13 to US $22.87 (median US $8.17). The majority (n=64, 93%) of the apps included were free of charge. Of these, 8 offered in-app purchases ranging from US $3.80 to US $12.70. Android apps (n=41, 93%) were more often free to access full functionality than iOS apps (n=15, 60%). Of the 41 apps reviewed by customers on a 5-star rating system, the median customer rating in 19 apps from iTunes (4.3 stars) was higher than in 22 apps from Google Play (4.1 stars). With respect to age or content rating, most of the included iOS apps (n=19, 76%) were downloaded without any age limitation: 4 apps were restricted to those of 12 years or above and 3 apps were restricted to those of at least 17 years old. However, all Android apps were labeled as suitable for all age groups.

In terms of developers, there is a mix of some health care groups and other private companies. A total of 24 (35%) apps contained advertisements, 10 (42%) of which were for products or medical companies targeting MSK disorders, leaving 14 (58%) random advertisements. In addition, all apps were asynchronous which means that they failed to deliver continuously updated application data to users. A total of 9 (13%) apps collect personally identifiable information from users and only 2 (22%) asked for their consent for collection. It appeared that none of the apps tracked user engagement. The characteristics of each app are presented in Multimedia Appendix 1.

MARS Quality Assessment

The mean MARS total score obtained from 69 applications was 2.4 out of 5 (SD 0.44). Table 1 summarizes the MARS total scores for each app. All apps were initially assessed using MARS for engagement, functionality, aesthetics, and information. Mean scores for each subscale (out of 5) were calculated. Of the 5 categories, apps scored highest in the functionality (mean 3.0, SD 0.55) domain, followed by aesthetics (mean 2.6, SD 0.61) and engagement (mean 2.1, SD 0.58). The information domain received the lowest score (mean 2.1, SD 0.46). The MARS total score and domain score for each app are shown in Multimedia Appendix 2.
Table 1. Mobile app rating scale scores (N=69 apps).

<table>
<thead>
<tr>
<th>MARS subscale</th>
<th>iOS, mean (SD)</th>
<th>Android, mean (SD)</th>
<th>Total, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engagement</td>
<td>2.4 (0.52)</td>
<td>1.9 (0.55)</td>
<td>2.1 (0.58)</td>
</tr>
<tr>
<td>Functionality</td>
<td>3.1 (0.57)</td>
<td>3.0 (0.49)</td>
<td>3.0 (0.55)</td>
</tr>
<tr>
<td>Aesthetics</td>
<td>2.9 (0.57)</td>
<td>2.4 (0.47)</td>
<td>2.6 (0.61)</td>
</tr>
<tr>
<td>Information</td>
<td>2.3 (0.49)</td>
<td>1.9 (0.40)</td>
<td>2.1 (0.46)</td>
</tr>
<tr>
<td>MARS overall score</td>
<td>2.6 (0.43)</td>
<td>2.3 (0.35)</td>
<td>2.4 (0.44)</td>
</tr>
</tbody>
</table>

aMARS: Mobile App Rating Scale.
bAverage of 4 objective subscales.

Intervention Contents for LBP

The LBP interventions embedded with the included apps were mainly classified into 3 categories (Table 2). Of these, most (n=47, 68.1%) of the apps offered only an exercise program, while 14.5% (n=10) apps provided patient education alone, and 13.0% (n=9) apps recommended patient education in combination with an exercise program. The remaining 2.9% (n=2) apps provided some psychological intervention for LBP in combination with an exercise program. Finally, only 1.4% (n=1) app-prescribed combinations of patient education, exercise, and psychological intervention.

Table 2. Number and percentage of category for low back pain interventions used in included apps.

<table>
<thead>
<tr>
<th>Main category</th>
<th>Value, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient education + exercise program + psychological intervention</td>
<td>1 (1.4)</td>
</tr>
<tr>
<td>Exercise program + patient education</td>
<td>9 (13.0)</td>
</tr>
<tr>
<td>Exercise program + psychological intervention</td>
<td>2 (2.9)</td>
</tr>
<tr>
<td>Patient education only</td>
<td>10 (14.5)</td>
</tr>
<tr>
<td>Exercise program only</td>
<td>47 (68.1)</td>
</tr>
</tbody>
</table>

More specifically, of the 69 apps included in this review, most (n=51, 73.9%) apps recommended muscle stretching as a self-management strategy for LBP. A total of 42 (60.9%) apps suggested muscle strengthening, 32 (46.4%) apps offered core stability exercises, and 19 apps (27.5%) recommended using yoga to manage LBP. Also, there were 17 (24.6%) apps providing information about LBP mechanisms, followed by advice to use medication (n=9, 13%), staying active (n=9, 13%), postural therapy (n=8, 11.6%), cold and heat therapy (n=8, 11.6%), and aerobic exercise (n=8, 11.6%). Only some apps mentioned manual therapy (n=4, 5.8%), cognitive behavioral therapy (n=4, 5.8%), meditation (n=4, 5.8%), mindfulness (n=3, 4.3%), McKenzie exercise (n=3, 4.3%), electrotherapy (n=3, 4.3%), acupuncture (n=3, 4.3%), and lifestyle advice (n=2, 2.9%). Concerning app intervention content, 14 (78%) of 18 interventions for LBP were aligned with the NICE guidelines and 4 included interventions that were not endorsed by the NICE guideline: postural therapy, electrotherapy, cold and heat therapy, and acupuncture [29]. Details of the interventions offered for LBP managed in the included apps are summarized in Table 3.
Table 3. Number and percentage of specific component for low back pain interventions used in included apps.

<table>
<thead>
<tr>
<th>Specific components</th>
<th>Value, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Understanding LBP&lt;sup&gt;a&lt;/sup&gt; mechanisms</td>
<td>17 (24.6)</td>
</tr>
<tr>
<td>Staying active</td>
<td>9 (13.0)</td>
</tr>
<tr>
<td>Postural therapy</td>
<td>8 (11.6)</td>
</tr>
<tr>
<td>Lifestyle advice</td>
<td>2 (2.9)</td>
</tr>
<tr>
<td>Electrotherapy</td>
<td>3 (4.3)</td>
</tr>
<tr>
<td>Cold and heat therapy</td>
<td>8 (11.6)</td>
</tr>
<tr>
<td>Manual therapy</td>
<td>4 (5.8)</td>
</tr>
<tr>
<td>Medication use</td>
<td>9 (13.0)</td>
</tr>
<tr>
<td>Core stability exercise</td>
<td>32 (46.4)</td>
</tr>
<tr>
<td>Muscle strengthening</td>
<td>42 (60.9)</td>
</tr>
<tr>
<td>Muscle stretching</td>
<td>51 (73.9)</td>
</tr>
<tr>
<td>McKenzie exercise</td>
<td>3 (4.3)</td>
</tr>
<tr>
<td>Aerobic exercise</td>
<td>8 (11.6)</td>
</tr>
<tr>
<td>Yoga</td>
<td>19 (27.5)</td>
</tr>
<tr>
<td>Mindfulness</td>
<td>3 (4.3)</td>
</tr>
<tr>
<td>Meditation</td>
<td>4 (5.8)</td>
</tr>
<tr>
<td>CBT&lt;sup&gt;b&lt;/sup&gt;</td>
<td>4 (5.8)</td>
</tr>
<tr>
<td>Acupuncture</td>
<td>3 (4.3)</td>
</tr>
</tbody>
</table>

<sup>a</sup>LBP: low back pain.  
<sup>b</sup>CBT: cognitive behavioral therapy.

**Theoretical and Risk-Related Approaches**

None of the 69 apps included in this review referred to or explained their theoretical care model. Only 1 app considered or incorporated a tailored approach to their intervention. A total of 18 (26%) of these apps provided an intervention program based on principles of gradual intervention progression. No app mentioned the age group for which their intervention content was appropriate and none set an age limit for their use. Even fewer, 11 (16%) apps offered safety checks for app users, including identifying red flags, signs, and symptoms of LBP that required medical attention or providing safety-netting advice if the back pain did not resolve or worsen (Multimedia Appendix 3).

**Discussion**

**Intervention Content in LBP Apps**

We attempted to benchmark the content of the included apps against the most recently published best practice guideline from NICE and found that most LBP self-management app components were consistent with those recommended by NICE. The NICE LBP guidelines [29] recommended the use of a group exercise program, including biomechanical, aerobic, mind-body, or a combination of approaches. Accordingly, muscle stretching, strengthening, core stability exercises, and yoga are the most common interventions in self-management apps. The findings of this review support a 2016 systematic review, which found that interventions in LBP selected by app developers were primarily based on clinical practice guidelines [31]. Similarly, the 2021 Cochrane review also reported moderate-certainty evidence suggesting that different types of exercise therapy are effective in treating LBP [32]. This means that most current LBP self-management apps offer evidence-based interventions.

However, while the content of most apps appears evidence-based, they are not often delivered in the context of a complex intervention and as such do not reflect the current LBP care models. We found that most self-management apps rely on exercise interventions, and very few apps incorporate social and psychological interventions for managing LBP. This suggests that current self-management apps emphasize a more biological care model to manage LBP rather than considering the influence of psychological and social factors in the development and maintenance of pain [33]. Research has emphasized interrelationships among biological changes, psychological status, and the sociocultural context and as such, they all need to be considered to understand the impact of chronic pain and its subsequent management [34]. The biopsychosocial model has been widely accepted as a holistic approach to increase efficacy and outcomes in managing chronic LBP [2,35]. Also, a complex multidisciplinary approach with a biopsychosocial model has been recommended in the early stages of LBP to reduce the likelihood of chronicity following acute LBP [36]. Thus, digital self-management interventions for LBP should consider adopting this model.
Quality Assessment in Self-Management Apps

Generally, apps for the self-management of LBP are of poor quality as assessed by MARS. Functionality (mean 3.0, SD 0.55) was the domain that scored the highest on the MARS test, as described by other authors [31,37]. This implies that the apps are functioning well, easy to learn and navigate, and efficient. However, the quality assessment revealed that these apps had low scores on the domains of “engagement” (mean 2.1, SD 0.58) and “aesthetics” (mean 2.6, SD 0.61). This indicates that the features that make the app equally engaging and important to a wide user base may have been overlooked. This was partly because most apps did not consider using specific strategies to increase involvement and aesthetics from the user’s point of view (eg, entertainment, interactivity, customization, layout, and graphics). Involving patient users during the development of these apps might better identify their needs and characteristics, and increase adherence to improve self-management and health outcomes [38].

The lowest score on information (mean 2.1, SD 0.46) indicated weakness in the quality and trustworthiness of information presented in the included apps. This was evidenced through assessing credibility (MARS item 18) and evidence base (MARS item 19). In terms of MARS item 18, most apps were developed by either commercial businesses with a vested interest or a legitimate source without verification (eg, has no web page), yet few are developed by credible health organizations (eg, government or universities). The lack of health professional involvement is a consistent issue highlighted by Rizwana, who has expressed concern about the accuracy and trustworthiness of in-app information [39]. In addition, none of the apps available for LBP management were evaluated using a randomized controlled trial. This was evident in MARS item 19, which assesses whether the app has been trialed or tested in the published scientific literature, and therefore the effectiveness and safety of these apps remain unknown. These results align with previous reviews on mHealth apps directed at pain management, in which a lack of scientific basis of the outcomes was found to support the use of such apps [40,41]. A possible explanation could be that most of the apps are of commercial rather than scientific origin, which suggests that the need to promote cross-disciplinary collaboration between academic and commercial institutions might help develop the evidence base for using such apps [37,40].

Theoretical and Risk-Related Approaches

Current self-management apps do not appear to have used a theoretical rationale in their development. This aligns with the findings from a 2018 review, which found the development of current self-management mHealth apps for patient education programs lacks the support of underpinning theory or framework [42]. An underpinning theoretical model is widely recognized as a crucial component of health interventions and is important when trying to understand key components of the intervention, how they interact, and the mechanisms of the intervention [43-45]. Systematic reviews of existing evidence demonstrate that interventions underpinned by theory are more effective than those that are not [46,47]. Additionally, as a complex multifactorial condition, the management of LBP should consider theory development and identifying the evidence base in accordance with the MRC framework for complex interventions [48].

Regarding personalized care, almost all included apps provided limited customized service. Tailored communications provide individuals with information that is relevant to them and that fits their situation. This can lead to increased perceived personal relevance, user engagement, more in-depth processing of information, and consequently, more desire and motivation to engage in the health behavior change [49]. The importance of personalization of mHealth apps was emphasized in a qualitative review in 2019, which expressed that mHealth apps should meet patients’ needs since they were created for use [22]. Also, our review suggests that the principle of gradual progression from simple to more advanced levels of intervention is not universal. Providing tools that help the user implement exercise progression ensures that the intervention progressively becomes more challenging to continually stimulate adaptations and maintain interest in the program [24]. The Coventry, Aberdeen, and London—Refined (CALO-RE) taxonomy of behavior change techniques highlighted the importance of setting graded tasks, breaking targeted behaviors into smaller, more manageable tasks, and facilitating progress in small increments [50].

Current apps do not evaluate the age limitation of intervention content for which LBP advice and treatments are appropriate, which may pose some risks for the users. Although age ratings in the App Store are often reported, it is based primarily on the degree to which an app contains sensitive information rather than on the applicability of an intervention to different age groups. It is also worth noting that the current NICE guideline for LBP published in 2016 is based on an adult population. A review from the Lancet showed that the evidence underpinning LBP guidelines is drawn almost exclusively from clinical trials of adults, and there are few trials on the treatment of back pain in children [51]. Thus, adult LBP care pathways may not be suitable for adolescents and children [52,53]. When applying these LBP management apps to nonadults, their scope of application needs to be considered. In addition, most included apps performed poorly in providing safety considerations. Safety check advice, including information on the natural history of the illness, advice on worrying symptoms to watch out for, and specific information on how and when to seek help, as well as advice about the follow-up of investigations and hospital referrals, can effectively address uncertainty in the process of LBP management [54]. A scoping review in 2020 discussed that safety concerns within apps were a primary concern [18]. Thus, such approaches will improve the likelihood of providing users with appropriate care and reduce clinical risks associated with self-management [55].

Evaluation Tool for LBP Apps

Although some self-management apps show relatively high scores on the MARS score, such as “Back Doctor/Pain Relief-1.03.24,” “Perfect Posture&Healthy back-1.5.2,” and “Back pain exercise at home-1.0.99,” they perform poorly on the theoretical and risk-related framework, such as the personalization of care and the age group targeted. Conversely,
some self-management apps (eg, BackTrainer-2.0) adopted a tailored approach and addressed risk-related issues, but MARS’ quality assessment scores were not high. Additionally, MARS has not effectively evaluated or included the biopsychosocial care model commonly used in chronic pain [34]. This indicates that current evaluation tools may be limited in their ability to assess LBP self-management apps comprehensively, and consequently, further study is needed to explore whether a holistic tool to evaluate LBP self-management apps is required.

Barriers to Holistic Digital Apps

Despite the overwhelming evidence recommending the use of holistic multidisciplinary interventions based on a biopsychosocial model for the self-management of LBP, the proliferation of back pain apps on the market that use simple interventions is concerning. It is hard to speculate why this is the case given the strong evidence for the use of a biopsychosocial model in the management of LBP. This may relate to current regulatory approvals. For an app to deliver a complex biopsychosocial intervention it would be classified as a medical device requiring such approvals. Further work is needed to explain why there are so many apps on the market that fail to address the model of care we seek in medical practice, and how current regulatory processes affect this.

Moreover, digital health interventions have been strongly advocated for and implemented in other domains and countries. Notably, draft guidance from NICE has informed digital programs using multidisciplinary models to assist the NHS in delivering specialized services for weight management [56]. Furthermore, the Federal Institute for Drugs and Medical Devices in Germany has authorized the entry of multidisciplinary health care apps with robust trial data into the market, with the overarching objective of empowering clinicians to prescribe health care apps for their patients [57,58]. This illustrates the feasibility of developing multidisciplinary mHealth apps for health care practitioners to recommend. However, it remains imperative to conduct a comprehensive evaluation of the apps using clinical studies to ascertain their efficacy and suitability for widespread prescription.

Limitations

Apps not specifically targeted for the self-management of LBP (eg, chronic pain, pain management, or MSK apps) were excluded from this review, which may result in potentially eligible apps being missed. While many LBP apps in different countries are available in app stores, our search was limited to the UK iTunes and Google Play stores for practical reasons relating to data-capturing capacity. Therefore, this review did not include apps exclusively available in other countries, and therefore may not represent the broader landscape relating to digital approaches for management of LBP. Only 1 independent reviewer assessed app quality using MARS and discussed uncertainties with 2 other research members until a consensus was reached, which may impact the reliability of the assessments.

Conclusions

In this review, we identified 69 apps related to the self-management of LBP and rated them using MARS. Most apps scored poor quality due to their approaches to engagement and information, and many emerged as tools for delivering passive information rather than active management apps. Most apps were aligned with guideline-based care. However, no app offered a holistic self-management intervention approach incorporating a biopsychological model. Most apps underestimated the importance of theoretical and risk-related aspects. Thus, current self-management apps for LBP are limited in what they offer.

Acknowledgments

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Imperial College London is the sponsor for the study and has no influence on the direction or content of the work. We are grateful to Rebecca Jones (Library Manager in Imperial College London) who offered advice and support for the building search strategy and screening process. We also wish to thank our colleagues in the MSK laboratory for their valuable and helpful comments and suggestions. Without these people, studying would have been impossible.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors’ Contributions

TZ, DS, and AM conceived and designed the study. TZ takes responsibility for the acquisition of the data and the analysis of the data. TZ drafted the paper. DS and AM revised the paper critically for important intellectual content. All authors approved the final version to be published.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Characteristics of apps for the self-management of low back pain.
Multimedia Appendix 2
The Mobile App Rating Scale overall and subscale scores.

Multimedia Appendix 3
The content and approach for low back pain interventions used in the included apps.

References


57. Toelle TR, Upadela-Fischler DA, Haas KK, Priebe JA. App-based multidisciplinary back pain treatment versus combined physiotherapy plus online education: a randomized controlled trial. NPJ Digit Med 2019;2:34 [FREE Full text] [doi: 10.1038/s41746-019-0109-x] [Medline: 31304380]

Abbreviations

CALO-RE: Coventry, Aberdeen, and London—Refined
LBP: low back pain
MARS: Mobile App Rating Scale
mHealth: mobile health
MRC: Medical Research Council
MSK: musculoskeletal
NICE: National Institute for Health and Care Excellence
Evaluation of Patient-Facing Mobile Apps to Support Physiotherapy Care: Systematic Review

Mark Merolli1,2, BPhysio (Hons), PhD; Jill J Francis3, BA, PhD; Patrick Vallance3,4, BPTh; Kim L Bennell1, BPTh, PhD; Peter Malliaras4, BPTh (Hons), PhD; Rana S Hinman1, BPhysio (Hons), PhD

1Centre for Health, Exercise & Sports Medicine, Department of Physiotherapy, School of Health Sciences, The University of Melbourne, Melbourne, Australia
2Centre for Digital Transformation of Health, The University of Melbourne, Melbourne, Australia
3School of Health Sciences, The University of Melbourne, Melbourne, Australia
4Department of Physiotherapy, Podiatry, Prosthetics and Orthotics, School of Allied Health, Human Service and Sport, La Trobe University, Melbourne, Australia

Corresponding Author:
Mark Merolli, BPhysio (Hons), PhD
Centre for Health, Exercise & Sports Medicine
Department of Physiotherapy, School of Health Sciences
The University of Melbourne
161 Barry Street, Parkville
Melbourne, 3052
Australia
Phone: 61 383443689
Email: merollim@unimelb.edu.au

Abstract

Background: Mobile health interventions delivered through mobile apps are increasingly used in physiotherapy care. This may be because of the potential of apps to facilitate changes in behavior, which is central to the aims of care delivered by physiotherapists. A benefit of using apps is their ability to incorporate behavior change techniques (BCTs) that can optimize the effectiveness of physiotherapeutic interventions. Research continues to suggest that despite their importance, behavior change strategies are often missing in patient management. Evaluating mobile apps that physiotherapists can use to drive behavior change may inform clinical practice and potentially improve patient outcomes. Examining the quality of apps and exploring their key features that can support behavior change and physiotherapy care are important aspects of such an evaluation.

Objective: The primary aim of this study was to describe the range of mobile apps in app stores that are intended for use by patients to support physiotherapy care. The secondary aims were to assess app quality, BCTs, and their behavior change potential.

Methods: A systematic review of mobile apps in app stores was undertaken. The Apple App Store and Google Play were searched using a 2-step search strategy, using terms relevant to the physiotherapy discipline. Strict inclusion and exclusion criteria were applied: apps had to be intended for use by patients and be self-contained (or stand-alone) without the requirement to be used in conjunction with a partner wearable device or another plugin. Included apps were coded for BCTs using the Behavior Change Technique Taxonomy version 1. App quality was assessed using the Mobile App Rating Scale, and the App Behavior Change Scale was used to assess the app’s potential to change behavior.

Results: In total, 1240 apps were screened, and 35 were included. Of these 35 apps, 22 (63%) were available on both the Apple App Store and Google Play platforms. In total, 24 (69%) were general in their focus (eg, not condition-specific), with the remaining 11 (31%) being more specific (eg, knee rehabilitation and pelvic floor training). The mean app quality score (Mobile App Rating Scale) was 3.7 (SD 0.4) of 5 (range 2.8-4.5). The mean number of BCTs identified per app was 8.5 (SD 3.6). BCTs most frequently included in the apps were instruction on how to perform a behavior (n=32), action planning (n=30), and self-monitoring of behavior (n=28). The mean behavior change potential score (App Behavior Change Scale) was 8.5 (SD 3.1) of 21 (range 3-15).

Conclusions: Mobile apps available to support patient care received from a physiotherapist are of variable quality. Although they contain some BCTs, the potential for behavior change varied widely across apps.

International Registered Report Identifier (IRRID): RR2-10.2196/29047
physiotherapy; physical therapy; digital health intervention; mobile app; behavior change technique; behavior change; exercise; systematic review; quality; rehabilitation; BCT; mHealth; mobile health; app; apps; physical activity; fitness; synthesis; syntheses; review methods; review methodology; search; searches; searching; systematic; mobile phone

Introduction

Background

Digital health in physiotherapy is an emergent area. It continues to gather speed with its use in both clinical practice and research growing exponentially [1-3]. Many digital modalities, with varied functions, are relevant to physiotherapy practice; these might include the delivery of timely digital patient information and resources through websites and patient portals [4], digital patient assessment using connected Bluetooth and wireless-enabled devices [5], digital models of care (telehealth) supported by video-based consultation [6-8], remote monitoring of patient status through wearables [9,10], and mobile health (mHealth) apps used to prescribe, monitor, and support home exercise programs [1,11].

Mobile apps are one type of digital health modality of particular interest because of the ubiquity of smartphone use and their ability to deliver digital therapeutics [12-14]. They are relatively inexpensive and thus scalable [15]. They are also worthy of scientific attention within the physiotherapy community because of their ability to support interventions targeting several aspects of behavior change, otherwise known as behavior change techniques (BCTs) [16-19]. Facilitating positive behavior change in people with health conditions is often central to the management approach of physiotherapists, aimed at deriving the best possible therapeutic benefit from an intervention. For example, behavior change may be required for a person to follow physiotherapist advice regarding activity pacing or to adhere to exercise and physical activity recommendations [1,16,19,20].

BCTs can be thought of as the components of an intervention that regulate the behavior by altering cause and effect [21]. As defined by Michie and Johnston [22], they may be thought of as the “active ingredients” that facilitate intended behaviors. Apps have the ability to incorporate and deliver numerous BCTs, which can mediate a behavioral target of physiotherapy care [23-25].

The app space continues to grow exponentially, with well over 300,000 health apps now available and over 200 health apps added to major app stores daily [18]. Concurrently, despite a growing evidence base regarding the role of mobile apps in physiotherapy contexts, there remains a dearth of high-quality research using validated and rigorous scientific methods assessing the quality of physiotherapy-specific mobile apps and their behavior change aspects (eg, BCTs used and behavior change potential) [11,25].

Objectives

The primary study objective was to describe mobile apps, intended for use by patients, that can support physiotherapy care. The secondary objective was to evaluate app quality [26] and any BCTs contained within them [21], and to evaluate the behavior change potential of the apps [24].

Methods

Study Design

The detailed protocol for this systematic review was previously published [25]. As this study is a systematic review of apps (in app stores) and does not involve human participants, ethics approval was not required. The review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement’s systematic review reporting principles, with minor adaptations as relevant for our search of app stores rather than research literature [27]. This method is common and adopted in other similar studies [18,28,29] (Multimedia Appendix 1) [30].

Search Strategy

Overview

A 2-phase search strategy was used to search the popular app stores (Apple App Store and Google Play), which was in line with other rigorous systematic reviews of health apps in app stores [29]. App store search strategy, keywords used, and steps can be seen in Textbox 1 [25]. Initial search was conducted in March 2021 and rerun in February 2023 to ensure further up-to-date coverage. To maintain the feasibility of the search, each search term combination was limited to the first 100 apps retrieved. This was also done to maintain the relevance of apps retrieved, which diminishes as the end of the search list is refreshed [28].

Textbox 1. Search strategy.

<table>
<thead>
<tr>
<th>Step 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>“physiotherapy,” “physio,” “physical therapy,” “physiotherapist,” “physical therapist,”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>“physiotherapy,” “physio,” “physical therapy,” “physiotherapist,” “physical therapist,”</td>
</tr>
</tbody>
</table>

and

Additionally, the websites of the top 10 member organizations of World Physiotherapy (the peak international physiotherapy body) based on number of members were searched for recommendation or endorsement of any specific apps that met our criteria. Only websites in English and with the relevant section not behind a paywall were searched.

Selection Process
The search was performed independently by 2 reviewers (MM and PV). The same reviewers independently screened the apps using 1 Apple and 1 Android device each. Screening was based on information in the respective app stores, including app title, description, and screenshots [31]. Apps eligible for full analysis were downloaded onto the devices for review by each reviewer independently, and any disagreements were resolved by discussion. A third reviewer (PM) was flagged as a mediator to resolve any nonconsensus, but this was not required [28].

Data Extraction
Overview
The 2 reviewers (MM and PV) independently extracted descriptive data about each app from the app stores, within the apps themselves, or from official websites of the apps (if readily apparent from information in the app store). All extracted data were computed in Microsoft Excel (Microsoft Corp). A full list of extracted descriptive characteristics is shown in Multimedia Appendix 2 [25].

Both reviewers (MM and PV) independently engaged with all of the app functions for a minimum of 10 minutes on each device, for familiarization. This allowed each reviewer to independently code and score the app quality, BCTs, and their behavior change potential.

Mobile App Quality
To appraise the quality of included apps, the 23-item Mobile App Rating Scale (MARS) was used [26]. MARS was used because of its reliability, simplicity, and wide applicability in mHealth research [26,32]. Its primary domains are engagement, functionality, aesthetics, and information, as well as a subjective rating of quality. Each domain has questions coded on a 5-point Likert scale (ranging from 1=inadequate to 5=excellent). Mean scores are calculated for each of the 4 domains, which are tallied and divided by the number of domains to produce an overall mean quality score out of 5. Mean scoring is used instead of total scores because individual items can be rated as “not applicable” [26].

Calculating the mean scores of the engagement, functionality, aesthetics, and information quality objective subscales and an overall mean app quality total score is how the MARS is scored. Mean scores instead of total scores are used because an item can be rated as “not applicable.”

Both reviewers (MM and PV) completed MARS training before scoring each app independently. Any disagreements were rectified via discussion. A third reviewer (PM) was flagged as a mediator to resolve any nonconsensus, but this was not required.

Coding BCTs
Both reviewers (MM and PV) underwent and received certification in coding BCTs using the Behavior Change Technique Taxonomy version 1 (BCTTv1). This was to increase their competence in identifying and assessing BCTs and to improve coding agreement [33]. Based on the minimum of 10 minutes they engaged with app functions, each reviewer independently coded the BCTs incorporated in each app using the BCTTv1, a framework of 93 BCTs created for investigating behavior change interventions using valid and reliable methods [21]. Furthermore, the BCTs in the BCTTv1 are further arranged into 16 clusters, each including similar BCTs [21]. This clustering was helpful for coders when developing a novel codebook (Multimedia Appendix 3), as it helped clarify coding decisions when examining the behavior change potential of included apps [25]. Any disagreements were discussed to achieve consensus. Where consensus was not able to be achieved, a third reviewer (JFF; who is a behavior change expert and one of the original developers of the BCTTv1) helped resolve any disagreements.

Behavior Change Potential
The App Behavior Change Scale (ABACUS) was used to evaluate the behavior change “potential” of each app, which is a scale of 21 items [24]. Apps are scored by identifying any of the 21 items and summing these item scores to give a score out of 21. The ABACUS focuses on 4 BCT clusters: knowledge and information, goals and planning, feedback and monitoring, and actions [24]. Using clustered BCTs identified, we developed the aforementioned novel coding rulebook to support reporting, and the 2 reviewers (MM and PV) scored behavior change potential together using the ABACUS. Any disagreements were discussed at the time, and a third reviewer (JFF) decided in the case of nonconsensus.

Data Synthesis
App characteristics are reported as descriptive and categorical data and a proportion (percentage), including app name, data privacy transparency (yes or no), companion app (yes or no), platform (Apple App Store, Google Play, or both), focus of the app (specific condition or region or whether more general), physiotherapy specialty (eg, musculoskeletal, pelvic health, and pediatric), target behaviors, simplified target behaviors, country of origin, developer qualifications (not clear or health care professional [HCP] or non-HCP), app version number, payment method (free or in-app purchases or one-off payment), and cost. Overall mean (SD) and individual domain scores are presented for MARS and ABACUS scores, as well as total BCT identification frequency and BCTs coded per app.

Results
Search
Our initial search identified a total of 1913 apps (Figure 1). This included 1302 unique apps from the Apple App Store, 600 apps from the Google Play Store, and a further 11 apps identified on the websites of professional physiotherapy associations. Overall, 834 were left after duplicates were removed. Following screening of app name, description, and screenshots, a further
790 were excluded. In total, 44 apps were screened fully, of which 19 were excluded. Reasons for exclusion included non-English language, costing more than Aus $10 (a currency exchange rate of Aus $1=US $0.72 is applicable), delivering its own service, irrelevant content, targeted at physiotherapist (clinician use) rather than the patient, being a bespoke or white-labeled app (ie, standard architectures licensed to a business or brand for private use), incompatibility with a current mobile device, and not a stand-alone app. This left 25 included apps for data extraction. Search update was rerun in February 2023, yielding an additional 405 unique apps. Following screening, this yielded a further 10 included apps for data extraction. Thus, a total of 35 apps were examined.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.

Descriptive Characteristics of Included Apps

App characteristics are described in Multimedia Appendix 2. Overall, 22 (63%) apps were available on both Apple App Store and Google Play platforms, and 12 (34%) only on Apple App Store. Only 1 (3%) app was solely available on Google Play. Country of origin was clear for 24 (69%) apps, with 6 apps originating from the United Kingdom, 6 from the United States, and 5 from Australia. Regarding price, 26 (74%) were free to download, and 9 had a one-off payment price ranging from Aus $1.49 to $9.99. Additionally, 10 apps offered in-app purchases (including subscriptions). Just under half of the apps (n=16, 46%) were not clear about the developer’s qualifications, a further 17 (49%) were developed by HCPs (ie, physiotherapists, orthopedic surgeons, exercise physiologists, and chiropractors), and 2 (6%) were developed by non-HCPs. Overall, 15 (43%) apps had an obvious physiotherapist companion feature (ie, dashboard, desktop software, or precise login for therapists to access and engage with patients), and 31 (89%) apps had information about privacy and security in the app store. The focus of most apps was general in nature (eg, not condition-specific; n=24, 69%), while 11 (31%) had a specific target (eg, knee rehabilitation and pelvic floor training). This included 3 apps focusing on pelvic health and 1 app each focusing on wrist or hand, knee, hemophilia, core stability, wheelchair-bound individuals, cerebral palsy, osteoarthritis, and back pain.

Assessment of Quality (MARS)

Table 1 shows app quality scores using the MARS. Individual app quality scores ranged from 2.8 to 4.5, with a mean score of 3.7 (SD 0.4) of a maximum of 5. The apps with the highest MARS scores were Squeezy: CF (4.5), Squeezy (4.4), AllyCare (4.3), Squeezy for Men (4.2), and TeleHab (4.2). Of the appraised MARS domains, “aesthetics” (mean 3.9, SD 0.6) and “functionality” (mean 3.8, SD 0.5) were the highest scoring sections, while “subjective app quality” (mean 3.4, SD 0.8) and “engagement” (mean 3.4, SD 0.5) were the lowest. To the best of our knowledge, only 4 (11%) apps met criterion 19 (has the app been trialed or tested and published in scientific literature), including ReHand, Hand Rehabilitation, Squeezy, Embodia, and PhysiApp. Only 1 (2.9%) app (ReHand, Hand Rehabilitation) has been evaluated in a randomized controlled trial (RCT), showing positive outcomes for physiotherapy patients [34,35].
Table 1. Mobile App Rating Scale scoring.

<table>
<thead>
<tr>
<th>App</th>
<th>Section A: engagement</th>
<th>Section B: functionality</th>
<th>Section C: aesthetics</th>
<th>Section D: information</th>
<th>Mean app quality score (section A to D)</th>
<th>Section E: subjective app quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Rehab Diary</td>
<td>3.6</td>
<td>3.0</td>
<td>2.7</td>
<td>3.2</td>
<td>3.1</td>
<td>2.5</td>
</tr>
<tr>
<td>AllyCare</td>
<td>4.4</td>
<td>4.0</td>
<td>4.7</td>
<td>4.2</td>
<td>4.3</td>
<td>4.5</td>
</tr>
<tr>
<td>Back Pain Diary</td>
<td>3.4</td>
<td>4.0</td>
<td>3.7</td>
<td>2.8</td>
<td>3.5</td>
<td>2.5</td>
</tr>
<tr>
<td>BlueJay Engage - Patient</td>
<td>3.8</td>
<td>4.0</td>
<td>4.3</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
</tr>
<tr>
<td>ComplexCore</td>
<td>2.4</td>
<td>4.0</td>
<td>4.0</td>
<td>3.3</td>
<td>3.4</td>
<td>3.0</td>
</tr>
<tr>
<td>CP-Fit</td>
<td>3.2</td>
<td>3.5</td>
<td>4.0</td>
<td>3.3</td>
<td>3.5</td>
<td>3.0</td>
</tr>
<tr>
<td>Embodia</td>
<td>3.8</td>
<td>4.0</td>
<td>4.3</td>
<td>3.6</td>
<td>3.9</td>
<td>4.0</td>
</tr>
<tr>
<td>ExorLive</td>
<td>3.6</td>
<td>3.3</td>
<td>3.0</td>
<td>3.1</td>
<td>3.3</td>
<td>3.0</td>
</tr>
<tr>
<td>Extensor</td>
<td>3.4</td>
<td>3.5</td>
<td>3.3</td>
<td>3.3</td>
<td>3.4</td>
<td>3.0</td>
</tr>
<tr>
<td>Guided Physio</td>
<td>3.4</td>
<td>4.3</td>
<td>4.3</td>
<td>3.8</td>
<td>4.0</td>
<td>4.0</td>
</tr>
<tr>
<td>HaemActive</td>
<td>3.6</td>
<td>3.5</td>
<td>4.7</td>
<td>3.6</td>
<td>3.8</td>
<td>4.0</td>
</tr>
<tr>
<td>Home Physio</td>
<td>3.0</td>
<td>4.0</td>
<td>4.3</td>
<td>3.3</td>
<td>3.7</td>
<td>2.5</td>
</tr>
<tr>
<td>My Exercise Messages</td>
<td>3.4</td>
<td>4.3</td>
<td>4.0</td>
<td>4.8</td>
<td>4.1</td>
<td>3.5</td>
</tr>
<tr>
<td>My Exercise Program</td>
<td>3.6</td>
<td>3.8</td>
<td>4.0</td>
<td>3.8</td>
<td>3.8</td>
<td>3.3</td>
</tr>
<tr>
<td>My Injury Exercises</td>
<td>2.8</td>
<td>4.0</td>
<td>4.0</td>
<td>3.5</td>
<td>3.6</td>
<td>3.3</td>
</tr>
<tr>
<td>OT App Lite</td>
<td>2.6</td>
<td>3.5</td>
<td>4.0</td>
<td>3.0</td>
<td>3.3</td>
<td>2.0</td>
</tr>
<tr>
<td>PhysioApp</td>
<td>3.4</td>
<td>4.3</td>
<td>4.3</td>
<td>3.7</td>
<td>3.9</td>
<td>4.0</td>
</tr>
<tr>
<td>Physiotools Trainer</td>
<td>3.0</td>
<td>3.5</td>
<td>3.3</td>
<td>3.3</td>
<td>3.3</td>
<td>3.3</td>
</tr>
<tr>
<td>Pocket Physio</td>
<td>2.4</td>
<td>2.5</td>
<td>3.0</td>
<td>3.3</td>
<td>2.8</td>
<td>2.3</td>
</tr>
<tr>
<td>PT Timer: Stretch &amp; Exercise</td>
<td>3.0</td>
<td>3.5</td>
<td>3.0</td>
<td>3.2</td>
<td>3.2</td>
<td>2.0</td>
</tr>
<tr>
<td>PT-Helper Pro</td>
<td>3.0</td>
<td>3.3</td>
<td>3.3</td>
<td>3.0</td>
<td>3.2</td>
<td>2.5</td>
</tr>
<tr>
<td>RecovAware Knee Health Fitness</td>
<td>4.2</td>
<td>3.8</td>
<td>4.3</td>
<td>4.2</td>
<td>4.1</td>
<td>4.0</td>
</tr>
<tr>
<td>Rehab Guru Client</td>
<td>3.8</td>
<td>3.8</td>
<td>4.3</td>
<td>3.8</td>
<td>3.9</td>
<td>4.0</td>
</tr>
<tr>
<td>ReHand, Hand Rehabilitation</td>
<td>3.6</td>
<td>4.3</td>
<td>4.3</td>
<td>4.1</td>
<td>4.1</td>
<td>4.0</td>
</tr>
<tr>
<td>Smart Therapist</td>
<td>3.4</td>
<td>3.5</td>
<td>3.3</td>
<td>3.5</td>
<td>3.4</td>
<td>3.3</td>
</tr>
<tr>
<td>Squeezzy: CF</td>
<td>4.0</td>
<td>5.0</td>
<td>4.3</td>
<td>4.5</td>
<td>4.5</td>
<td>4.5</td>
</tr>
<tr>
<td>Squeezzy for Men</td>
<td>3.6</td>
<td>4.8</td>
<td>4.3</td>
<td>4.3</td>
<td>4.2</td>
<td>4.5</td>
</tr>
<tr>
<td>Squeezzy</td>
<td>4.0</td>
<td>5.0</td>
<td>4.3</td>
<td>4.3</td>
<td>4.4</td>
<td>5.0</td>
</tr>
<tr>
<td>Switchback Health</td>
<td>3.2</td>
<td>3.8</td>
<td>3.7</td>
<td>3.5</td>
<td>3.5</td>
<td>3.0</td>
</tr>
<tr>
<td>TeleHab</td>
<td>4.0</td>
<td>4.5</td>
<td>4.7</td>
<td>3.7</td>
<td>4.2</td>
<td>4.0</td>
</tr>
<tr>
<td>Track Rehab</td>
<td>3.4</td>
<td>3.3</td>
<td>3.3</td>
<td>3.8</td>
<td>3.5</td>
<td>3.0</td>
</tr>
<tr>
<td>TrackActive Pro - Patient App</td>
<td>3.4</td>
<td>3.5</td>
<td>3.7</td>
<td>3.8</td>
<td>3.6</td>
<td>3.5</td>
</tr>
<tr>
<td>VR steps Home rehabilitation</td>
<td>4.0</td>
<td>3.8</td>
<td>3.7</td>
<td>3.3</td>
<td>3.7</td>
<td>3.3</td>
</tr>
<tr>
<td>Wheelchair Exercises</td>
<td>3.0</td>
<td>3.8</td>
<td>4.0</td>
<td>3.5</td>
<td>3.6</td>
<td>3.3</td>
</tr>
<tr>
<td>YRMOVE</td>
<td>3.0</td>
<td>3.5</td>
<td>3.0</td>
<td>2.6</td>
<td>3.0</td>
<td>2.3</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>3.4 (0.5)</td>
<td>3.8 (0.5)</td>
<td>3.9 (0.6)</td>
<td>3.6 (0.5)</td>
<td>3.7 (0.4)</td>
<td>3.4 (0.8)</td>
</tr>
</tbody>
</table>

**Behavior Change: Target Behaviors and BCTs**

The most frequently observed target behaviors in the apps included recording information about exercise (n=28, 80%), performing therapeutic exercise (general; n=25, 71%), and communicating with a health professional (n=12, 34%; Multimedia Appendix 2). Other observed target behaviors included performing therapeutic exercise (men’s health, women’s health, hand therapy, knee, and perioperative), connecting with a health professional (make an appointment), connecting with friends, general self-care, and postsurgical self-care.
Coded BCTs by frequency can be observed in Multimedia Appendix 4, with a glossary of BCTs presented in Multimedia Appendix 5. The mean number of BCTs identified per app was 8.5 (SD 3.6). The apps with the highest number of unique BCTs were AllyCare (16), My Exercise Messages (14), ReHand, Hand Rehabilitation (14), and PhysiApp. The most frequently coded BCTs were BCT 4.1 (instruction on how to perform a behavior; coded in n=32, 91% apps), BCT 1.4 (action planning; coded in n=30, 86% apps), BCT 2.3 (self-monitoring of behavior; coded in n=28, 80% apps), BCT 2.2 (feedback on behavior; coded in n=27, 77% apps), BCT 6.1 (demonstration of the behavior; coded in n=27, 77% apps), BCT 7.1 (prompts or cues; coded in n=25, 71% apps), and BCT 9.1 (credible source; coded in n=25, 71% apps).

Assessment of Behavior Change Potential (ABACUS)

Table 2 shows ABACUS scores for each app. The behavior change potential of included apps was a mean of 8.5 (SD 3.1) of a maximum of 21 (range 3-15). Of the 4 domains assessed by the ABACUS, section 1 (knowledge and information: mean score 3.0, SD 0.9) and section 3 (feedback and monitoring: mean score 2.9, SD 1.6) scored the highest, while section 2 (goals and planning: mean score 0.3, SD 0.8) and section 4 (actions: mean score 2.4, SD 1.0) scored the lowest. The apps (n=21) with the highest ABACUS scores included AllyCare (n=15, 71%), My Exercise Messages (n=14, 67%), ExorLive Go (n=14, 67%), A Rehab Diary (n=13, 62%), and both BlueJay Engage (n=12, 57%) and ReHand, Hand Rehabilitation (n=12, 57%).
Table 2. App Behavior Change Scale scoring.

<table>
<thead>
<tr>
<th>Application</th>
<th>Section 1: Knowledge and information</th>
<th>Section 2: Goals and planning</th>
<th>Section 3: Feedback and monitoring</th>
<th>Section 4: Actions</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Rehab Diary</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>AllyCare</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>Back Pain Diary</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>BlueJay Engage - Patient</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>ComplexCore</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>CP-Fit</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Embodia</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>ExorLive</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>Extensor</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>Guided Physio</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>HaemActive</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Home Physio</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>My Exercise Messages</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>My Exercise Program</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>My Injury Exercises</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>OT App Lite</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>PhysiApp</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>Physiotools Trainer</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Pocket Physio</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>PT Timer: Stretch &amp; Exercise</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>PT-Helper Pro</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>RecovAware Knee Health Fitness</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Rehab Guru Client</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>ReHand, Hand Rehabilitation</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Smart Therapist</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Squeezy: CF</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Squeezy for Men</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Squeezy</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>Switchback Health</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>TeleHab</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>Track Rehab</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>TrackActive Pro - Patient App</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>VR steps Home rehabilitation</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Wheelchair Exercises</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>YRMOVE</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>3.0 (0.9)</td>
<td>0.3 (0.8)</td>
<td>2.9 (1.6)</td>
<td>2.4 (1.0)</td>
<td>8.5 (3.1)</td>
</tr>
</tbody>
</table>
**Discussion**

**Principal Findings**

**Implications**

The primary aim of this study was to describe apps, intended for use by patients, that can support physiotherapy care. The secondary aims were to evaluate app quality, examine BCTs they contained, and evaluate behavior change potential. The findings of this study offer valuable insights into the current landscape of apps that may be used to support physiotherapy care and shed light on their ability to facilitate behavior change and potentially improve patient outcomes.

We identified 35 eligible apps, highlighting the popularity and increasing recognition of digital strategies for supplementing physiotherapy care [2,36]. It must also be noted that there is a growing number of apps within the broader exercise medicine and rehabilitation science space that may not have met the specific eligibility criteria for this systematic review but may still be of value in supporting physiotherapy care [1,18]. The rising number of available mobile apps necessitates careful and considered evaluation and selection of apps by clinicians and researchers to ensure that patients are accessing high-quality, effective, and safe apps that are evidence-based [1,11,37,38].

**App Characteristics**

Analysis of the app descriptive characteristics revealed interesting findings. Just under two-thirds of apps were cross-platform compatible (available on both the Google Play Store and Apple App Store). As smartphone ubiquity continues to grow, broader availability may contribute to wider accessibility and increase the uptake of apps to supplement physiotherapy care by both patients and physiotherapists [39]. The physiotherapy profession has recognized the value of mHealth tools to serve under-resourced communities and regions [40]. It is encouraging to note that in terms of price, the majority of the apps included in our review were free to download for patients. It must also be acknowledged that our review only examined apps that cost Aus $10 or less, which is in line with similar studies of apps for the management of arthritis, back pain, and persistent pain [29,31,32,41]. This follows similar research, suggesting that health consumers are less likely to buy health apps costing more than this [42]. However, there may well be more expensive apps available for purchase that were not included in our review. Conversely, while there are many seemingly free-to-download apps, several of the included apps require the physiotherapist to prescribe the patient a program from a companion app or dashboard before they can use it (eg, PhysiApp and TeleHab). In situations like this, the use of apps to support care is contingent on factors such as willingness to engage with digital health, acceptance, and digital health literacy of treating physiotherapists [2,37,38].

Relevant to an evidence-based profession like physiotherapy, the credibility of apps to support care remains an important consideration, and research suggests that this is a central factor in user engagement and satisfaction [43,44]. Less than half of the apps 16 (46%) provided clear information about the qualifications or background of the developers. Similarly, only 17 (49%) apps were clearly developed by health professionals (including physiotherapists). Research suggests that apps developed by or with health professionals may contribute to improved efficacy and safety, outcomes, and evidence-based care delivered by the apps [45]. However, while this may be true, a counterargument stands that while said to be developed by health professionals, it remains unclear what evidence was used in developing these apps.

**Quality**

A key aim of this study was to assess the quality of mobile apps relevant to physiotherapy care. While app quality is somewhat subjective, and measurement scales continue to emerge (eg, the recently released Deakin Health E-technologies Assessment Lab framework) [46], the MARS remains the most used, published, and validated [1,26,29]. Overall mean app quality in this study was 3.7 (SD 0.4), which is similar to previous research validating the MARS [47]. The authors reported a mean health app quality score of 3.74 (SD 0.6) of 5 and considered this as “moderate” quality. Notably, there is no universally accepted threshold for interpreting MARS scores. This suggests that the current landscape of physiotherapy apps has a little way to go in terms of quality. Our MARS data for the domains of “aesthetics” and “functionality” suggested that the included apps generally offer a good user experience and technical performance. These findings align with a quality appraisal of mobile apps for self-management of persistent pain conditions [29] but contrast to the relatively low scores obtained for apps specific to the management of low back pain [32].

Given that the physiotherapy profession is an evidence-based discipline, it is alarming that we found that only 4 (11%) of included apps met MARS criterion 19 (has the app been trialed or tested and published in scientific literature). Further, only 1 (3%; ReHand, Hand Rehabilitation) has been examined in an RCT [34,35,48]. However, the SMS-based precursor, which informed the development of My Exercise Messages, has undergone rigorous RCTs, showing positive patient outcomes [48,49]. An RCT studying the My Exercise Messages app is currently underway, and the protocol has been published [50]. As digital physiotherapy practice evolves, a combination of (1) greater assessment and scrutiny of digital health tools (eg, apps) through standardized validated measures and (2) high-quality well-designed RCTs and systematic reviews will assist in building evidence for and ultimately trust in digital tools that can support physiotherapy care [51-53].

**Behavior Change**

Another aim was to determine the BCTs included in apps and the apps’ potential for behavior change. Facilitating patient behavior change is an important aspect of physiotherapy care that may determine the effectiveness of many physiotherapy interventions [54-56]. Analysis of target behaviors suggested that the apps included in this review most commonly targeted behaviors linked to recording information about exercise, performing therapeutic exercise, and communicating with a health professional. This is encouraging, as these behaviors are often central to the core aims of many physiotherapeutic interventions [55,57]. The most frequently occurring BCTs coded within the included apps were instruction on how to
perform a behavior, action planning, self-monitoring of behavior, feedback on behavior, demonstration of the behavior, prompts or cues, and credible source, and these align very well to digital behavior change interventions using apps reported in literature on physical activity in osteoarthritis, various musculoskeletal conditions, and falls prevention exercise [58-60].

The potential for behavior change (based on ABACUS findings) could only be considered modest, given the mean ABACUS score of 8.5/21 (SD 3.1; 41%) across the 35 apps evaluated. This is similar to a recent comprehensive analysis of mobile apps to support behavior change in patients with a chronic disease or multimorbidity. In that study, mean ABACUS score was 8.07/21 (38%) [19]. Given that the developers of the ABACUS [24] have highlighted that the clinical implications of ABACUS score are still to be determined, it is not possible to determine whether the apps included in this study are effective or not at changing behavior or if they are likely to be effective at improving clinical outcomes from physiotherapy care. Future prospective research is required to evaluate if the apps included in our review can change patient behavior over time.

Limitations
The study has limitations. The review only analyzed apps in the English language, which limits the generalizability of the findings. Importantly, the search strategy had specific and stringent inclusion or exclusion criteria, which may lead to interpretative bias. For instance, apps had to be directly identifiable as being supportive of care between a patient and their physiotherapist. While several apps in app stores may indirectly support physiotherapy (eg, general exercise or informational apps), these were out of scope. Similarly, our search was based on a specified set of keywords, and thus, retrieval may not have captured all relevant apps. Furthermore, some apps originally identified in the search strategy were not able to be trialed for several reasons such as no free trial readily available, not available in the select region, or no response by developers for limited-time access by the study authors. A similar scenario occurred for apps that were subscription-based. In these cases, free trials were examined where possible, which may have not included full app functionality, thus potentially leading to more conservative quality, BCT, and behavioral potential scores. Regarding app quality, this was not appraised by patients themselves in this study. Should this have occurred, a different version of the MARS, the user-MARS, would have been required [61]. It is possible that patients may achieve a different quality rating for the apps. This is an area worthy of future research.

In addition, our protocol involved apps being trialed for a minimum of 10 minutes [25,29]. Vaghefi and Tulu [62] suggest “... most users tend to withdraw from mHealth apps before the end of the first week.” This raises the possibility that engaging with each app for a limited amount of time can make it difficult to get a complete picture of an app, with certain features and BCTs not being immediately apparent. This was observed firsthand in the case of “My Exercise Messages,” which scored 14/21 on the ABACUS in this study but scored 17/21 when rated by its developers who are intimately familiar with all the features of the entire 24-week app program [50]. The discrepancy is likely due to the fact that the more BCTs become apparent, the longer a user engages with the app over 6 months. A similar issue lies in the possibility that important descriptive data about an app may be missed or inaccurate because we relied on limited data sources. For instance, not all the necessary detailed information about an app can be found in the app itself or within its app store description. In select cases, this information may be identifiable if publications about the app exist. However, our protocol was deliberately designed to more closely mimic how a lay user would likely find and interact with an app through an app store. For these reasons, the findings reported in this study may be conservative for some apps, and it is possible that quality and behavior change scores would be higher with proper in-depth use. We thus urge readers to use caution in interpreting our findings.

Finally, this study focused primarily on app descriptive characteristics, quality, BCTs, and behavior change potential. It was not an aim of this review to evaluate the clinical effectiveness of these apps in changing behavior or their efficacy toward improving patient health outcomes.

Conclusions
Mobile apps available to support patient care received from a physiotherapist are of variable quality and contain relatively few BCTs. The potential for behavior change varied widely across apps. This study has provided the first comprehensive examination of mobile apps specifically supporting the care of patients receiving physiotherapy.

Acknowledgments
This study was supported by a University of Melbourne Early Career Researcher Grant Scheme (MM). KLB is supported by a National Health & Medical Research Council Leadership Investigator Grant (1174431). RSH is supported by a National Health & Medical Research Council Senior Research Fellowship (1154217).

Conflicts of Interest
RSH and KLB designed the app My Exercise Messages, which was included in this systematic review. They were not involved in the screening or data extraction or analysis processes for any app in this review. KLB also reports receiving royalties from Wolter Kluwers for production of the UpToDate knee osteoarthritis clinical guidelines.

Multimedia Appendix 1
References


24. McKay FH, Sulykerman S, Dunn M. The app behavior change scale: creation of a scale to assess the potential of apps to promote behavior change. JMIR Mhealth Uhealth 2019;7(1):e11130 [FREE Full text] [doi: 10.2196/11130] [Medline: 30681967]


57. Bassett SF. Bridging the intention-behaviour gap with behaviour change strategies for physiotherapy rehabilitation non-adherence. N Z J Physiother 2015;43(3):105-111 [FREE Full text] [doi: 10.15619/nzjp/43.3.05]


Abbreviations

ABACUS: App Behavior Change Scale  
BCT: behavior change technique  
BCTTv1: Behavior Change Technique Taxonomy version 1  
MARS: Mobile App Rating Scale  
mHealth: mobile health  
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses  
RCT: randomized controlled trial

©Mark Merolli, Jill J Francis, Patrick Vallance, Kim L Bennell, Peter Malliaras, Rana S Hinman. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 04.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Review

Quality and Accessibility of Home Assessment mHealth Apps for Community Living: Systematic Review

Jung-hye Shin1, BFA, MS, PhD; Rachael Shields1, BS, MLA; Jenny Lee1; Zachary Skrove2, BS; Ross Tredinnick3, BS, MS; Kevin Ponto1,3, BS, MS, PhD; Beth Fields2, BS, MS, PhD

1Department of Design Studies, University of Wisconsin-Madison, Madison, WI, United States
2Department of Kinesiology, University of Wisconsin-Madison, Madison, WI, United States
3Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, WI, United States

Corresponding Author:
Jung-hye Shin, BFA, MS, PhD
Department of Design Studies
University of Wisconsin-Madison
1300 Linden Dr
Madison, WI, 53706-1524
United States
Phone: 1 608 890 3704
Email: jshin9@wisc.edu

Abstract

Background: Home assessment is a critical component of successful home modifications, enabling individuals with functional limitations to age in place comfortably. A high-quality home assessment tool should facilitate a valid and reliable assessment involving health care and housing professionals, while also engaging and empowering consumers and their caregivers who may be dealing with multiple functional limitations. Unlike traditional paper-and-pencil assessments, which require extensive training and expert knowledge and can be alienating to consumers, mobile health (mHealth) apps have the potential to engage all parties involved, empowering and activating consumers to take action. However, little is known about which apps contain all the necessary functionality, quality appraisal, and accessibility.

Objective: This study aimed to assess the functionality, overall quality, and accessibility of mHealth home assessment apps.

Methods: mHealth apps enabling home assessment for aging in place were identified through a comprehensive search of scholarly articles, the Apple (iOS) and Google Play (Android) stores in the United States, and find.io. The search was conducted between November 2022 and January 2023 following a method adapted from PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). Reviewers performed a content analysis of the mobile app features to evaluate their functionality, overall quality, and accessibility. The functionality assessment used a home assessment component matrix specifically developed for this study. For overall quality, the Mobile Application Rating Scale (MARS) was used to determine the apps’ effectiveness in engaging and activating consumers and their caregivers. Accessibility was assessed using the Web Content Accessibility Guidelines (WCAG) 2.1 (A and AA levels). These 3 assessments were synthesized and visualized to provide a comprehensive evaluation.

Results: A total of 698 apps were initially identified. After further screening, only 6 apps remained. Our review revealed that none of the apps used thoroughly tested assessment tools, offered all the functionality required for reliable home assessment, achieved the “good” quality threshold as measured by the MARS, or met the accessibility criteria when evaluated against WCAG 2.1. However, DIYModify received the highest scores in both the overall quality and accessibility assessments. The MapIt apps also showed significant potential due to their ability to measure the 3D environment and the inclusion of a desktop version that extends the app’s functionality.

Conclusions: Our review revealed that there are very few apps available within the United States that possess the necessary functionality, engaging qualities, and accessibility to effectively activate consumers and their caregivers for successful home modification. Future app development should prioritize the integration of reliable and thoroughly tested assessment tools as the foundation of the development process. Furthermore, efforts should be made to enhance the overall quality and accessibility of these apps to better engage and empower consumers to take necessary actions to age in place.
Introduction

Home modifications are essential procedures for individuals with various functional limitations, enabling them to live independently within their own community. Traditionally seen as targeted biopsychosocial interventions, these modifications aim to address the functional limitations experienced by aging adults and individuals living with disabilities in their home environments. Additionally, home modifications are frequently used as part of hospital discharge planning after medical treatments such as geriatric and stroke rehabilitation [1,2]. Conducting a timely home assessment using a valid instrument and promptly implementing home modifications is crucial in assisting individuals recovering from medical procedures. These steps help maintain their functional abilities and ensure a reasonable quality of life in their homes [1,3,4].

The current gold standard for home modifications necessitates a systematic home assessment conducted by trained professionals, often performed by occupational therapists, as a prerequisite [3]. However, accessing such services remains challenging for many consumers [2,5,6]. Several contributing factors include (1) the lack of standardized and validated assessments and limited knowledge of best practices [7]; (2) the shortage of professionals trained to conduct these assessments, particularly in rural areas [6,8-10]; (3) the consumer’s perceived burden from participating in comprehensive assessments [11,12]; and (4) the complexity and cost involved in conducting home assessments [13-16].

Reliable and validated home assessment tools do exist, albeit in a paper-and-pencil format. In a systematic review of the psychometric properties of available home accessibility assessment tools, Patry et al [7] identified several tools that meet critical psychometric properties, including In-Home Occupational Performance Evaluation (I-HOPE) [17], I-HOPE Assist [18], Housing Enabler (HE) [19,20], Comprehensive Assessment and Solution Process for Aging Residents [21], and Home and Community Environment [22]. However, challenges still exist in using these tools, such as the laborious and time-consuming measurement process and the difficulty in getting reimbursement for the cost of an occupational therapist’s time [23]. The lack of objective environmental measurement has also been identified as a weakness of these tools, except for the HE [24].

The cost barrier and limitations in objectively measuring the physical environment have prompted researchers to explore the use of teleconferencing for remote home assessment tools [8,9,25]. More recently, several entities have started developing mobile health (mHealth) apps that integrate 3D modeling [26], virtual reality [27], and augmented reality (AR) tools [28,29] to measure, store, and share spatial data required for home modification solutions. However, what remains less understood and documented are the functionality, quality, and accessibility of mHealth apps. This is problematic as the number of mHealth apps for home assessment continues to increase, and there is no available evidence-informed guidance on which ones to use and why.

Therefore, the objective of this study was to systematically identify and evaluate publicly available mHealth apps available in the United States that focus on home modification in the context of aging in place, using three important criteria: (1) comprehensiveness of functionality, (2) overall quality leading to consumer engagement and follow-up actions, and (3) accessibility. A well-developed tool with all necessary functions can help professionals and consumers perceive that the app possesses the features and qualities they need to support collaboration with home modification providers in achieving desired goals [10,30].

Methods

App Identification

The research team conducted a systematic search across multiple information sources, including a database search of peer-reviewed journal papers on home assessment, the Apple (iOS) and Google Play (Android) stores in the United States, and fnio. The team followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses [31]) guidelines whenever applicable (Multimedia Appendices 1 and 2) and referred to reviews focused on mHealth apps [32-34]. The initial search was conducted between November 2022 and January 2023.

The first author (JS) conducted a search of 4 databases, including Academic Search Premier, APA PsycInfo, Consumer Health Complete—EBSCOhost, and MEDLINE. The search terms used were (“home assessment” or “home modifications” or “home mods” or “home adaptations”) AND “occupational therapy” AND (technology or application or computer or tablet or mobile phone or smartphone or internet). Titles and abstracts were then reviewed using the following criteria:

- Years considered: 1990-2022
- Language: English
- Publication status: Published
- Publication type: Includes articles in peer-reviewed journals, encompassing all types of publications
- Home assessment focus: Accessibility, covering parts of the entire house
- Types of functional limitations: All forms of functional limitations, including both physical and mental

https://mhealth.jmir.org/2024/11/e52996
Exclusion criteria: Gaming apps for occupational therapy or medical training and exclusive use of technology for communication (telehealth)

In addition to articles on individual apps, the database search yielded 3 recent research publications conducting a meta-analysis of home assessment tools [10,23,35], prompting further manual searches.

The app store search was carried out by JL. JL systematically conducted individual searches on the Google Play store using a Samsung Galaxy S21 phone and on the Apple App Store using an iPhone 11. The search terms used were consistent with those used in the database search. To ensure comprehensive coverage of potentially relevant apps, the same method was applied to searches on the fnd.io website by the author RS.

All 3 reviewers (JS, RS, and JL) convened to establish common exclusion criteria and reach consensus on the final list of apps for analysis. The exclusion criteria encompassed (1) apps not intended for home assessment; (2) unstable operations that impeded effective use, such as frequent crashes and errors; (3) regional restrictions that limited access to certain apps for US users; and (4) apps with dubious objectives, such as prioritizing product promotion over facilitating home modifications for enhanced accessibility.

Data Extraction

Between January 2023 and March 2023, three distinct tools were used to evaluate the quality of the chosen apps: an app component matrix developed for this study, the Mobile App Rating Scale (MARS) [36], and the Web Content Accessibility Guidelines (WCAG) 2.1 (created by the World Wide Web Consortium) [37].

The analysis of app components focused on evaluating the features, capabilities, and operations of each home assessment tool to understand its usefulness compared with traditional paper-and-pencil evaluations [23]. The primary objective of this evaluation was to assess the potential of each app to be effectively used by professionals or consumers in the field. Traditional and validated home assessment tools, such as I-HOPE [17] and HE [38,39], typically enable evaluators to assess the functional limitations of residents and evaluate the physical aspects of the home environment to identify necessary adjustments for the identified functional limitations. These assessments typically do not encompass suggestions for subsequent home modifications, as such considerations lie beyond the purview of home assessments. Nonetheless, given that a significant number of the reviewed apps featured recommendations, the incorporation of recommendations was introduced into the component matrix for this study. Overall, the examination centered on assessing whether each app empowers users to appraise functional limitations, conduct home environment assessments (through checklists or measurements), generate assessment outcome reports, and offer recommendations.

The MARS is a reliable and objective tool used for classifying and assessing the overall quality of mHealth apps [36]. Unlike star ratings in app stores or subjective app reviews, the MARS provides a systematic approach to evaluate mHealth apps, offering a more comprehensive and reliable measure of their quality. Studies conducted by Stoyanov et al [36] and Terhorst et al [40] have reported a high level of construct and concurrent validity, as well as reliability and objectivity, with an intraclass correlation coefficient ranging between 0.82 and 0.85. This indicates a strong level of consistency among different MARS raters, further highlighting the reliability of the scale. Moreover, the MARS has been used by researchers to assess the ability of mHealth apps to engage and activate patients [34].

The MARS consists of 3 main components: App Quality Questions, App Subjective Quality Questions, and App Specific Questions. The App Quality Questions cover various categories to provide a comprehensive evaluation of the app. These categories include engagement (A), functionality (B), aesthetics (C), and information quality (D). The engagement category assesses factors such as fun, interest, individual adaptability, interactivity, and target group. Functionality focuses on the app’s performance, usability, navigation, and gestural design. Aesthetics evaluates the layout, graphics, and visual appeal of the app. Information quality examines the accuracy, quantity, and quality of information provided, including the credibility and evidence base of the app.

In addition to the App Quality Questions, the MARS includes App Subjective Quality Questions to capture the reviewer’s personal opinion and the perceived impact on the user. The reviewer’s personal opinion (E) covers aspects such as app recommendation, willingness to pay for it, anticipated frequency of usage, and an assigned star rating. The perceived impact on the user (F) assesses how the app affects the user’s knowledge, attitudes, intentions to change, and the likelihood of actual change in the target health behavior.

Each question in the MARS is aligned with a 5-point scale (1=inadequate, 2=poor, 3=acceptable, 4=good, 5=excellent). This scale provides a standardized framework for rating the app’s performance across different categories. The unique structure and scale of the MARS allow reviewers to holistically evaluate mobile apps, considering both objective quality indicators and subjective assessments.

The research team also used the WCAG 2.1 [37]. These guidelines are designed to make web content more accessible to all users but is highly relevant to web and nonweb mobile phone content [41]. Considering that the target population of home assessment apps includes people with functional limitations, ensuring accessibility is deemed critical for the successful deployment of these apps. In comparison with the MARS rating, which primarily provides a general evaluation of mHealth apps, the WCAG assessment offers a comprehensive framework for evaluating accessibility criteria.

The WCAG 2.1 has four criteria categories: (1) perceivable, (2) operable, (3) understandable, and (4) robust. The “perceivable” category highlights the importance of users being able to perceive all presented information using their available senses. “Operable” refers to an interface that is easily navigable and usable by a wide range of users. The “understandable” category includes criteria ensuring that users should have no difficulty comprehending both the content and the user interface. Lastly, “robust” ensures that the content is consistently and accurately comprehensible.
interpreted by a diverse range of user agents, including assistive technologies.

WCAG 2.0 was initially released in December 2008 and was adopted into Section 508 of the Rehabilitation Act (29 USC 794d) in 2018. Any project receiving federal funds must adhere to Section 508 of the Rehabilitation Act. WCAG has 3 conformance levels: A, AA, and AAA, with level AAA being the highest level. To understand how WCAG operates, it is crucial to recognize that content meeting a higher level of compliance also satisfies all the criteria of a lower level. Additionally, it is worth highlighting that achieving full compliance with high-level standards is uncommon among apps, mainly due to the inherent difficulties and resource-intensive nature involved in meeting these criteria. This study used both the A and AA levels of WCAG 2.1 to evaluate the accessibility of all identified apps.

Data Analysis

The identified mHealth apps underwent testing and assessment by 3 reviewers (RS, JL, and ZS), with each being evaluated one at a time. All apps were installed and tested using an iPhone 13. The diverse training backgrounds of the 3 reviewers represented a mix of professionals likely to use or be involved in app development. ZS brings 2 years of experience in occupational therapy. RS has training and professional experience in architecture and landscape architecture, with substantial knowledge of building accessibility. JL’s training encompasses interior design, user interface design, and graphic design expertise; she also has substantial knowledge of building accessibility.

The reviewers individually examined each app and determined if it included assessment components commonly found in traditional paper-and-pencil home assessment tools: functional limitations, physical environment assessment area, mode of measurement (checklist vs measurement), final report, and recommendations. As they evaluated these components, they considered observations during app usage and information from the app store. The components were agreed on during a postassessment meeting before being included in a matrix.

The MARS rating procedures followed the recommendations specified in the original study [36]. All 3 reviewers familiarized themselves with the MARS and watched the training video to gain a thorough understanding of its components and dimensions. Before assessing apps, they engaged in extensive discussion and consensus building to clarify the meaning and relevance of each MARS assessment item.

Subsequently, the reviewers extensively used each selected app for review to gain a comprehensive understanding of its features, functionality, content, and user experience. After individually assessing the apps, the reviewers convened to reach a consensus on the final scores and tabulated the mean scores. The mean scores were compiled from each section of the MARS, namely, (A) engagement, (B) functionality, (C) aesthetics, and (D) information. The app quality mean score was calculated by averaging the scores from these 4 sections. App subjective quality (E) and app-specific (F) scores were separately averaged and treated as distinct measures. The questions in the (F) section, which evaluate the perceived impact of the app on the user’s knowledge, attitudes, intentions to change, and the likelihood of actual change in the target health behavior, were specifically responded to considering the target behavior of home assessment and modification, aligning them with the scope of this study. The agreed-on ratings were then compiled into a matrix to facilitate comparisons and analysis of ratings and data across the apps.

The accessibility assessment involved checking apps against all WCAG 2.1 criteria. During a preassessment meeting, the reviewers engaged in a detailed discussion of each of the 50 WCAG criteria. They collaboratively developed concise descriptions in the newly created WCAG evaluation form (Multimedia Appendix 3) and established a consensus on how to assess the apps based on these criteria. Individually, the reviewers assigned a pass, fail, or NA (not applicable) designation to each of the 50 WCAG criteria. The NA designation was used when a particular criterion covered a function that was not present in the app being evaluated.

In the postassessment meeting for each app, the reviewers convened to deliberate on the designations for each WCAG criterion and worked together to reach a consensus. Subsequently, the passing percentage for each WCAG category (perceivable, operable, understandable, and robust) was calculated, along with an overall passing percentage. The passing percentages were determined by excluding any criteria assigned an NA designation from the calculation.

The results of the quality and accessibility appraisal were visualized in a map. The map represents the quality appraisal using the MARS on the horizontal axis and WCAG 2.1 on the vertical axis. To enhance the objectivity of the MARS as a measure of app quality [36], we used the average score from the objective MARS items on the horizontal axis. On the vertical axis, we considered 2 levels of WCAG 2.1, namely, A and AA. Apps that scored higher on both parameters indicate high quality and accessibility, suggesting a greater potential for user engagement and activation [34,36].

Results

Identification

The initial database search yielded 104 articles, which were then reduced to 36 after eliminating duplicates. Among these, 3 meta-analyses on home assessment tools were included. After a comprehensive review of all identified articles from the database search, 3 mHealth apps were identified: HESTIA, MapIt Mobile, and Magicplan. After an initial assessment of downloadable apps, HESTIA was excluded due to its ongoing development and unavailability for download.

On engaging with the development team of MapIt, it was discovered that the desktop version of the app offered significantly enhanced functionality compared with its mobile counterpart. Because of the distinct interfaces between the 2 versions, separate evaluations were conducted for each. Although MapIt Desktop is not classified as an mHealth app, it was included in the evaluation due to its close association with the mobile app and its robustness in offering 3D
measurement capabilities in the context of home assessment, which were not found in any other apps.

The initial app store search yielded a substantial 690 apps. After reviewing titles, descriptions, preview photos, and keywords, 23 apps remained. The fnd.io search yielded a total of 253 apps, which were then narrowed down to 16 apps using the same screening method used for the app store search. After removing duplicates from all 3 sources, 18 apps remained, all of which were downloaded for eligibility evaluation. Eventually, 12 apps were eliminated due to exclusion criteria, leaving 6 apps for detailed analysis: BEAT-D, DC Carehomes, DIYModify, HomeFit AR, MapIt Mobile, and MapIt Desktop (Figure 1).

![Figure 1. Identification process of mobile health apps for home assessment in the United States.](https://mhealth.jmir.org/2024/1/e52996)

### Characteristics of the Included Apps

The characteristics of the included apps are presented in Table 1. Of the 6 apps reviewed, 4 (67%) apps were developed in university settings, 1 (17%) app was developed by a private company, and 1 app was developed by AARP, a nonprofit organization in the United States. Geographically, 3 (50%) apps (BEAT-D, DC Carehomes, and DIYModify) were developed in Australia, whereas 2 (33%) apps (MapIt Mobile and MapIt Desktop) were developed in Canada. The latter 2 apps were developed by the same entity, and they are functionally complementary to each other. Of the 6 apps, 1 (17%) app (HomeFit AR) was developed in the United States.

All the apps were designed for the iOS environment. Moreover, DC Carehomes and DIYModify were also developed for the Android platform. All apps were free to download. None of the reviewed apps were categorized as medical products nor had they published trials evaluating the effectiveness of the apps. Among the 6 apps, only 2 (33%) apps (MapIt Mobile and MapIt Desktop) allowed the users to take actual measurements of the environment. These apps provided a feature for users to gather specific measurements. In contrast, all other apps were questionnaire-based interactive decision-making tools that asked a series of questions to the users as part of the home assessment process (Tables 1 and 2).
Table 1. Operating characteristics and summary of identified mobile health apps.

<table>
<thead>
<tr>
<th>App name</th>
<th>Logo</th>
<th>Platform and operating system</th>
<th>Developer</th>
<th>Assessment characteristics: app summaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEAT-D</td>
<td>![Logo]</td>
<td>Mobile, tablet, and iOS</td>
<td>University of Wollongong, Australia</td>
<td>Questionnaire: a guided questionnaire assessment for buildings designed to accommodate individuals with dementia. The user’s responses are compiled into a comprehensive report, which identifies areas that need improvement in order to reduce confusion, agitation, and depression.</td>
</tr>
<tr>
<td>DC Carehomes</td>
<td>![Logo]</td>
<td>Mobile, web, tablet, iOS, and Android</td>
<td>Private Company: Hammond Care, Australia</td>
<td>Questionnaire: a guided questionnaire assessment for care homes, units, or households catering specifically to individuals with dementia. The app generates a comprehensive report that offers recommendations based on the assessment findings.</td>
</tr>
<tr>
<td>DIYModify</td>
<td>![Logo]</td>
<td>Mobile, iOS</td>
<td>University of New South Wales, Australia</td>
<td>Questionnaire: an interactive decision-making tool that concentrates on 5 particular home modifications and provides guidance. It helps users select appropriate product types that match their needs and offers instructions on taking necessary measurements before shopping for home modifications. The app includes real-life stories of individuals who have undergone these specific adaptations, allowing users to learn from their experiences.</td>
</tr>
<tr>
<td>HomeFit</td>
<td>![Logo]</td>
<td>Mobile, tablet, and iOS</td>
<td>Nonprofit: AARP, United States</td>
<td>Questionnaire: an interactive decision-making tool assists users in identifying potential home improvements for aging in place. The app generates a comprehensive report, including tips, suggestions, and a checklist, based on the user’s responses. The checklist distinguishes between tasks suitable for do-it-yourself and those requiring professional assistance. While the app uses AR(^a) to recognize specific features such as a kitchen sink, it does not use AR for actual measurement purposes.</td>
</tr>
<tr>
<td>MapIt</td>
<td>![Logo]</td>
<td>Mobile, tablet, iOS, and Android</td>
<td>University of Sherbrooke, Canada</td>
<td>Measurement: the app uses AR and the LiDAR(^b) sensor on the phone to create a 3D scan of a room. Users can then add measurements to specific areas of interest within the scan, catering to accessibility needs. The scan can be exported and viewed in the MapIt Desktop version, enhancing the overall measurement experience.</td>
</tr>
<tr>
<td>MapIt Desktop</td>
<td>![Logo]</td>
<td>Windows desktop, Mac OS X, and Windows</td>
<td>University of Sherbrooke, Canada</td>
<td>Measurement: this desktop app leverages MapIt on an iPhone to capture 3D scans of rooms, which can subsequently be imported to facilitate space measurements.</td>
</tr>
</tbody>
</table>

\(^a\)AR: augmented reality.  
\(^b\)LiDAR: Light Detection and Ranging.

Table 2. Functional components and mode of assessment of the reviewed apps.

<table>
<thead>
<tr>
<th>App name</th>
<th>Functional assessment</th>
<th>Environmental assessment</th>
<th>Summary report</th>
<th>Recommendation</th>
<th>Mode of assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEAT-D</td>
<td>—(^a)</td>
<td>N(^b)</td>
<td>N</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>DC Carehomes</td>
<td>—</td>
<td>N</td>
<td>N</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>DIYModify</td>
<td>—</td>
<td>T(^c)</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>HomeFit</td>
<td>—</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>Yes</td>
</tr>
<tr>
<td>MapIt Mobile</td>
<td>—</td>
<td>Generic</td>
<td>Generic</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>MapIt Desktop</td>
<td>—</td>
<td>Generic</td>
<td>Generic</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

\(^a\)Not available. \(^b\)N: nontargeted. \(^c\)T: targeted.

Component Analysis of the Included Apps

Table 2 includes a summary of the component analysis. Among the 6 apps assessed, none of them considered the functional limitations of the resident. Consequently, none of the apps enabled the evaluation of the physical environment tailored to the resident’s individual functional capacities.
DIYModify and HomeFit AR stood out by delivering concentrated assessments for certain critical spaces, notably the entrance, bathroom, kitchen, and bedroom, designated as “targeted” (T) in Table 2. DIYModify, in particular, enabled users to assess essential elements within entrance and bathroom areas, while HomeFit emphasized assessment of the bathroom, kitchen, and bedroom areas. However, it is worth noting that no single app assessed all of the key areas comprehensively. Conversely, the remaining apps either lacked specific evaluations for the designated target spaces—although they did include questions related to those areas, classified as “nontargeted” (N)—or presented generalized assessment tools adaptable to any area, categorized as “Generic” in Table 2.

The reporting modules within these apps should ideally furnish a concise overview of assessment findings on each evaluation’s conclusion, aiding users in comprehending the assessment outcomes and strategizing potential home modifications. Among the 6 apps under scrutiny, only HomeFit, DC Carehomes, and BEAT-D yielded a comprehensive report after completion of the assessment. Notably, HomeFit AR, DIYModify, DC Carehomes, and BEAT-D offered recommendations. Conversely, DIYModify omitted the provision of a report, whereas both MapIt Mobile and Desktop were deficient in both report and recommendation functionalities.

### Quality Appraisal of the Included Apps: MARS

The MARS scores for the 6 parameters of (A) engagement, (B) functionality, (C) aesthetics, (D) information, (E) subjective rating of the app overall, and (F) subjective ratings of app-specific features are presented in Table 3. Very few apps received a rating of “good” (4 or above [36]) across the measured parameters, although many of them achieved an “acceptable” (3 or above and below 4) range.

#### Table 3. Mobile Application Rating Scale (MARS) objective and subjective quality criteria and the assessment result.

<table>
<thead>
<tr>
<th>App name</th>
<th>Objective quality</th>
<th>Subjective quality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(A) engagement</td>
<td>(B) functionality</td>
</tr>
<tr>
<td>BEAT-D</td>
<td>1.6</td>
<td>3.75</td>
</tr>
<tr>
<td>DC Carehomes</td>
<td>2</td>
<td>3.5</td>
</tr>
<tr>
<td>DIYModify</td>
<td>3</td>
<td>3.75</td>
</tr>
<tr>
<td>HomeFit AR</td>
<td>2.8</td>
<td>3</td>
</tr>
<tr>
<td>MapIt Mobile</td>
<td>3</td>
<td>2.75</td>
</tr>
<tr>
<td>MapIt Desktop</td>
<td>3.2</td>
<td>3.5</td>
</tr>
</tbody>
</table>

<sup>a</sup> Apps considered “good” (4 or above) under subcategories of the MARS assessment criteria.

When using the MARS objective quality criteria, none of the apps were rated as good in the categories of (A) engagement and (B) functionality. In the (C) aesthetic category, only DIYModify achieved a good rating with a score of 4. In the (D) information category, 2 apps, namely DC Carehomes and DIYModify, scored 4 and were thus classified as good. However, when considering the overall objective quality of the apps (mean of A, B, C, and D), none of the apps reached the threshold of 4.

Ratings for the MARS subjective quality were even lower. None of the reviewed apps reached the threshold of “good” in the (E) subjective rating of the app overall, and only half of them reached an “acceptable” level. Apps performed similarly in (F) subjective ratings on the app-specific features. However, DIYModify achieved an exceptionally high score of 4.7 out of 5. The incorporation of real-life video stories showcasing practical and effective home modifications within the app contributed to the high score, as it significantly enhanced the app’s potential to positively influence the user’s knowledge, attitudes, and actual behavior change in relation to home assessment.

When considering all reviewed apps together and focusing solely on the MARS objective quality criteria, the apps showed a tendency to perform better in the (C) aesthetics category (range 3–4; mean 3.50, SD 0.35) and (D) information category (range 2.84–4; mean 3.46, SD 0.46). However, their performances in (A) engagement were much lower (range 1.6–3.2; mean 2.6, SD 0.64; Figure 2). This discrepancy may be attributed to the “customization” score under the (A) engagement category, which scored the lowest (mean 1.33, SD 0.82 out of 5). On the other hand, the “gestural design” under the (B) functionality category scored the highest (mean 4, SD 0.63 out of 5), contributing to a slightly higher overall “functionality” score.
Accessibility Appraisal of the Included Apps: WCAG 2.1

The results from the accessibility evaluation using WCAG 2.1 are presented in Table 4. None of the reviewed apps conformed to either the A or AA version of WCAG 2.1. Conformance to these standards means that there is no content that violates the success criteria [42]. Considering that the multitude of criteria and any single issue such as a broken link or lack of voice-over recognition of a piece of text can cause an app to fail, this outcome is not surprising.

When evaluated against the overall success criteria for WCAG 2.1A, which is the most used standard in the field to meet basic accessibility requirements, apps achieved a conformance rate ranging from 65% to 86%. However, this range dropped to 53% to 71% when evaluated against WCAG 2.1AA. BEAT-D and DIYModify received the highest ratings in both evaluations (Table 4).

It is worth noting that all apps fulfilled at least 1 or 2 subcriteria of the WCAG. For example, DIYModify met all criteria under the “understandable” and “robust” categories of both the A and AA versions of WCAG 2.1. BEAT-D also met both criteria but only for WCAG 2.1A. Furthermore, all apps met the “robust” criteria of WCAG 2.1A (Table 4).

On examining the individual assessment items, we found that all apps passed at least a couple of items in each success criterion. In the “perceivable” category, all apps successfully met the assessment items of info and relationships (1.3.1) and meaningful sequence (1.3.2). Similarly, in the “operable” category, all apps passed the assessment items of 3 flashes (2.3.1) and pointer cancellation (2.5.2). Moving to the “understandable” category, the apps fulfilled the assessment items of language of page (3.1.1), language of parts (3.1.2), on focus (3.2.1), consistent navigation (3.2.3), and consistent identification (3.2.4). Lastly, in the “robust” category, the apps satisfied the assessment items of parsing (4.1.1) and name, role, value (4.1.2).

However, none of the reviewed apps managed to pass 3 assessment items. These items include resize text (1.4.4), which evaluates the ability to zoom in and enlarge text; reflow (1.4.10), which assesses the ability to reflow and adjust the content to fit the screen when zoomed in; and text spacing (1.4.12), which examines the ability to customize text characteristics. All these criteria are measured against the WCAG 2.1 AA level standards.

Table 4. Web Content Accessibility Guidelines (WCAG) 2.1 assessment success criteria and the assessment.

<table>
<thead>
<tr>
<th>App name</th>
<th>Success criteria for WCAG 2.1 A</th>
<th>Success criteria for WCAG 2.1 AA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Perceivable Operable Understand-able</td>
<td>Robust</td>
</tr>
<tr>
<td></td>
<td>Perceivable Operable Understand-able</td>
<td>Robust</td>
</tr>
<tr>
<td>BEAT-D</td>
<td>4/5 7/9 5/5⁴</td>
<td>2/2²</td>
</tr>
<tr>
<td>DC Carehomes</td>
<td>4/5 6/10 73/5</td>
<td>2/2²</td>
</tr>
<tr>
<td>DIYModify</td>
<td>6/9 6/7 4/4⁴</td>
<td>2/2²</td>
</tr>
<tr>
<td>HomeFit AR</td>
<td>5/5⁴ 4/8 2/4</td>
<td>2/2²</td>
</tr>
<tr>
<td>MapIt Desktop</td>
<td>1/4 7/10 3/4</td>
<td>2/2²</td>
</tr>
<tr>
<td>MapIt Mobile</td>
<td>1/4 8/8⁴ 3/5</td>
<td>2/2²</td>
</tr>
</tbody>
</table>

⁴Apps fulfilled a subcriterion of the WCAG 2.1. These items show that all apps met at least 1 or 2 subcriteria of the WCAG 2.1 despite their failure to meet WCAG 2.1 in its entirety.

Figure 2. Box and whisker plot of the MARS objective quality assessments of all apps. MARS: Mobile App Rating Scale.
The Objective Quality and Accessibility: Visual Synthesis of MARS×WCAG

The synthesis of the results from the MARS objective quality and WCAG assessments is visually represented in Figure 3. This visualization illustrates the performance of different apps based on the MARS overall objective quality appraisal and the WCAG 2.1 accessibility criteria. Although none of the apps met the thresholds to be considered both accessible and “good” according to WCAG 2.1 A and MARS, respectively, all of them were fairly close to these thresholds. Notably, DIYModify and MapIt Desktop achieved high scores for both accessibility and the MARS objective quality. While BEAT-D performed well in terms of accessibility, there is room for improvement in its overall MARS quality. On the other hand, HomeFit AR, MapIt Mobile, and DC Carehomes scored lower in both accessibility and MARS objective quality (Figure 3).

Discussion

Principal Findings

mHealth apps are expected to empower users with the capability to assess an individual’s functional capacities and the environmental conditions crucial for comprehensive home evaluations [43]. Key areas such as the entrance, bathroom, kitchen, and bedroom hold significant importance in home assessments, which is reflected in conventional paper-and-pencil assessment tools and should thus be integral components of the app [21,38]. Furthermore, these apps should not only offer a succinct summary of the assessment but also motivate users to take up the subsequent steps after assessment, fostering engagement among all parties involved, including health care and housing professionals, as well as consumers and their caregivers. The manner in which these apps facilitate these processes should embody comprehensiveness, engagement, and accessibility.

Our findings demonstrate that, currently, there are no apps available in the United States that meet all of these criteria. Specifically, none of the apps allowed for the assessment of functional limitations of consumers, a crucial element in identifying areas requiring assessment. The MARS ratings revealed that all the apps were near the lower threshold of the acceptable range, with the exception of DIYModify; however, none of them reached the “good” range. The low scores in the “engagement” category, in particular, need further exploration as significant factors contributing to these apps’ lower ratings. Furthermore, none of the reviewed apps met the accessibility criteria.

Despite these findings, our team notes that DIYModify scored the highest in our multidimensional assessment. We also observed that when used collectively, MapIt Mobile, MapIt Desktop, and the creator’s instruction website demonstrate strong potential. Although we assessed them separately based on the parameters of our review, we found that using the entire suite together was highly effective for visualizing multiple measurements on a 3D scan. Offering an alternative viewing option on a larger screen device could prove beneficial for apps with complex user interfaces or content. On the other hand, the other apps (HomeFit AR, BEAT-D, and DC Carehomes) were
primarily questionnaire based and could have been easily accomplished without the need for an app.

**Comparison With Prior Work**

Our findings align with previous meta-analyses of home assessment tools, encompassing both paper-and-pencil formats as well as technology-assisted formats [10,23]. These studies have consistently revealed a lack of comprehensive and user-friendly technology tools that can be used to assess the home environment in relation to the functional abilities of its residents. Using technology to assess home environments in order to enhance accessibility and prevent falls and other injuries has remained challenging despite the rapid advancements in 3D modeling, virtual reality, and AR over the past few decades [10,23].

Within the limited pool of tools identified in previous studies, the majority were either pilot studies or exploratory qualitative studies [10]. Only a small fraction of these tools successfully transitioned into commercially available products, as confirmed by our search process across multiple app stores. Furthermore, review studies evaluating the efficacy of home assessment tools in all formats consistently demonstrated that the traditional paper-and-pencil assessment method was more effective in identifying issues [10]. This indicates a continued preference for the traditional method over digital alternatives among occupational therapy professionals. The findings of our study, which revealed that none of the reviewed apps allowed users to assess functional limitations, are in line with these earlier observations. This limitation hampers the effectiveness of the apps in detecting problematic areas and assessing accessibility.

Our study expands on previous research by evaluating the overall quality and accessibility of mHealth home assessment tools, emphasizing their significance in promoting consumer engagement and their follow-up actions. Specifically, we found that all of the reviewed apps met the minimum acceptable quality, but none reached the threshold of “good” quality. Additionally, none of the apps met the accessibility criteria as measured by the WCAG. This is a glaring omission as the home assessment and the subsequent modification are to help aid individuals with functional limitations in their home. To facilitate the active engagement and informed decision-making of older adults and individuals with functional limitations in their health care, as well as to empower them to undertake necessary home modifications, the development of apps that prioritize engagement, activation, and accessibility becomes imperative.

We anticipate that meeting this goal will remain challenging in the foreseeable future. Developing a mobile app may seem straightforward on the surface but can quickly become a multimillion-dollar project for several reasons. First, the development of mHealth apps for home assessment, with a focus on reliability, precision, and user-friendliness, necessitates robust interdisciplinary collaboration involving occupational therapists, building professionals, and user interface and user experience design experts.

Second, the inherent limitations in precision with current 3D scanning and AR technology, along with the need to meet the requirements of various devices, quickly add another layer of complexity to the endeavor [44-47]. Third, creating user-friendly apps requires extensive usability testing across diverse populations. In the case of home assessment, this involves testing with individuals exhibiting various functional limitations and their caregivers, contributing to the overall complexities and high cost of conducting such studies.

In contrast, the current funding landscape shows a tendency to prioritize research emphasizing cutting-edge scientific discovery or direct health outcomes with large-scale clinical trials. While the benefits of home modification have been studied either through indirect measures such as falls and emergency department visits via secondary data analysis [48,49], or a clinical trial [50], measuring the direct health outcome of the home assessment itself remains rather obscure.

Additionally, the transition from discovery to commercialization, as discussed earlier, introduces additional intricacies. Deploying the app in the market sustainably necessitates ongoing support, updates, and maintenance, contributing to the long-term cost of app development, which academic endeavors are not well suited for, often impeding the provision of free or affordable consumer apps. Above all, the lack of awareness and appreciation of the benefits that come from home assessments and home modification in the general public appears to be a key hindrance [11,12], discouraging adequate investment in this critical domain. With the rapid aging of the population and increasing interest and awareness from both the consumer market and the government alike, we hope that adequate resources are invested, fostering innovations in academia and the commercial realm alike.

**Strengths and Limitations**

Our review boasts several strengths. First, our comprehensive search encompassed scholarly databases, the US Google Play Store, Apple App Store, and ind.io, ensuring a thorough exploration of available resources. The convergence of these searches instilled confidence in the thoroughness of our efforts. Furthermore, the significant disparity between the results of the scholarly database search and the app store search shed light on the challenges associated with translating scholarly endeavors into practical applications through the commercialization process.

Another strength lies in the complementary use of multiple rigorous assessments, focusing on both quality and accessibility. For instance, the BEAT-D app achieved a high passing percentage for WCAG criteria, yet it ranked lower in the MARS evaluation. On the other hand, MapIt Desktop had a lower score in WCAG, but ranked higher in the MARS. This discrepancy emphasizes the importance of conducting both assessments, particularly in the context of home assessment tools where individuals with functional limitations play a crucial role.

Finally, while evaluating apps with the MARS assessment is standardized and relatively straightforward, assessing apps based on WCAG criteria requires a more substantial time investment to grasp each of the 50 criteria, which is extensively explained on the WCAG’s website. To facilitate the use of WCAG, our research team has developed an evaluation form that includes
concise summaries of each criterion, which can be used in future studies, streamlining the evaluation process (Multimedia Appendix 1).

The review also presents certain weaknesses that deserve attention. First, despite our efforts to conduct a comprehensive search for all available home assessment tools, the inclusion of apps was limited by their availability in the US market. It is worth highlighting that several apps, discovered via our database search, find.io search, and personal connections, demonstrate promise but remain unavailable on US app stores.

Second, our testing of multiplatform apps was focused solely on the iPhone versions of all reviewed apps. This limitation was observed during the assessment of WCAG criterion 1.3.4 (orientation), which assesses the device’s capacity to transition between portrait and landscape modes. Notably, the iPhone variant of BEAT-D did not meet this criterion, while the iPad version might have passed had it been evaluated. To ensure a more comprehensive assessment, it would be advantageous to test these apps on all compatible devices, including those running on Android operating systems.

Finally, while assessing apps with the MARS and WCAG provided valid insights into the overall quality and accessibility of apps based on established criteria, future studies will have to take into consideration consumer-level feedback, particularly focusing on the firsthand experiences of those with various functional capacities and their caregivers.

Conclusions
A proficient home assessment tool, designed to engage consumers, health care providers, and housing professionals, should offer reasonable functionality and possess objective quality and accessibility. However, our findings bring to light that none of the currently available home assessment mHealth apps in the United States align with these benchmarks. None of the apps offered sufficient methods to assess individuals’ functional capacity and conduct comprehensive environmental assessments, and they fell short of meeting the WCAG accessibility criteria. Furthermore, although every app reached an “acceptable” level, none of them attained a “good” level in the MARS quality evaluations.

Acknowledgments
The review received funding from the RRF Foundation of Aging, the Fall Research Competition, and the School of Human Ecology at the University of Wisconsin-Madison.

Authors’ Contributions
JS and RS played a leading role in conceptualizing, designing, and planning the review. App identification involved the contributions of JS, RS, and JL. Content analysis and quality or accessibility appraisal were conducted by RS, JL, and ZS under the supervision of JS. The initial draft of the manuscript was prepared by JS, incorporating the key technical findings provided by RS. All authors contributed to the manuscript by providing comments and suggestions for improvement, ultimately approving the final version.

Conflicts of Interest
None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[PDF File (Adobe PDF File), 106 KB - mhealth_v121e52996_app1.pdf]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 for abstracts checklist.
[PDF File (Adobe PDF File), 65 KB - mhealth_v121e52996_app2.pdf]

Multimedia Appendix 3
Accessibility evaluation form adapted from Web Content Accessibility Guidelines (WCAG) 2.1.
[DOC File, 101 KB - mhealth_v121e52996_app3.doc]

References


42. Cooper M, Kirkpatrick A, O'Connor J. Understanding conformance, understanding WCAG 2.0: a guide to understanding AA Success Criteria. URL: https://www.w3.org/TR/WCAG21/conformance.html [accessed 2024-02-01]

Abbreviations

AR: augmented reality
HE: Housing Enabler
I-HOPE: In-Home Occupational Performance Evaluation
MARS: Mobile Application Rating System
mHealth: mobile health
NA: not applicable
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
WCAG: Web Content Accessibility Guidelines

© Jung-hye Shin, Rachael Shields, Jenny Lee, Zachary Skrove, Ross Tredinnick, Kevin Ponto, Beth Fields. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 11.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Original Paper

Short-Term Effect of a Health Promotion Intervention Based on the Electronic 12-Hour Dietary Recall (e-12HR) Smartphone App on Adherence to the Mediterranean Diet Among Spanish Primary Care Professionals: Randomized Controlled Clinical Trial

Luis María Béjar1, PhD; Pedro Mesa-Rodríguez2, BSc; María Dolores García-Perea3, PhD

1Department of Preventive Medicine and Public Health, Institute of Anatomy, School of Medicine, University of Seville, Seville, Spain
2Camas Health Center, Seville, Spain
3Virgen Macarena University Hospital, Seville, Spain

Corresponding Author:
Luis María Béjar, PhD
Department of Preventive Medicine and Public Health
Institute of Anatomy
School of Medicine, University of Seville
3rd Floor, Sánchez-Pizjuán Avenue
Seville, 41009
Spain
Phone: 34 954551771
Fax: 34 954556481
Email: lmbprado@us.es

Abstract

Background: The World Health Organization has called for addressing the growing burden of noncommunicable diseases (NCDs) by promoting healthy lifestyles among the population. Regarding patient health, primary care professionals (PCPs) are the first line of care who can positively influence patients’ behavior and lifestyle habits. However, a significant percentage of PCPs do not lead a healthy lifestyle. Therefore, addressing their health behaviors may be the key to substantially increasing health promotion advice in general practice. The Mediterranean diet has been extensively studied, and there is strong evidence of it being a dietary pattern for the prevention of NCDs, in addition to its significant environmental, sociocultural, and local economics benefits.

Objective: This study focused only on the dietary aspect of the PCPs’ lifestyle. The primary objective was to evaluate the effect of using the Electronic 12-Hour Dietary Recall (e-12HR) smartphone app to improve diet, specifically to promote adherence to the Mediterranean diet (AMD), among PCPs. The secondary objectives were to establish the usability of the e-12HR app and to determine AMD among PCPs.

Methods: An individual-level randomized, controlled, and single-blind clinical trial was conducted with 2 parallel groups: a control group (CG), using the nonfeedback version of the e-12HR app, and an intervention group (IG), using the feedback version of the e-12HR app. The level of human involvement was fully automated through the use of the app. There was a 28-day follow-up period. Participants were PCPs (medicine or nursing) recruited offline at one of the selected primary care centers (Andalusia, Spain), of both sexes, over 18 years old, possessing a smartphone, and having smartphone literacy.

Results: The study response rate was 73% (71 of 97 PCPs), with 27 (38%) women and 44 (62%) men: 40 (56%) PCPs in the CG and 31 (44%) in the IG. At baseline, AMD was medium (mean Mediterranean Diet Serving Score [MDSS] index 9.45, range 0-24), with 47 (66%) PCPs with a medium/high MDSS index. There were significant statistical improvements (CG vs IG, in favor of the IG) at week 4 (no significant statistical differences at baseline): +25.6% for the MDSS index ($P=.002$) and +213.1% for the percentage with a medium/high MDSS index ($P=.001$). In relation to specific food groups, there were significant statistical improvements for fruits (+33.8%, $P=.02$), vegetables (+352%, $P=.001$), nuts (+184%, $P=.02$), and legumes (+75.1%, $P=.03$). The responses to the usability rating questionnaire were satisfactory.
Conclusions: The results support recommending the use of the e-12HR app as a tool to contribute to improving diet and preventing NCDs among PCPs, while positively influencing patient dietary behavior and preventing diet-related NCDs among patients.

Trial Registration: ClinicalTrials.gov NCT05532137; https://clinicaltrials.gov/study/NCT05532137

KEYWORDS
primary care professionals; Mediterranean diet; smartphone applications; smartphone apps; health promotion; Mediterranean diet adherence; food group

Introduction

Major chronic noncommunicable diseases (NCDs; eg, cardiovascular diseases, cancer, chronic respiratory diseases, and diabetes) are responsible for 74% of deaths worldwide, making them the leading cause of preventable mortality [1]; they also reduce patients’ well-being and activity, which contributes to a poor quality of life, disability, and reduced productivity [1]. Notably, these NCDs share common risk factors, such as an unhealthy diet, smoking, harmful alcohol consumption, and physical inactivity [1], all of which are behavioral or lifestyle-related factors that are potentially modifiable [2]. For this reason, the World Health Organization has called for addressing the growing burden of NCDs by promoting healthy lifestyles [1,3].

Health care workers are a subgroup of the population where lifestyle promotion is essential for 3 main reasons. The first reason is their own health: although health care professionals do their best to provide exceptional patient care, they often fail to prioritize their own self-care [4], resulting in unhealthy behaviors associated with a high risk of NCDs. Studies of health care workers’ lifestyles in hospitals, for example, showed high rates of overweight/obesity [5-7], low fruit and vegetable consumption [6,7], low physical activity [5-8], tobacco [5,7] and alcohol consumption [6,7], and high levels of stress and insufficient rest [4-6]. In addition, almost half had more than 2 of these risk factors [5,6].

The second reason is the success of health care organizations: health care workers with better and healthier habits have been shown to have higher personal and job satisfaction and fewer sickness-related absences from work [9]. The third reason is patient health: several studies have described that health care professionals who practice healthy behaviors offer more advice on healthy habits to patients who come to their practice, that they have more assertive attitudes when counselling patients, and they provide more comprehensive and aggressive counselling, which can positively influence patients’ health [10-14].

Regarding patient health, primary care is the cornerstone for preventing NCDs through health education and plays an essential role in the success of therapeutic medicine [11]. Primary care professionals (PCPs) are the first line of care who can positively influence patient behavior and lifestyle habits [15], and indeed, patients perceive PCPs (doctors and nurses) as the most trusted source of health information and advice on healthy lifestyles [13]. However, a significant percentage of PCPs, as with hospital health care professionals, do not lead a healthy lifestyle: there is a high prevalence of NCD risk factors, such as smoking [10,16], alcohol abuse [10,16], overweight/obesity [17], physical inactivity [5,10,17], inadequate fruit and vegetable intake [10,11,17], added salt intake [17], and high consumption of ultraprocessed foods [18], sugars, and fats [10]. Therefore, addressing PCPs’ health behaviors may be the key to substantially increasing health promotion advice in general practice [12].

As the evidence shows, PCPs are exposed to several risk factors of NCDs; this study focused only on the dietary aspect of the PCPs’ lifestyle. As a starting point, the research team posed the following question: How can the dietary habits of PCPs be improved?

The research team has previously evaluated the effectiveness of a smartphone app called the Electronic 12-Hour Dietary Recall (e-12HR) in improving diet in Spanish university students (health science [19,20] and non–health science [20]). The main hypothesis of this study was that the use of this app among PCPs can have an influence on improving their diet, as has already been evidenced in previous research among university students [19,20], the null hypothesis being that the use of the app has no influence on improving the diet of PCPs.

To the best of our knowledge, this study is the first to evaluate the use of a smartphone app to improve diet among PCPs. The primary study objective was to evaluate the effect of using the e-12HR app on improving diet among PCPs, specifically to promote adherence to the Mediterranean diet (AMD). The Mediterranean diet has been extensively studied, and there is strong evidence of it being a dietary pattern for the prevention of NCDs [21-25], in addition to its significant environmental, sociocultural, and local economics benefits [26-28]. In addition, secondary objectives were to establish the usability of the e-12HR app and determine AMD among PCPs.

Methods

Overview of the Study

The study was an individual-level randomized, controlled, and single-blind clinical trial with 2 parallel groups: a control group (CG) and an intervention group (IG). All participants used an app called the e-12HR app, with different versions for the CG and the IG. In the CG, participants used the nonfeedback version of the e-12HR app, and in the IG, participants used the feedback version of the e-12HR app (see the Intervention section). The
level of human involvement was fully automated through the use of the app.

All research was carried out in 3 basic health zones of the Andalusian Health Service (Andalusia, Spain, South of Europe): Camas, Coria del Río, and San Juan de Aznalfarache. Several primary care centers were selected in the 3 zones: the Camas, Santiponce, Valencina de la Concepción, Castilleja de Guzmán, Coca de la Piñera, Carambolo, and Pañoleta health centers in Camas; the Coria del Río health center in Coria del Río; and the San Juan de Aznalfarache and Gelves health centers in San Juan de Aznalfarache. The study ran for 28 days, and participant recruitment took place offline from September to October 2022.

Inclusion criteria for the study were both sexes, age over 18 years, possession of a smartphone (iOS or Android operating system), smartphone literacy, and a PCP (medicine or nursing) at one of the selected primary care centers. Exclusion criteria included food intolerance, chronic disease, or pregnancy (due to the possibility of requiring specific dietary recommendations).

Ethical Considerations

The study was conducted according to the guidelines laid down in the Declaration of Helsinki, and all procedures involving study participants were approved by the Andalusian Biomedical Research Ethics Portal (PEIBA) on March 30, 2022 (identifier: 2813-N-21). The trial was registered at ClinicalTrials (identifier: NCT05532137). Written informed consent was obtained from all participants.

Participant Enrollment

The project was publicized in the selected primary care centers by a member of the research team, and individual talks were scheduled with interested health care professionals. In each of the individual talks, the study protocol was explained, including the objectives, risks, and benefits of the research, and an email address for the study was provided.

To participate in the study, it was necessary for interested health care professionals to send an email to the designated address, indicating their “interest in participating in the study” and the primary care center where they worked. After receiving the email, a member of the research group sent candidates a series of documents necessary to be able to participate in the study: (1) an informed consent form; (2) a form with personal information (sex, date of birth, primary care center, weight, height, and smoking status), with documents 1 and 2 to be completed, signed, and returned to the same email address; (3) instructions for downloading the e-12HR app (free to download from the App Store or the Play Store); (4) an image of the Mediterranean diet pyramid (with recommendations for consumption by food group); (5) a personal alphanumeric code; and (6) a user’s guide with detailed information for using the app. Document 6 was the only one that differed depending on whether it was intended for participants in the CG (nonfeedback e-12HR version) or the IG (feedback e-12HR version), and obviously, document 5, which contained the personal code, was unique for each participant.

Throughout the study, participants could contact the research team by email with any questions, including questions to reduce the likelihood of harm.

Participant Allocation

In each basic health zone, the recruited participants were randomized into 1 of 2 groups (CG or IG) in a ratio of 1:1 as follows: the participant who sent the first email was assigned to the CG, the participant who sent the second email was assigned to the IG, and so on.

This study was single-blind because, due to the nature of the intervention, the PCPs could not be blinded. However, the investigator who performed the statistical analysis of the data was blinded throughout the study. In addition, each participant only had access to 1 version of the app: the personal codes of the participants assigned to the CG activated only the nonfeedback e-12HR version, while the personal codes of the participants assigned to the IG activated only the feedback e-12HR version.

Intervention

The structure and functions of the e-12HR app (nonfeedback and feedback versions) have been described in detail by Béjar et al [19]. In this study, we used e-12HR version 3.0. The e-12HR app did not undergo changes throughout the study. In brief, the nonfeedback e-12HR version allows the user to collect food consumption data; however, this version does not provide any feedback to users to promote the Mediterranean diet (ie, this version of the app presents a single function: diet determination). The feedback e-12HR version allows for the collection of food consumption data, and as an additional automatic function, every 7 days the app issues personalized feedback on how to improve AMD (ie, this version of the app has 2 functions: determining the diet and providing feedback to improve AMD). The feedback provided came in 3 parts: (1) the AMD index score: specifically, the Mediterranean Diet Serving Score (MDSS) index [29] (range 0-24); (2) the image of a traffic light: the MDSS index score was divided into 1 of 3 levels (low: score 0-8, red light; medium: score 9-15, orange light; high: score 16-24, green light) [30]; and (3) recommendations for consumption by food group [19]. The nonfeedback e-12HR version did not provide any of the 3 parts of the feedback, as they were exclusive to the e-12HR feedback version. See Multimedia Appendix 1 for real images of the e-12HR app (nonfeedback and feedback versions).

Follow-up and Outcome Measures

To analyze the effect of the intervention (CG nonfeedback e-12HR vs IG feedback e-12HR), 4 follow-up points were established: week 1 (baseline), week 2, week 3, and week 4. At each follow-up point, the research team manually calculated the MDSS index for each of the 2 groups from the data provided by the e-12HR app. The method for calculating the MDSS index has been described in detail elsewhere [19].

The main result variable was the change in the total MDSS index at weeks 2, 3, and 4 of monitoring, while the secondary result variables were the personal information variables, the MDSS index at week 1 (baseline), and the answers to the
usability rating questionnaire for the e-12HR app (see the Usability Rating Questionnaire for the e-12HR section).

The MDSS index at week 1 (baseline) was used to determine AMD among the PCPs (a secondary objective of the study). To relativize the data, the MDSS index of the PCPs was compared to the MDSS index of health sciences students. For a proper comparison, the MDSS index was obtained using the same app (e-12HR) and during the same follow-up period (recruitment period: September-October 2022) for both students and health care professionals.

Usability Rating Questionnaire for the e-12HR

After the 4-week study period, a member of the research team sent a new email to each PCP who had completed the follow-up; this new email contained a usability rating questionnaire for the e-12HR app [19] (see Multimedia Appendix 2).

Statistical Analysis

The sample size was estimated for the main result variable. Assuming SD=2.7 points, dropout rate=20.6% (from a previous study on use of the e-12HR app among health science university students [19]), α=.05, and β=.20 (bilateral test), 82 participants (n=41, 50%, per group) were needed to detect an increase of 2 points in the MDSS index (CG versus IG). The sample size was calculated using nQuery Advisor Release 7.0 (Statsols).

Quantitative variables were expressed as means (SD), and qualitative variables were displayed as numbers (percentages). The nonparametric Kolmogorov-Smirnov test was used for the test for normality.

For unpaired samples and quantitative variables, the Student t test or the nonparametric Mann-Whitney U test was used, and the chi-square test (or Fisher exact test) was used for the comparison of proportions.

For paired samples, quantitative variables, and two groups, the Student t test or the nonparametric Wilcoxon test was used, penalizing P values with Bonferroni adjustment for multiple comparisons. For 3 or more groups, the ANOVA test or the nonparametric Friedman test was used.

P<.05 was considered significant, except for multiple comparisons using Bonferroni penalization: P<.02 (.05/3).

All statistical analyses were performed using the SPSS statistical software package version 26.0 (SPSS Inc).

Results

Sample and Adherence to the Study

The sequence of allocating participants to the 2 study groups is detailed in Figure 1. In total, 97 PCPs signed the informed consent form (n=50, 52%, in the CG and n=47, 48%, in the IG). Of them, 26 (27%; n=10, 38%, in the CG and n=16, 62%, in the IG) were considered nonresponsive because they did not complete the study’s 4-week follow-up period (Figure 1). The nonresponsive individuals were not included in the later statistical analysis (ie, in this study, per protocol analysis was applied).

Figure 1. Flowchart of the study. e-12HR: Electronic 12-Hour Dietary Recall.
The study response rate was 73% (71 of 97 participants), with 40 of 50 (80%) participants in the CG and 31 of 47 (66%) participants in the IG (Figure 1). Participants did not report any harm or unintended effects throughout the study.

Table 1. Characteristics of participants who completed the study on the short-term effects of a health promotion intervention based on the e-12HR smartphone app on AMD among Spanish PCPs.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>All participants (N=71)</th>
<th>CG (n=40)</th>
<th>IG (n=31)</th>
<th>P value&lt;sup&gt;f&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>43.2 (11.6)</td>
<td>45.1 (11.1)</td>
<td>40.7 (12.1)</td>
<td>.07&lt;sup&gt;g&lt;/sup&gt;</td>
</tr>
<tr>
<td>&lt;40, n (%)</td>
<td>34 (47.9)</td>
<td>16 (40.0)</td>
<td>18 (58.1)</td>
<td>.13&lt;sup&gt;h&lt;/sup&gt;</td>
</tr>
<tr>
<td>≥40, n (%)</td>
<td>37 (52.1)</td>
<td>24 (60.0)</td>
<td>13 (41.9)</td>
<td>—</td>
</tr>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>27 (38.0)</td>
<td>16 (40.0)</td>
<td>11 (35.5)</td>
<td>.70&lt;sup&gt;i&lt;/sup&gt;</td>
</tr>
<tr>
<td>Male</td>
<td>44 (62.0)</td>
<td>24 (60.0)</td>
<td>20 (64.5)</td>
<td>—</td>
</tr>
<tr>
<td><strong>BMI (kg/m&lt;sup&gt;2&lt;/sup&gt;)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>25.0 (4.0)</td>
<td>25.3 (4.4)</td>
<td>24.7 (3.4)</td>
<td>.76&lt;sup&gt;g&lt;/sup&gt;</td>
</tr>
<tr>
<td>&lt;25, n (%)</td>
<td>41 (57.7)</td>
<td>23 (57.5)</td>
<td>18 (58.1)</td>
<td>.96&lt;sup&gt;h&lt;/sup&gt;</td>
</tr>
<tr>
<td>≥25, n (%)</td>
<td>30 (42.3)</td>
<td>17 (42.5)</td>
<td>13 (41.9)</td>
<td>—</td>
</tr>
<tr>
<td><strong>Smoking status, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>64 (90.1)</td>
<td>37 (92.5)</td>
<td>27 (87.1)</td>
<td>.45&lt;sup&gt;h&lt;/sup&gt;</td>
</tr>
<tr>
<td>Yes</td>
<td>7 (9.9)</td>
<td>3 (7.5)</td>
<td>4 (12.9)</td>
<td>—</td>
</tr>
<tr>
<td><strong>Physical activity status (minutes/week), n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥150</td>
<td>45 (63.4)</td>
<td>24 (60.0)</td>
<td>21 (67.7)</td>
<td>.50&lt;sup&gt;h&lt;/sup&gt;</td>
</tr>
<tr>
<td>&lt;150</td>
<td>26 (36.6)</td>
<td>16 (40.0)</td>
<td>10 (32.3)</td>
<td>—</td>
</tr>
</tbody>
</table>

<sup>a</sup>e-12HR: Electronic 12-Hour Dietary Recall.
<sup>b</sup>AMD: adherence to the Mediterranean diet.
<sup>c</sup>PCP: primary care professional.
<sup>d</sup>CG: control group.
<sup>e</sup>IG: intervention group.
<sup>f</sup>P<.05 considered significant.
<sup>g</sup>Evaluated with the Mann-Whitney U test.
<sup>h</sup>Evaluated with the chi-square test.
<sup>i</sup>Not applicable.

No significant statistical differences were observed in the personal variables studied between the CG and the IG (Table 1). PCPs (n=47, 66%, doctors and n=24, 34%, nurses) recorded their daily consumption using the e-12HR app for a total of 1988 days (N=71 participants × 28 recording days). The app differentiated 19 food groups. Thus, a total of 37,772 daily food group consumption data points were collected during the study.

There were no significant statistical differences in the personal variables between responsive (those who completed the study) and nonresponsive (those who did not complete the study) participants.

**MDSS Index**

As previously mentioned, the MDSS index was calculated manually by the research team (CG and IG, weeks 1, 2, 3, and 4) [19]. During the process, the research team corrected obvious errors made by the PCPs: for example, when the app asked for the number of standard servings of a certain food group consumed on the current day, if the participant entered a value of 150, it was considered that the data referred to milliliters or grams (instead of standard servings). In any case, only 0.2% (91 of 37,772) of the recorded data were corrected.

At week 1 of the monitoring period (baseline), that is, before the IG received the first feedback from the e-12HR app (only week 1 was considered because the feedback for IG in weeks 2, 3, and 4 could affect the alteration of the usual dietary intake), participants did not report any harm or unintended effects throughout the study.
the PCPs had a mean MDSS index of 9.45 (SD 2.32), which corresponds to a medium level of adherence [30]; moreover, two-thirds of them (n=47, 66%) had a medium/high MDSS value (≥9) at baseline (week 1) [30].

**Effect of the Intervention With the e-12HR App in Terms of Variation in the MDSS Index and Number of Participants With Medium/High (≥12) MDSS Index**

Tables 2 and 3 show the MDSS index, and Table 2 also shows the number of participants with a medium/high MDSS index (≥12) in the CG and the IG throughout the 4 weeks of follow-up. We decided to use the value of 12 (instead of 9) [30] due to the high percentage of PCPs (n=47, 66%) who at baseline (week 1) already had an MDSS index≥9; therefore, using the value of 9 would have made it difficult to observe statistically significant differences between the CG and the IG.

### Table 2. MDSS\(^a\) index for the CG\(^b\) and the IG\(^c\) and number of participants with a medium/high (≥12) MDSS index throughout the 4 weeks of follow-up.

<table>
<thead>
<tr>
<th>Variables and week number</th>
<th>CG (n=40)</th>
<th>IG (n=31)</th>
<th>(P) value(^d)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MDSS index, mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 1</td>
<td>9.30 (2.40)</td>
<td>9.65 (2.23)</td>
<td>.54(^e)</td>
</tr>
<tr>
<td>Week 2</td>
<td>8.98 (2.84)</td>
<td>10.81 (2.82)</td>
<td>.009(^e)</td>
</tr>
<tr>
<td>Week 3</td>
<td>9.08 (2.45)</td>
<td>10.94 (3.05)</td>
<td>.008(^e)</td>
</tr>
<tr>
<td>Week 4</td>
<td>9.30 (2.59)</td>
<td>11.68 (3.61)</td>
<td>.002(^f)</td>
</tr>
<tr>
<td>(P) value(^e)</td>
<td>.99</td>
<td>.01</td>
<td>—</td>
</tr>
<tr>
<td><strong>Participants with a medium/high (≥12) MDSS index, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 1</td>
<td>10 (25.0)</td>
<td>10 (32.3)</td>
<td>.50</td>
</tr>
<tr>
<td>Week 2</td>
<td>7 (17.5)</td>
<td>13 (41.9)</td>
<td>.02</td>
</tr>
<tr>
<td>Week 3</td>
<td>5 (12.5)</td>
<td>13 (41.9)</td>
<td>.005</td>
</tr>
<tr>
<td>Week 4</td>
<td>7 (17.5)</td>
<td>17 (54.8)</td>
<td>.001</td>
</tr>
<tr>
<td>(P) value(^e)</td>
<td>.41</td>
<td>.07</td>
<td>—</td>
</tr>
</tbody>
</table>

\(^a\)MDSS: Mediterranean Diet Serving Score.  
\(^b\)CG: control group.  
\(^c\)IG: intervention group.  
\(^d\)\(P\) value in columns: MDSS index—intragroup differences (CG and IG) throughout the 4 weeks of follow-up in the study, evaluated with the Friedman test; number of participants with a medium/high (≥12) MDSS index—intragroup differences (CG and IG) in week 1 versus week 4, evaluated with the chi-square test. \(P<.05\) was considered significant.  
\(^e\)Evaluated with the Student \(t\) test.  
\(^f\)Evaluated with the Mann-Whitney \(U\) test.  
\(^g\)\(P\) value in rows: MDSS index—intergroup differences (CG versus IG) in each of the 4 study weeks; number of participants with a medium/high (≥12) MDSS index—intergroup differences (CG versus IG) in each of the 4 study weeks, evaluated with the chi-square test. \(P<.05\) was considered significant.  
\(^h\)Not applicable.
Table 3. Comparison of the MDSS\(^a\) index in weeks 2, 3, and 4 of follow-up with that in week 1 (baseline) for the CG\(^b\) and the IG\(^c\).

<table>
<thead>
<tr>
<th>Group and week</th>
<th>MDSS index, mean (SD)</th>
<th>(P) value(^d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 1</td>
<td>9.30 (2.40)</td>
<td>Reference</td>
</tr>
<tr>
<td>Week 2</td>
<td>8.98 (2.84)</td>
<td>.34(^e)</td>
</tr>
<tr>
<td>Week 3</td>
<td>9.08 (2.45)</td>
<td>.58(^e)</td>
</tr>
<tr>
<td>Week 4</td>
<td>9.30 (2.59)</td>
<td>.99(^f)</td>
</tr>
<tr>
<td>IG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 1</td>
<td>9.65 (2.23)</td>
<td>Reference</td>
</tr>
<tr>
<td>Week 2</td>
<td>10.81 (2.82)</td>
<td>.005(^f)</td>
</tr>
<tr>
<td>Week 3</td>
<td>10.94 (3.05)</td>
<td>.004(^e)</td>
</tr>
<tr>
<td>Week 4</td>
<td>11.68 (3.61)</td>
<td>.001(^f)</td>
</tr>
</tbody>
</table>

\(^a\)MDSS: Mediterranean Diet Serving Score.\n\(^b\)CG: control group.\n\(^c\)IG: intervention group.\n\(^d\)Intragroup differences (CG and IG) in week 1 versus weeks 2, 3, and 4. \(P<.02 (.05/3)\) was considered significant (penalizing \(P\) values with Bonferroni adjustment for multiple comparisons).\n\(^e\)Evaluated with the Wilcoxon test.\n\(^f\)Evaluated with the Student \(t\) test.

Regarding intragroup modifications, there were no significant statistical differences in the MDSS index in the CG, either throughout the 4 weeks of study, ranging from mean 9.30 (SD 2.59) in weeks 1 and 4 to mean 8.98 (SD 2.84) in week 2 (Table 2) or in weeks 2, 3, and 4 when compared to week 1 (baseline; Table 3). In the IG, there were significant statistical differences in the MDSS index throughout the 4 weeks of study, ranging from mean 9.65 (2.23) in week 1 to mean 11.68 (SD 3.61) in week 4 (Table 2). Compared to week 1 (baseline), the differences were statistically significant from week 2 onward: with 1.16, 1.29, and 2.03 points of improvement at weeks 2, 3, and 4, respectively (Table 3). There were no significant statistical differences in the number of participants with a medium/high (\(\geq 12\)) MDSS index (week 1 vs week 4) in either the CG or the IG.

Regarding intergroup modifications, there were significant statistical differences for both the MDSS index and the number of participants with a medium/high (\(\geq 12\)) MDSS index in the CG versus the IG (in favor of the IG) from week 2 onward (no significant differences in week 1). For the MDSS index, we found 1.83, 1.86, and 2.38 points of improvement at weeks 2, 3, and 4, respectively; for the number of participants with a medium/high (\(\geq 12\)) MDSS index, we found 24.4, 29.4, and 37.3 percentage points of improvement at weeks 2, 3, and 4, respectively (Table 2).

**Effect of the Intervention With the e-12HR App in Terms of Variation in Food Groups**

Table 4 shows the number of participants who met the consumption criteria for each food group [19] in the CG and the IG throughout the 4 weeks of follow-up.
Table 4. Number of participants who met the consumption criteria of the MDSS index for each food group throughout the 4 weeks of follow-up (CG n=40, IG n=31).

<table>
<thead>
<tr>
<th>Food group</th>
<th>MDSS index consumption criteria and study group</th>
<th>Week 1</th>
<th>Week 2</th>
<th>Week 3</th>
<th>Week 4</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fruits (1-6 servings/day)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>26 (65.0)</td>
<td>27 (67.5)</td>
<td>27 (67.5)</td>
<td>27 (67.5)</td>
<td>.81e</td>
<td></td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>24 (77.4)</td>
<td>27 (87.1)</td>
<td>28 (90.3)</td>
<td>28 (90.3)</td>
<td>.17e</td>
<td></td>
</tr>
<tr>
<td>P value</td>
<td>.26f</td>
<td>.06g</td>
<td>.02g</td>
<td>.02g</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vegetables (≥2 servings/day)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>1 (2.5)</td>
<td>3 (7.5)</td>
<td>3 (7.5)</td>
<td>4 (10.0)</td>
<td>.36i</td>
<td></td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>3 (9.7)</td>
<td>8 (25.8)</td>
<td>10 (32.3)</td>
<td>14 (45.2)</td>
<td>.002e</td>
<td></td>
</tr>
<tr>
<td>P value</td>
<td>.31j</td>
<td>.05j</td>
<td>.01f</td>
<td>.001f</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cereals (1-6 servings/day of breakfast cereals, pasta, rice, and bread)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>40 (100.0)</td>
<td>37 (92.5)</td>
<td>39 (97.5)</td>
<td>37 (92.5)</td>
<td>.241i</td>
<td></td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>31 (100.0)</td>
<td>30 (96.8)</td>
<td>30 (96.8)</td>
<td>30 (96.8)</td>
<td>.99i</td>
<td></td>
</tr>
<tr>
<td>P value</td>
<td>—</td>
<td>.63j</td>
<td>.99j</td>
<td>.63j</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Olive oil (1-4 servings/day)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>31 (77.5)</td>
<td>33 (82.5)</td>
<td>33 (82.5)</td>
<td>34 (85.0)</td>
<td>.390e</td>
<td></td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>26 (83.9)</td>
<td>26 (83.9)</td>
<td>28 (90.3)</td>
<td>27 (87.1)</td>
<td>.99h</td>
<td></td>
</tr>
<tr>
<td>P value</td>
<td>.50j</td>
<td>.88f</td>
<td>.50j</td>
<td>.99j</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Milk and dairy products (1-3 servings/day)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>33 (82.5)</td>
<td>32 (80.0)</td>
<td>32 (80.0)</td>
<td>33 (82.5)</td>
<td>.99e</td>
<td></td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>26 (83.9)</td>
<td>27 (87.1)</td>
<td>27 (87.1)</td>
<td>28 (90.3)</td>
<td>.71i</td>
<td></td>
</tr>
<tr>
<td>P value</td>
<td>.88g</td>
<td>.43g</td>
<td>.43g</td>
<td>.50j</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nuts (1-2 servings/day)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>3 (7.5)</td>
<td>2 (5.0)</td>
<td>2 (5.0)</td>
<td>5 (12.5)</td>
<td>.71i</td>
<td></td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>3 (9.7)</td>
<td>5 (16.1)</td>
<td>3 (9.7)</td>
<td>11 (35.5)</td>
<td>.02e</td>
<td></td>
</tr>
<tr>
<td>P value</td>
<td>.99j</td>
<td>.23j</td>
<td>.65j</td>
<td>.02g</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fermented beverages (0-2 serving/day of wine and beer)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>36 (90.0)</td>
<td>36 (90.0)</td>
<td>37 (92.5)</td>
<td>37 (92.5)</td>
<td>.99i</td>
<td></td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>27 (87.1)</td>
<td>27 (87.1)</td>
<td>29 (93.5)</td>
<td>29 (93.5)</td>
<td>.67i</td>
<td></td>
</tr>
<tr>
<td>P value</td>
<td>.72j</td>
<td>.72j</td>
<td>.99j</td>
<td>.99j</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potatoes (≤3 servings/week)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>32 (80.0)</td>
<td>27 (67.5)</td>
<td>26 (65.0)</td>
<td>25 (62.5)</td>
<td>.08e</td>
<td></td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>23 (74.2)</td>
<td>19 (61.3)</td>
<td>17 (54.8)</td>
<td>19 (61.3)</td>
<td>.28e</td>
<td></td>
</tr>
<tr>
<td>P value</td>
<td>.56f</td>
<td>.59g</td>
<td>.39g</td>
<td>.92g</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Legumes (≥2 servings/week)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>15 (37.5)</td>
<td>13 (32.5)</td>
<td>17 (42.5)</td>
<td>14 (35.0)</td>
<td>.82e</td>
<td></td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>13 (41.9)</td>
<td>18 (58.1)</td>
<td>22 (71.0)</td>
<td>19 (61.3)</td>
<td>.13e</td>
<td></td>
</tr>
<tr>
<td>Food group</td>
<td>MDSS index consumption criteria and study group</td>
<td>Week 1</td>
<td>Week 2</td>
<td>Week 3</td>
<td>Week 4</td>
<td>$P$ value $^d$</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-------------------------------------------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>----------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.02$^e$</td>
<td>.02$^e$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Eggs (2-4 servings/week)</strong></td>
<td></td>
<td>.26$^e$</td>
<td>.22$^e$</td>
<td>.24$^e$</td>
<td>.99$^b$</td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>21 (52.5)</td>
<td>24 (60.0)</td>
<td>22 (55.0)</td>
<td>26 (65.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>19 (61.3)</td>
<td>20 (64.5)</td>
<td>17 (54.8)</td>
<td>17 (54.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Fish (≥2 servings/week)</strong></td>
<td></td>
<td></td>
<td></td>
<td>.99$^j$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>31 (77.5)</td>
<td>28 (70.0)</td>
<td>32 (80.0)</td>
<td>35 (87.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>28 (90.3)</td>
<td>27 (87.1)</td>
<td>26 (83.9)</td>
<td>27 (87.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>White meat (2-3 servings/week)</strong></td>
<td></td>
<td>.04$^e$</td>
<td>.14$^e$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>20 (50.0)</td>
<td>11 (27.5)</td>
<td>10 (25.0)</td>
<td>11 (27.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>10 (32.3)</td>
<td>9 (29.0)</td>
<td>12 (38.7)</td>
<td>5 (16.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Red meat (&lt;2 servings/week of pork, beef, lamb, and processed meat)</strong></td>
<td></td>
<td>.39$^e$</td>
<td>.54$^e$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>9 (22.5)</td>
<td>9 (22.5)</td>
<td>7 (17.5)</td>
<td>6 (15.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>6 (19.4)</td>
<td>10 (32.3)</td>
<td>7 (22.6)</td>
<td>8 (25.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sweets (≤2 servings/week)</strong></td>
<td></td>
<td></td>
<td></td>
<td>.26$^e$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>19 (47.5)</td>
<td>15 (37.5)</td>
<td>20 (50.0)</td>
<td>20 (50.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>15 (48.4)</td>
<td>15 (48.4)</td>
<td>14 (45.2)</td>
<td>17 (54.8)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$^a$MDSS: Mediterranean Diet Serving Score.  
$^b$CG: control group.  
$^c$IG: intervention group.  
$^d$P value in columns: intergroup differences (CG vs IG) in each of the 4 study weeks. $P<.05$ was considered significant.  
$^e$Evaluated with the chi-square test.  
$^f$P value in rows: intragroup differences (CG versus IG) in week 1 versus week 4. $P<.05$ was considered significant.  
$^g$Evaluated with the chi-square test.  
$^h$Not applicable.  
$^i$Evaluated with the Fisher exact test.

Regarding intergroup modifications, statistically significant differences (CG vs GI) were observed throughout the study period in 4 food groups: fruits (weeks 3 and 4), vegetables (weeks 2, 3, and 4), nuts (week 4), and legumes (weeks 2, 3, and 4). In these 4 food groups, at week 4, the number of participants meeting the recommendations was higher in the IG compared to the CG, with 22.8% for fruits, 35.2% for vegetables, 23.0% for nuts, and, finally, 26.3% for legumes (Table 4).

Regarding intragroup modifications, statistically significant differences (week 1 vs week 4) were observed in the CG for white meat and in the IG for vegetables and nuts.

**Usability Rating Questionnaire for the e-12HR App**

Of 71 participants, 45 (63%) returned the completed questionnaire: 25 (63%) of 40 from the CG and 20 (65%) of 31 from the IG. The responses of the 45 PCPs are shown in Table 5.
Table 5. Responses to the usability rating questionnaire for the e-12HR app (CG n=25, IG n=20).

<table>
<thead>
<tr>
<th>Questions and groups</th>
<th>Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Easy to complete (strongly agree + agree)</td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>25 (100)</td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>20 (100)</td>
</tr>
<tr>
<td>P value</td>
<td>—</td>
</tr>
<tr>
<td>2. Understandable questions (strongly agree + agree)</td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>24 (96)</td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>20 (100)</td>
</tr>
<tr>
<td>P value</td>
<td>.99</td>
</tr>
<tr>
<td>3. Understandable feedback only for the IG (strongly agree + agree)</td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>—</td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>18 (90)</td>
</tr>
<tr>
<td>P value</td>
<td>—</td>
</tr>
<tr>
<td>4. I would be willing to complete again (strongly agree + agree)</td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>20 (80)</td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>12 (60)</td>
</tr>
<tr>
<td>P value</td>
<td>0.141</td>
</tr>
<tr>
<td>5. Time to complete (≤3 minutes/day)</td>
<td></td>
</tr>
<tr>
<td>CG, n (%)</td>
<td>19 (76)</td>
</tr>
<tr>
<td>IG, n (%)</td>
<td>15 (75)</td>
</tr>
<tr>
<td>P value</td>
<td>.99</td>
</tr>
</tbody>
</table>

a e-12HR: Electronic 12-Hour Dietary Recall.

b CG: control group.

c IG: intervention group.

d Differences between subgroups. P < .05 was considered significant.

e Not applicable.

f Evaluated with the Fisher exact test.

No statistically significant differences were observed for any of the questions on the questionnaire (CG vs IG). All PCPs indicated that the e-12HR app was easy to complete, and most of them responded that the app contained questions that were understandable (CG n=24, 96%; IG n=20, 100%) and that feedback was understandable (only for the IG, n=18, 90%). Furthermore, a large percentage were willing to use the app again (CG n=20, 80%; IG n=12, 60%). Completion of the app could be considered to have taken 3 minutes or less (CG n=19, 76%; IG n=15, 75%).

Discussion

Principal Findings

In relation to the main objective, there were significant statistical differences between the 2 groups in this study. At week 4 (no significant differences in week 1, baseline), the values were higher in the IG compared to the CG by 25.6% for the MDSS index (Table 2); by 213.1 for the number of participants with a medium/high (≥12) MDSS index (Table 2); and by 33.8% for fruits, 352.0% for vegetables, 184.0% for nuts, and 75.1% for legumes for the number of participants meeting the recommendations for specific food groups (Table 4).

Regarding the secondary objectives, first, the answers to the questions of the usability rating questionnaire for the e-12HR app were satisfactory. According to the questionnaire, the daily use time of the app was about 3 minutes or less per day for most respondents (Table 5). When working with smartphone apps, usability is an important aspect to consider. According to health care professionals, there are 3 principal criteria for selecting a “nutrition and diet” app for clients/patients, which are [31] ease of use (satisfactory data were obtained in this study), free of charge, and validation (the e-12HR app is free to download and has been previously validated [32-36]). Second, at baseline (week 1), AMD for PCPs was medium (mean MDSS index 9.45, SD 2.32) and 66% of participants had a medium/high MDSS index (≥9).
Overview

To begin with, workplace interventions are an excellent strategy to promote a healthy diet, considering that health care professionals spend long hours in their professional activity and often have 1 or more meals during their working day. At the hospital level, interventions have been implemented to facilitate access to and choice of healthy foods during the working day, such as modifying the availability of foods served in the canteen, subsidizing the cost of fresh fruits and vegetables (which are often more expensive than less healthy alternatives) [37], or implementing traffic light labeling (green: healthy; yellow: less healthy; red: unhealthy) [37,38]. In a study by Thorndike et al. [38] (in a hospital in Boston, Massachusetts, USA), their intervention also included personalized automated messages using a platform that automatically generated 2 weekly emails with feedback on previous purchases in the hospital cafeteria and lifestyle advice. Significant statistical increases were observed in green-labeled food purchases and decreases in red-labeled food purchases among the IG compared to the CG in the hospital cafeteria throughout the study period.

Thorndike et al’s [38] intervention was based on information about food eaten only in the hospital cafeteria (without considering other food consumed outside the hospital), so its scope was limited. However, to date, no interventions to promote a healthy diet among PCPs have been implemented; for example, the workplace intervention strategies discussed before would be difficult to implement in Spain because health centers are widely distributed throughout the territory and do not usually have a cafeteria or restaurant. Considering these difficulties as a possible alternative strategy, this study was the first to assess the ability of a smartphone app to improve the dietary habits among health care professionals (specifically, PCPs). Several randomized controlled clinical trials have used an app to improve AMD in Spanish adults, such as patients of health care centers (the EVIDENT II app [39,40] and the SalBi Educa Nutrition app [41]) and patients with type 2 diabetes mellitus (the EVIDENT II app [42]), but not in health care professionals.

Comparison With Prior Work

As previously mentioned, the e-12HR app has also been evaluated among Spanish university students (health sciences and non–health sciences) [19,20]. In relation to the main objective, the results obtained by the e-12HR app among PCPs compared to university students were (1) similar for the MDSS index (the increase among PCPs was 25.6% for the MDSS index, as shown in Table 2, and among university students was 17.4% [19] or 25.7% [20]), (2) more positive for the number of participants with a medium/high MDSS index (the increase among PCPs was 213.1%, as shown in Table 2, and among university student was 61.9% [19] or 74.5% [20]), and (3) less positive regarding the number of participants meeting the recommendations for specific food groups (improvements in 4 food group among PCPs, as shown in Table 4, and in 7 food groups among Spanish university students).

Regarding the secondary objectives, first, similar results were obtained among Spanish university students for the answers to the questions of the usability rating questionnaire for the e-12HR app. Second, at baseline (week 1), the mean MDSS index of 9.45 (SD 2.32) and the number of participants with a medium/high MDSS index (66%) among PCPs (Table 1) were higher compared with the data from health sciences students [20] during the same follow-up period (mean MDSS index 7.59, SD 2.72; percentage of participants with a medium/high MDSS index=33.4%). Significant statistical differences were found (P<.05) for both the MDSS index, which was evaluated with the Mann-Whitney U test, and the number of participants with a medium/high MDSS index, which was evaluated with the chi-square test: PCPs showed an improvement of 24.5% for the MDSS index and 98.2% for the number of participants with a moderate/high MDSS index (PCPs vs health science university students). This comparison must be made with caution, since the sample of PCPs was made up of doctors and nurses and the sample of health sciences university students was made up of students from the faculties of medicine and pharmacy. In addition, in a previous study by Sentenach-Carbo et al [43] among Spanish PCPs, the number of participants with medium/high AMD was lower: 55% versus 66.2%. It should, however, be considered that the adherence index used by both studies was different: the MDSS index was used in this study, and the validated 14-point Mediterranean diet adhesion screener was used in the Prevention with Mediterranean Diet (PREDIMED) study.

Conclusion

At baseline, Spanish PCPs presented medium AMD (measured as the MDSS index and the number of participants with a
Throughout the study period, in the short term, the use of the e-12HR app (an easy-to-implement and low-cost intervention) showed moderate improvements in the MDSS index and remarkable improvements in the number of participants with a medium/high MDSS index; in addition, PCPs responded positively to questions about the usability of the app. These results support recommending the use of the e-12HR app as a tool to contribute to improving diet and preventing NCDs among PCPs, which, at the same time, could positively influence patient dietary behavior and prevent diet-related NCDs among the patients. From the point of view of health care organizations, the prevention of NCDs among PCPs could, in addition, lead to higher personal and job satisfaction and fewer sickness-related absences from work; for this reason, health organizations themselves should be more involved in the recommendations to use tools such as the one analyzed in this study among their own workers.

Acknowledgments
This work was supported in part by the SAMFyC Foundation (grant number TD180/21). The SAMFyC Foundation had no role in the design, analysis or writing of this article. The research team would like to thank the PCPs who participated in this study.

Authors' Contributions
LMB performed the conception and design of the study, developed the app, analyzed and interpreted data, and wrote the paper. PMR and MDGP were involved in data collection and interpretation of the data and contributed to drafting the manuscript. The authors/evaluators are the owners and developers of the e-12HR app.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Real images of the e-12HR app (nonfeedback and feedback versions). e-12HR: Electronic 12-Hour Dietary Recall.

Multimedia Appendix 2
Usability rating questionnaire for the e-12HR app. e-12HR: Electronic 12-Hour Dietary Recall.

Multimedia Appendix 3
Consort checklist.

References


Abbreviations

AMD: adherence to the Mediterranean diet
CG: control group
e-12HR: Electronic 12-Hour Dietary Recall
IG: intervention group
MDSS: Mediterranean Diet Serving Score
NCD: noncommunicable disease
PCP: primary care professional
Original Paper

Quality Assessment of Smartphone Medication Management Apps in France: Systematic Search

Mickael Toïgo¹, MD; Julie Marc¹, MSc, MD; Maurice Hayot², MD, PhD; Lionel Moulis³,⁴, MD; Francois Carbonnel¹,⁵, MD, PhD

¹Department of General Practice, Univ Montpellier, Montpellier, France
²PhyMedExp, Univ Montpellier, CNRS, INSERM, CHU Montpellier, Montpellier, France
³Clinical Research and Epidemiology Unit, Department of Public Health, Univ Montpellier, CHU Montpellier, Montpellier, France
⁴Pathogenesis and Control of Chronic and Emerging Infections, Univ Montpellier, INSERM, EFS, University of Antilles, Montpellier, France
⁵Desbrest Institute of Epidemiology and Public Health, Univ Montpellier, INSERM, Montpellier, France

Corresponding Author:
Francois Carbonnel, MD, PhD
Desbrest Institute of Epidemiology and Public Health
Univ Montpellier, INSERM
Campus Santé, IURC, 641 avenue du Doyen Gaston Giraud
Montpellier, 34090
France
Phone: 33 684014834
Email: Francois.carbonnel@umontpellier.fr

Abstract

Background: Adherence to medication is estimated to be around 50% for chronically ill patients in high-income countries. Improving the effectiveness of adherence interventions could have a far greater impact on population health than any improvement in specific medical treatments. Mobile health (mHealth) is one of the most effective solutions for helping patients improve their medication intake, notably through the use of mobile apps with reminder systems. With more than 327,000 apps available in the mHealth field, it is difficult for health care professionals and patients alike to choose which apps to recommend and use.

Objective: We aim to carry out a systematic search of medication management smartphone apps available in France that send reminders to patients and assess their quality using a validated scale.

Methods: Mobile apps were identified in October and November 2022 after a systematic keyword search on the 2 main app download platforms: App Store (Apple Inc) and Google Play Store. Inclusion criteria were free availability, date of last update, and availability in French. Next, 2 health care professionals independently evaluated the included apps using the French version of the Mobile App Rating Scale (MARS-F), an objective scoring system validated for assessing the overall quality of apps in the mHealth field. An intraclass correlation coefficient was calculated to determine interrater reliability.

Results: In total, 960 apps were identified and 49 were selected (25 from the App Store and 24 from the Google Play Store). Interrater reliability was excellent (intraclass correlation coefficient 0.92; 95% CI 0.87-0.95; P<.001). The average MARS-F score was 3.56 (SD 0.49) for apps on the App Store and 3.51 (SD 0.46) for those on the Google Play Store, with 10 apps scoring above 4 out of 5. Further, 2 apps were tested in at least one randomized controlled trial and showed positive results. The 2 apps with the highest ratings were Mediteo rappel de médicaments (Mediteo GmbH) and TOM rappel medicaments, pilule (Innovation6 GmbH), available on both platforms. Each app’s MARS-F score was weakly correlated with user ratings on the App Store and moderately correlated on the Google Play Store.

Conclusions: To our knowledge, this is the first study that used a validated scoring system to evaluate medication management apps that send medication reminders. The quality of the apps was heterogeneous, with only 2 having been studied in a randomized controlled trial with positive results. The evaluation of apps in real-life conditions by patients is necessary to determine their acceptability and effectiveness. Certification of apps is also essential to help health care professionals and patients identify validated apps.

(JMIR Mhealth Uhealth 2024;12:e54866) doi:10.2196/54866
KEYWORDS
medication adherence; mobile apps; telemedicine; reminder system; behavioral therapy; mHealth; mobile health; app; apps; applications; smartphone apps; medication; medications; adherence; search; searches; searching; systematic; App Store; Google Play; French; reminder; reminders; MARS; quality; Mobile App Rating Scale; mobile phone

Introduction
Therapeutic adherence is defined by the World Health Organization as “the extent to which the behaviors of a person required to take medication, follow a diet and/or change lifestyle correspond to the recommendations agreed with a healthcare professional” [1]. It is estimated to be around 50% for people with chronic diseases in high-income countries [1]. The result is an increased risk of complications, hospitalization, and mortality for these patients, as well as consequently higher health care costs [2-5]. Improving the effectiveness of adherence interventions could potentially have a far greater impact on population health than any improvement in specific medical treatments [1].

Mobile health (mHealth) is a promising strategy to optimize therapeutic adherence [6]. mHealth covers medical and public health practices based on mobile devices such as cell phones, patient monitoring systems, personal digital assistants, and other wireless devices [7]. It is dominated by the use of health and wellness smartphone apps, the number of which continues to grow yearly [8]. In 2021, over 327,000 mHealth apps were listed in digital stores [8].

In France, almost 90% of French people aged older than 12 years owned a smartphone in 2022 and 72% of them had downloaded apps [9]. For their part, French general practitioners seem to be in favor of mHealth apps or devices and ready to prescribe these in their practice, but at the same time stress the importance of validating the use of these apps through randomized clinical trials and certification by health care professionals [10].

Several studies have confirmed that the use of a medication management app improves adherence to therapy, notably by sending reminders, even in older patients with no experience of using new technologies [11-14]. However, no study evaluating the quality of medication management apps via a validated score has been published to date.

The main objective of this study was to identify medication management apps to improve patient adherence and assess their quality, using a validated scale. The secondary objective of this study was to identify high-quality medication reminder apps and to provide recommendations to any patient needing to take one or more medications, regardless of pathology.

Methods
Overview
This involved a systematic search of smartphone apps with content evaluation, carried out between October 1, 2022, and June 20, 2023. It was reported in line with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist for reporting a systematic review (items not pertinent to a systematic search of apps were considered not applicable) [15].

App Selection
Apps were searched from October 1 to November 30, 2022, on the App Store and Google Play Store. These 2 platforms were used by over 99% of mobile users to download apps in 2022 [16].

The keywords searched on the download platforms were “medication reminder” ("rappel de médicament"), “medication monitoring” ("suivi de médicament"), “treatment reminder” ("rappel de traitement"), “treatment monitoring” ("suivi de traitement"), and “pillbox” ("pilulier"). The selection of search terms was based on existing studies, which were adapted after several prior search attempts on digital download platforms [17]. We also included native health apps, automatically present on iOS and Android phones without the need to download them. Being preinstalled on smartphones, they cannot be identified by our search equations on app stores. However, they are available and easily accessible to users. Android currently has no native health app.

Inclusion criteria were as follows: (1) medication reminder app; (2) availability in the French language; (3) free (apps free for only a trial period were not selected); (4) updated within the last 2 years (since 2022, Google Play Store removes apps without updates for more than 2 years [18], and for the App Store, this period is 3 years for apps with few downloads in the last 12 months [19]); and (5) not specific to a single treatment (eg, the contraceptive pill) or to the treatment of a particular pathology. Inclusion criteria were based on the data available on the apps’ presentation page, that is, the title, description, images, and general information about the app. Apps were excluded after download in the event of (1) or (2) unavailability on download platforms at the time of evaluation by one of the evaluators.

The decision to assess exclusively free apps was grounded in the primary target audience of patients encountering diverse and, at times, multiple impediments to consistent medication adherence. The high cost of an app is one of the main reasons why it is not downloaded [20]. In addition, users frequently report hidden costs as the main reason for discontinuation [20]. So, considering the price as a dissuading factor for the adoption of health apps, proposing a paid app to this population appeared inadequate [21].

Evaluation via the French Version of the Mobile App Rating Scale App
The Mobile App Rating Scale (MARS) was used to evaluate the apps. This score was previously validated for the overall quality assessment of an app in the field of mHealth [22]. The French version of this score, MARS-F, has also been validated for use with French-language apps [23].
MARS-F is a 19-item questionnaire divided into 4 objective evaluation sections (A to D): engagement (5 items: entertainment, interest, personalization, interactivity, and adaptation to target group), functionality (4 items: app performance, ease of use, navigation, and app gesture design), aesthetics (3 items: layout, graphic design, and overall visual appeal), and information (7 items: accuracy of app description, precise app objectives, quality of information, quantity of information, visual information, credibility of information, and scientific evidence; Multimedia Appendix 1). Each item is rated with a Likert scale from 1 ("inadequate") to 5 ("excellent"). When specific objectives (item 14) were not delineated, and pertinent information (items 15, 16, and 17) or scientific evidence (item 19) was lacking, the respective items were rated as "not applicable." Consequently, these items were excluded from consideration in the overall scoring process. In total, MARS-F is a score out of 5, corresponding to the average of these 4 sections. Section E, not included in the overall MARS-F calculation, evaluates the apps’ subjective quality via 4 items and is described separately. The MARS-F also includes a preliminary descriptive section on the apps’ characteristics, including users’ ratings on the platforms, the apps’ target objectives, the strategies used, its affiliation (commercial, governmental, and academic), as well as the technical aspect (password protection, possible sharing, and internet access required).

In total, 2 family medicine residents evaluated each app independently. Before starting, they assisted with a training video on the use of MARS-F and trained in real-life conditions by evaluating 3 apps not selected for this study. The scores for each item were then discussed one by one to ensure a similar understanding on the part of the 2 evaluators.

Apps available on the 2 platforms were evaluated independently on iOS and Android. Each app was tested and used for at least 15 minutes. Evaluations took place from February to June 2023.

Statistical Analysis
App characteristics were described using frequencies and proportions. To determine interrater reliability, a 2-way mixed-effects model intraclass correlation coefficient (ICC) was calculated for the mean of the raters. An ICC below 0.5 was considered poor, moderate between 0.5 and 0.75, good between 0.75 and 0.90, and excellent if above 0.90 [24]. The average of the scores given by the 2 raters or evaluators was used for the final rating of each app. Apps were compared according to their respective MARS-F quality score. The results were presented as mean (SD) and median and quartiles.

The correlation between the average rating of app users on download platforms and the MARS-F obtained was measured by Spearman correlation after a normality test. Regarding the correlation between MARS-F and user ratings, we excluded apps with a limited number of ratings on platforms, as they may not be highly representative of users. The choice of the threshold was determined following a sensitivity analysis, wherein the correlation was computed for various thresholds: five ratings on platforms, 10, 20, 30, 50, and so on. The selected threshold was the lowest one for which both the lower and upper thresholds yielded similar results. The threshold of 20 user ratings was finally selected. The correlation between reviewers’ subjective assessment of the apps via item 23 ("What is your overall star rating for the application?") and the MARS-F obtained for each app was also calculated by Pearson correlation after a normality test. The correlation was judged as very strong from 1 to 0.9, strong from 0.9 to 0.7, moderate from 0.7 to 0.5, weak from 0.5 to 0.3, and very weak from 0.3 to 0.

All statistical analyses were performed with EasyMedStat (version 3.29; EasyMedStat).

Ethical Considerations
The Research Ethics Committee of the University of Montpellier approved this research project (UM 2022-006bis; Multimedia Appendix 2). This study was not funded.

Results
App Selection
After a keyword search and the addition of native apps, 480 apps were identified on the App Store and 1191 on the Google Play Store. A total of 25 apps meeting the inclusion and exclusion criteria were selected from the App Store and 24 from the Google Play Store (Figure 1).
Description of the Apps

In total, 51% (n=25) of the apps evaluated were available on the App Store and 49% (n=24) on the Google Play Store, 10 of which were common and available on both platforms. Further, 40 had been rated by users on both platforms (23 on the App Store and 17 on the Google Play Store).

Although all the apps were available free of charge, some had a paid version, as in the case of 15 (60%) App Store apps and 12 (50%) Google Play Store apps. The affiliation of the developers was not always known, but the majority were commercial (24/25, 96% on the App Store; 21/24, 87.5% on the Google Play Store).

Technical aspects according to the download platform are summarized in Table 1.
Table 1. Technical aspects of apps according to platform.

<table>
<thead>
<tr>
<th>Feature</th>
<th>App Store (n=25), n (%)</th>
<th>Google Play (n=24), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allows sharing</td>
<td>12 (48)</td>
<td>10 (42)</td>
</tr>
<tr>
<td>Has an app-linked community</td>
<td>3 (12)</td>
<td>1 (4)</td>
</tr>
<tr>
<td>Has the ability to password-protect</td>
<td>10 (40)</td>
<td>3 (12.5)</td>
</tr>
<tr>
<td>Requires a login ID</td>
<td>4 (16)</td>
<td>2 (8)</td>
</tr>
<tr>
<td>Sends reminders</td>
<td>25 (100)</td>
<td>24 (100)</td>
</tr>
<tr>
<td>Needs internet access to work</td>
<td>10 (40)</td>
<td>8 (33)</td>
</tr>
</tbody>
</table>

Evaluation Reliability

MARS-F interrater reliability for all apps was excellent (ICC 0.92; 95% CI 0.87-0.95; P<.001).

For each section, interrater reliability was excellent for engagement (ICC 0.92; 95% CI 0.88-0.95; P<.001) and for subjective app quality (ICC 0.95; 95% CI 0.92-0.97; P<.001). It was good for aesthetics (ICC 0.81; 95% CI 0.70-0.89; P<.001) and the information section (ICC 0.86; 95% CI 0.78-0.92; P<.001). Finally, interrater reliability was only moderate for the functionality part (ICC 0.70; 95% CI 0.52-0.82; P<.001).

Quality of Apps Based on MARS-F

The mean MARS-F score was 3.56 (SD 0.49) for iOS apps and 3.51 (SD 0.46) for Android apps (Table 2). The full set of scores by the app is presented in Multimedia Appendix 3.

Table 2. Average score, range, and median per section.

<table>
<thead>
<tr>
<th>Variable</th>
<th>iOS (n=25)</th>
<th>Android (n=24)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score, median (Q1-Q3)</td>
<td>3.3 (2.8-3.6)</td>
<td>3.15 (0.628)</td>
</tr>
<tr>
<td>Score, mean (SD)</td>
<td>3.19 (0.616)</td>
<td>3.15 (0.616)</td>
</tr>
<tr>
<td>Section A: engagement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score, median (Q1-Q3)</td>
<td>3.3 (2.8-3.6)</td>
<td>3.15 (0.628)</td>
</tr>
<tr>
<td>Score, mean (SD)</td>
<td>3.29 (0.708)</td>
<td>3.15 (0.628)</td>
</tr>
<tr>
<td>Section B: functionality</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score, median (Q1-Q3)</td>
<td>4.38 (4-4.62)</td>
<td>4.17 (0.431)</td>
</tr>
<tr>
<td>Score, mean (SD)</td>
<td>4.26 (0.448)</td>
<td>4.17 (0.431)</td>
</tr>
<tr>
<td>Section C: aesthetics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score, median (Q1-Q3)</td>
<td>3.6 (3.21-3.98)</td>
<td>3.51 (0.455)</td>
</tr>
<tr>
<td>Score, mean (SD)</td>
<td>3.56 (0.485)</td>
<td>3.51 (0.455)</td>
</tr>
<tr>
<td>Section D: information</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score, median (Q1-Q3)</td>
<td>3.3 (2.83-3.8)</td>
<td>3.33 (0.561)</td>
</tr>
<tr>
<td>Score, mean (SD)</td>
<td>3.29 (0.708)</td>
<td>3.33 (0.561)</td>
</tr>
<tr>
<td>MARS-F: global quality</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score, median (Q1-Q3)</td>
<td>2.38 (1.5-3.62)</td>
<td>2.55 (1.19)</td>
</tr>
<tr>
<td>Score, mean (SD)</td>
<td>2.52 (1.12)</td>
<td>2.55 (1.19)</td>
</tr>
<tr>
<td>Section E: subjective quality</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Quality of Apps Based on MARS-F

The mean MARS-F score was 3.56 (SD 0.49) for iOS apps and 3.51 (SD 0.46) for Android apps (Table 2). The full set of scores by the app is presented in Multimedia Appendix 3.

General App Features

The iOS apps with the highest MARS-F scores were TOM rappel medicament, Pilule (MARS-F score: 4.37; Innovation6 GmbH), then Mediteo rappels de médicaments from (MARS-F score: 4.34; Mediteo GmbH), followed by Rappels de médicaments from (MARS-F at 4.13; smartpatient GmbH). The Android apps with the highest MARS-F scores were Mediteo rappels de médicaments (MARS-F score: 4.32), followed by TOM Rappel medicaments, pilule (MARS-F score: 4.24) and Rappel de pilule et medicament (MARS-F score: 4.09; Medisafe). The MARS-F ranking of apps by platform is available in Multimedia Appendix 4.

In total, 84% (n=21) of the apps available on the App Store had a MARS-F score above 3, with 5 (20%) scoring above 4. On Google Play Store, 79% (n=19) of apps had a MARS-F score higher than 3, including 5 (21%) with a score higher than 4.

The “scientific evidence” item was completed for 4 (16%) apps on the App Store and 2 (8%) apps on the Google Play Store, 2 of which were common to both platforms.

Correlation Between MARS-F and User Ratings

In total, 23 (92%) apps were rated by users on the App Store, with an average rating of 4.38/5 (range 2.9-5.0) and an average number of ratings of 475 (range 1-3900; Multimedia Appendix 5). On Google Play Store, 17 (71%) apps were rated by users, with an average score of 4.17/5 (range 3.3-4.8) and an average number of ratings of 25,062 (range 24-223,000).

After performing sensitivity analyses, tests looking for a correlation between a given app’s MARS-F and average user rating were carried out on apps with at least 20 reviews on the platforms. The correlation was weak on the App Store (r=0.46; P=.12) and moderate on the Google Play Store (r=0.55; P=.02).

Correlation Between MARS-F and Reviewers’ Subjective Evaluation

The correlation between MARS-F and item 23 (“What is your overall star rating of the app?”) was considered very strong for iOS apps (r=0.93; P<.001) and strong for Android apps (r=0.88; P<.001).

The mean score for this item was 2.90 and 2.98 on iOS and Android, respectively, which is below the respective mean MARS-F.
Discussion

Principal Results

The main objective of our study was to identify medication management apps to improve patient adherence. In total, 960 apps were identified: 358 (37.3%) on App Store and 602 (62.7%) on Google Play Store. This imbalance in favor of the Google Play Store has already been observed in several studies [25,26]. However, after selection, the number of apps was balanced 51% (n=25) on the App Store and 49% (n=24) on the Google Play Store, so it seems that the Google Play Store search engine offers more apps for the same keyword. This balanced proportion is consistent with a 2019 medication management app evaluation study in which 328 apps available in English were studied [17]. In this study, 53.4% and 46.6% of apps were retrieved from the Google Play Store and the App Store, respectively [17]. Our study evaluated the content of these apps according to the information available on app download stores without downloading them.

Our study highlighted a cybersecurity issue: only 40% (n=10) of apps identified on iOS had password-protected access, and 4% (n=1) required a login ID. These figures were even lower for Android apps, at 12.5% (n=3) and 8.3% (n=2), respectively. This was also the case for pain management apps, with 39% of apps evaluated allowing password protection and 44% requiring login [27].

The overall quality of the medication reminder apps evaluated is considered acceptable to moderate, with 84% (n=21) of apps on the App Store achieving a MARS-F higher than 3/5 with an average of 3.56 (SD 0.49) and 79% (n=19) on Google Play Store with an average MARS-F of 3.51 (SD 0.46). We have no point of comparison in the absence of any previous French-speaking or English-speaking study evaluating medication reminder apps by a validated score. Nevertheless, these results were expected, as they are consistent with other French studies on health apps for smoking cessation (mean MARS-F of 3.49, SD 0.57) for 14 apps [28], nutrition (mean MARS-F of 3.26, SD 0.43) for 15 apps [29], mental health (mean MARS-F of 3.16, SD 0.43) for 12 apps [30], or oral hygiene (mean MARS-F of 3.20, SD 0.38) for 9 apps [31].

All the other studies reported that the functionality section of the MARS scale had the highest ratings, which was also the case in our study, with an average of 4.26 (SD 0.45) for the App Store and 4.17 (SD 0.43) for the Google Play Store for this section. This shows that most apps are functional, which also coincides with the fact that only 1 app was excluded for malfunction in our study. In contrast, the section with the lowest average score was the engagement section (average of 3.19, SD 0.62, on the App Store and 3.15, SD 0.63, on the Google Play Store), followed by the information section (average of 3.29, SD 0.71, and 3.33, SD 0.56, on the App Store and Google Play Store, respectively), in line with the literature [32-37]. Several other studies have noted lower scores for the information section [25,27,30,31,38-41]. Nevertheless, these 2 sections are regularly cited as being those for which we find lower averages. Further, 1 exception is the study by Salehinejad et al [42], for which the highest-rated section was the information section. This is a special case, as the evaluation was concerned with COVID-19 management apps, which were probably created for information purposes in the first place.

About the information section, the lack of involvement from health care professionals, universities, or governmental organizations, may explain the observed low scores, thereby lowering the credibility of the apps. This point is assessed in the MARS-F by item 18 (“Does the application come from a legitimate source (specified in the application store description or in the application itself?)”). Often difficult to ascertain [16], most apps were affiliated with a commercial company, with only 14.6% of medication management apps and 15.2% of apps for patients with genitourinary tumors showing involvement by health care professionals [43]. Involvement by health care professionals was sometimes mentioned on the apps’ websites, but without explaining their precise role or degree of involvement. The “scientific evidence” item (item 19) was completed for 4 different apps out of the 49 evaluated in our study, 2 of which were common to both the App Store and Google Play Store. These 2 apps are the only ones to have been tested in at least one randomized controlled trial with positive results: smartpatient GmbH or MyTherapy’s Rappels de médicaments [44] and Medisafe’s Rappel de pilule et medicament [45]. They are ranked among the best apps on the 2 platforms according to their MARS-F obtained in this study. This lack of scientific validation of health care apps is a recurring theme [25,29,32,33,37].

In our study, user ratings of digital app stores were higher than MARS-F, as expected [27,29]. The absence or low correlation between user rating and MARS score has also been described previously [33,46]. To our knowledge, only the study by Chen et al [47] found a link between the quality of apps assessed by the MARS score and user rating for drug information apps. The average MARS score of apps with higher user star ratings was significantly higher than for apps with lower user star ratings (3.38, SD 0.64, vs 3.05, SD 0.64, P<0.001) [47]. However, this may be explained by their study design: 3 out of 7 assessors were not health care professionals, which may be a bias.

The correlation between item 23 (“What is your overall star rating for the app?”) and the MARS-F was very strong for apps on iOS and strong for apps on Android. These results are consistent with other studies dealing with nutrition and mental health apps, respectively [29,30]. It is interesting to note that the subjective rating of the reviewers correlated with the overall quality of the app assessed by an objective scale, unlike the user rating. It is challenging to predict whether this is due to the evaluators’ experience or the fact that they have delved deeper into the evaluation of apps using the MARS score.

Strengths and Limitations

The main limitation of this study is the mobile app sector itself since this study is a reflection of the supply and quality of apps for a specific period. We note, for example, that 6 apps (3 on the App Store and 3 on the Google Play Store) were excluded because they were no longer available a few months after they were identified and downloaded for evaluation. This difficulty had already been highlighted in a study that aimed to evaluate apps for pregnant women at 2-year intervals, in 2017 and then
in 2019. One of the things that stood out was that the best app in 2017 was no longer available in 2019 at the time of the second evaluation [40]. The presence of new versions following updates is another element that can influence the quality of an app over time. App selection by only one of the evaluators is also a limitation, as it is possible that some eligible apps have not been identified. In terms of evaluation, the apps were assessed after they had been in use for a minimum of 15 minutes, so additional strengths or weaknesses of each app may not have been detected due to this limited duration.

Our choice to evaluate only free apps is also open to criticism. It could be argued that this choice is induced by a bias of the health care system in which the authors of this paper, who are French, operate. France has indeed the lowest share of out-of-pocket health expenditure among all EU countries [48]. Yet, medication nonadherence is a global problem. Costs attributed to “all causes” nonadherence range from US $5271 to US $52,341 [49]. The high cost of a paid app is dissuasive and hidden costs have been demonstrated as one of the main reasons for discontinuation of the use of an app [20,21]. To our knowledge, the superiority of a paid medication management app over a free one has yet to be demonstrated.

Finally, the MARS score was used because its use is simple, validated, and widespread in the evaluation of health care apps. However, it has several limitations. The first is the absence of data security and privacy evaluation criteria, although the presence of a password and login is indicated in the descriptive section. These points are nevertheless a concern for users of health care apps [50]. The second limitation is the absence of a threshold for judging app quality. We have described it here as moderate, as in several other studies, which found similar average scores, but this was not described when this tool was created [51].

To our knowledge, no other study has evaluated medication management apps using a validated scale. The identification method combined with the inclusion and exclusion criteria enabled an exhaustive analysis of free medication reminder apps available to French patients. These were not selected based on user ratings or the number of downloads from digital stores, which made it possible to evaluate apps that are not promoted on platforms but may nevertheless be relevant to patients. The weak correlation between app quality and user rating found in this study supports this approach. Apps available for iOS and Android devices were evaluated independently on each operating system, as there may be differences in terms of updating or functionality depending on the device used. Finally, the independent testing by 2 evaluators is a strong point, particularly with the observed excellent interevaluator reliability, enabling result confirmation.

**Perspectives**

This work is the first step toward facilitating more in-depth studies on top-rated and best-quality apps. App evaluation by patients with the user version of the MARS [52] would be relevant, even though there is currently no validated French version of this scale. The long-term use and benefits of these apps need to be studied in randomized clinical trials, to verify their acceptability and whether or not they improve therapeutic adherence and clinical outcomes in patients undergoing long-term treatment.

The results of this study, added to other studies on app evaluation in the health care field, show the possible ways to improve existing apps and give leads for the creation of new ones. Functionality is paramount, and this point already seems to have been achieved for the majority of apps currently available. The areas that need improvement relate above all to engagement and the information available, of which gamification is 1 avenue to be explored [53]. Data security and privacy protection are also important for patients and should not be neglected.

Currently, France is investing in the use of digital health, particularly for patients with chronic diseases. There is an ongoing project to list more than 50 apps offering exchanges with *Mon Espace Santé* by 2026. Therefore, it is vital to set up certification for existing apps to help doctors in their recommendations [10]. The creation of an app with the help of health care professionals and validated by the Agence du Numérique en Santé would also be a solution that would enable doctors to know which apps to recommend, thus facilitating its use by patients.

**Conflicts of Interest**

None declared.

**Multimedia Appendix 1**

Mobile App Rating Scale, French version (MARS-F).

[PDF File (Adobe PDF File), 190 KB - mhealth_v12i1e54866_app1.pdf ]

**Multimedia Appendix 2**

Ethics committee approval.

[PDF File (Adobe PDF File), 280 KB - mhealth_v12i1e54866_app2.pdf ]

**Multimedia Appendix 3**

The full set of scores by app.

[DOCX File (.40 KB - mhealth_v12i1e54866_app3.docx ]
References


8. Évaluation des Applications dans le champ de la santé mobile (mHealth)—État des lieux et critères de qualité du contenu médical pour le réferencement des services numériques dans l'espace numérique de santé et le bouquet de services des professionnels. La Plaine, Saint-Denis: Haute Autorité de Santé; 2021.


27. Salazar A, de Sola H, Failde I, Moral-Munoz JA. Measuring the quality of mobile apps for the management of pain: systematic search and evaluation using the mobile app rating scale. JMIR mHealth uHealth 2018;6(10):e10718 [FREE Full text] [doi: 10.2196/10718] [Medline: 30361196]


32. Gutiérrez NB, Ramallo HR, González MF, Martín LAK. Smartphone apps for patients with hematologic malignancies: systematic review and evaluation of content. JMIR mHealth uHealth 2022;10(9):e35851 [FREE Full text] [doi: 10.2196/mhealth.35851] [Medline: 36125860]


46. Wu X, Xu L, Li PF, Tang TT, Huang C. Multipurpose mobile apps for mental health in Chinese App Stores: content analysis and quality evaluation. JMIR mHealth uHealth 2022;10(1):e34054 [FREE Full text] [doi: 10.2196/34054] [Medline: 34982717]


49. Mohammadzadeh N, Khenarinezhad S, Ghazanfarisavadkoohi E, Safari MS, Pahlevanynejad S. Evaluation of M-Health applications use for patients: systematic literature review. JMIR Serious Games 2020;8(2):e16096 [FREE Full text] [Medline: 32347811]


**Abbreviations**

- **ICC**: intraclass correlation coefficient
- **MARS**: Mobile App Rating Scale
- **MARS-F**: Mobile App Rating Scale, French version
- **mHealth**: mobile health
- **PRISMA**: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
The Association of eHealth Literacy Skills and mHealth Application Use Among US Adults With Obesity: Analysis of Health Information National Trends Survey Data

George Shaw Jr¹, PhD; Bianca A Castro², MSc; Laura H Gunn¹, PhD; Keith Norris³, MD, PhD; Roland J Thorpe Jr⁴, PhD

¹Department of Public Health Sciences, School of Data Science, University of North Carolina at Charlotte, Charlotte, NC, United States
²Department of Public Health Sciences, University of North Carolina at Charlotte, Charlotte, NC, United States
³The University of California Los Angeles Division of General Internal Medicine and Health Services Research, University of California, Los Angeles, Los Angeles, CA, United States
⁴Department of Health, Behavior and Society, Johns Hopkins University, Baltimore, MD, United States

Corresponding Author:
George Shaw Jr, PhD
Department of Public Health Sciences
School of Data Science
University of North Carolina at Charlotte
9201 University City Blvd.
Charlotte, NC, 28223-0001
United States
Phone: 1 7045620021
Email: gshaw11@charlotte.edu

Abstract

Background: Physical inactivity and a poor diet are modifiable behaviors that contribute to obesity. Obesity is a well-recognized risk factor for chronic diseases, including diabetes. Mobile health (mHealth) apps can play an important adjuvant role in preventing and treating chronic diseases and promoting positive health behavior change among people with obesity, and eHealth literacy skills have the potential to impact mHealth app use.

Objective: The purpose of this study was to explore the associations between the 2 dimensions, access and application, of eHealth literacy skills and mHealth app use among US adults (≥18 years of age) with obesity (BMI ≥30 kg/m²).

Methods: Data were obtained from February to June 2020 using the Health Information National Trends Survey 5. A total of 1079 respondents met the inclusion criteria of adults with obesity and owners of smartphones. Individual associations between mHealth app use and sociodemographic variables were explored using weighted chi-square and 2-tailed t tests. A multivariable weighted logistic regression model was fitted, and adjusted odds ratios (ORs) of using mHealth apps with corresponding 95% CIs were reported across multiple sociodemographic variables. An Ising model-weighted network visualization was produced. A receiver operating characteristic curve was calculated, and the area under the curve was reported with the corresponding Delong 95% CI.

Results: A majority of respondents were female (550/923, 59.6%) or non-Hispanic White (543/923, 58.8%). Individuals in households earning less than US $50,000 comprised 41.4% (382/923) of the sample. All sociodemographic variables were found to be univariately significant at the 5% level, except employment and region. Results from the multivariable weighted logistic regression model showed that the adjusted odds of using an mHealth app are 3.13 (95% CI 1.69-5.80) and 2.99 (95% CI 1.67-5.37) times higher among those with an access eHealth literacy skill of using an electronic device to look for health or medical information for themselves and an application eHealth literacy skill of using electronic communications with a doctor or doctor’s office, respectively. Several sociodemographic variables were found to be significant, such as education, where adjusted ORs comparing subgroups to the lowest educational attainment were substantial (ORs ≥7.77). The network visualization demonstrated that all eHealth literacy skills and the mHealth app use variable were positively associated to varying degrees.

Conclusions: This work provides an initial understanding of mHealth app use and eHealth literacy skills among people with obesity, identifying people with obesity subpopulations who are at risk of a digital health divide. Future studies should identify equitable solutions for people with obesity (as well as other groups) and their use of mHealth apps.
Introduction

Overview
Physical inactivity and poor dietary behaviors are modifiable behaviors that contribute to obesity [1,2]. Recent studies show that obesity affects nearly 42% of the US population aged 20 years or older [3], with an associated excess annual estimated medical cost of upwards of US $170 billion in 2019 [4]. Obesity is a well-recognized risk factor for chronic diseases, including diabetes, cardiovascular diseases, and cancer, and a significant cause of premature morbidity and mortality [2,5]. Wang et al [6] demonstrated the importance of reducing the weight of patients classified as obese, which is a major contributor to the increased incidence of type 2 diabetes mellitus. Obesity is a complex and multifaceted disease, extending far beyond the realm of individual behavior or mere lifestyle choices. Genetic, environmental, and socioeconomic factors in the development of obesity [7]. However, mobile health (mHealth) apps can play an important adjunct role in preventing and treating chronic diseases and promoting positive health behavior change among individuals with obesity [8,9], with several factors influencing users’ acceptance of and engagement with mHealth apps [10-12]. Previous intervention studies have used mHealth apps to promote health behavior change among individuals with obesity [13,14]. Participants from the reference studies found that the use of mHealth made the intervention helpful and benefited weight loss when used in conjunction with other weight loss intervention methods. mHealth apps provide access to health information and can extend this access to underserved groups, particularly those at higher risk of chronic diseases [15]. However, many individuals with chronic diseases like obesity fail to engage in mHealth app use [16].

mHealth and eHealth Literacy
Istepanian et al [17] defined mHealth as mobile computing, medical sensors, or communication technologies designed for health care. Recent interventions involving mHealth apps provide evidence of improvement in participants’ self-care and disease self-management [18,19]. It is estimated that more than 350,000 mHealth apps are available on the market [20]. Despite the number of digital health interventions targeting weight management, the level to which users actively and regularly engage with those apps entails user engagement early in the intervention design process [21]. While health literacy is a broader concept associated with the ability of individuals to obtain and understand health information to make rational health decisions, eHealth literacy comprises the complex navigation of health care information from internet sources [22,23]. Kontos et al [24] showed that people with lower levels of education were less likely to use the internet to communicate with a doctor or use health information on their mobile devices. Moreover, national and international studies have shown that weight-management apps are beneficial for improving weight loss [25,26]. Previous studies have recommended focusing on understanding the level of health literacy of recipients who may use these apps, particularly those interested in weight loss interventions. Understanding eHealth literacy skills is critical when evaluating health information from mHealth apps and the subsequent application of the knowledge gained [27,28].

eHealth literacy is defined as seeking, finding, understanding, and appraising health information from electronic sources and applying the knowledge acquired to address or solve a health-related problem [29,30]. More recently, eHealth literacy has been conceptualized in dimensions including access to digital services and the application of services and information that satisfy users [31]. People lacking or with a low level of health literacy or eHealth literacy skills benefit less from digital health information and health informatics interventions [32,33], as low eHealth literacy skills were found to impact mHealth app use when mediated through mHealth app efficacy [23]. By contrast, those with higher levels of health literacy and eHealth literacy skills report a positive connection between mHealth app use and health outcomes [34].

mHealth apps are operated across a diverse group of users. This includes individuals with varying incomes, ages, races, ethnicities, and educations [35]. Curating data that include this and other defining personal characteristics require a significant number of resources. Few studies of eHealth literacy and mHealth app use have included a diverse group of participants. Through a national data set of noninstitutionalized adults, this study uses a diverse national data set. The purpose of this study is to explore the potential associations between the 2 dimensions, access and application, of eHealth literacy skills and mHealth app use among a diverse group of US adults (aged ≥18 years) with obesity (BMI ≥30 kg/m²).

Theoretical Underpinning
Seeking health information has become a typical behavior among people of all ages and health conditions [36]. The information obtained when factoring in a person’s eHealth literacy skills, has the potential to influence health outcomes. This is particularly important for people with obesity. People with obesity have lower self-confidence in managing their health. However, people with obesity view communication with their physician as helpful with self-care weight management [37]. Therefore, people with obesity may improve their confidence over time in managing their weight if they have the eHealth literacy skills to seek information from their physician electronically. There have been numerous information behavior theories and models developed to understand how individuals seek and use information [38]. For example, Zare-Farashbandi and Lalazaryan [39] designed their health information acquisition model based on 6 stages of information seeking. The model acknowledges that the information-seeking process can be iterative and that there is a need for a feedback loop in the search process. However, the model does not consider personal or contextual factors affecting information seeking.
Longo’s model of health information considers the effects of personal and contextual factors on the information-seeking behavior of patients [40]. Focusing on patients with chronic diseases, the model was significant in depicting the output process of information-seeking for patients [39]. These theories and models have also considered various social contexts and population groups, such as older individuals, patients with cancer, prisoners, and diabetics [38,41-43]. However, this study is informed by Lenz’s [41] Information Seeking Model, which is the foundation of many fundamental tenets of recent models and instruments aimed at measuring eHealth literacy skills.

According to this framework, in which information gathering is part of the decision-making process, individuals follow 6 stages to seek health information. First, they receive a stimulus from their previous disease experience or the environment. Second, they establish their informational goals, including sources, time available, and the type of information needed. Third, the person decides whether not or not to actively access the information they want. The decision is based on the individual’s previous knowledge, background, and the expected cost-benefit of the action. The fourth stage is of particular interest, characterized by the information-seeking action itself. This step is correlated to the eHealth literacy dimension of access, and it could be an in-depth search or superficial information gathering, depending on the person’s need and previous attempts. The access dimension corresponds to having the availability of digital services that suit people’s needs and work correctly [44]. The fifth stage corresponds to information achievement and interpretation. This step is related to the eHealth literacy dimension of application, in which the individual understands and appraises or applies the information obtained. In this final stage, people may have to decide on the adequacy of the acquired information [39]. Understanding health information seeking through the potential associations of eHealth literacy skills and mHealth app use may provide insights into how people with chronic conditions with chronic conditions such as obesity can access and apply the information they seek [45]. Using this framework, the objective of this study is to explore the potential associations between the 2 dimensions, access and application, of eHealth literacy skills and mHealth app use among a diverse group of US adults with obesity.

Methods

Data

The Health Information National Trends Survey (HINTS) was used to explore the potential association between mHealth app use and eHealth literacy skills. HINTS has been administered every few years by the National Cancer Institute since 2003, and the data sets that have been made publicly available are used for evaluating health information access and use among US adults [24,46,47]. HINTS collects representative data about noninstitutionalized US adults’ knowledge, access, attitudes, and use of cancer- and health-related information. The survey uses a 2-stage stratified random sampling that selects households from residential addresses in the United States and then selects 1 adult within each household [24,46]. This study used the HINTS 5, Cycle 4 data set. The data were collected from February to June 2020 and comprised responses from 3865 participants. Despite the COVID-19 impact on society, the response rate for the survey remained high. The response rate (37%) for the survey remained relatively high and was even higher than prepandemic HINTS 5 surveys, which experienced response rates of at most 33% [48]. However, COVID-19 impacted the time frame in which the data are typically collected. Individuals included in this analysis were those who indicated ownership of a tablet, smartphone, or both and self-declared a BMI ≥30 kg/m² (obese). The dependent variable was based on respondents’ answers to the following item: “In the past 12 months, have you used any of these health or wellness apps?” The binary variable derived was used to indicate those who reported using any health or wellness apps within the past 12 months and those who did not.

The main independent variables representing eHealth literacy skill’s access and application dimensions were the following four items pertaining to eHealth information and services, connecting to the common stem of “In the past 12 months, have you used a computer, smartphone, or other electronic means to do any of the following: (1) looked for health or medical information for yourself; (2) used email or the internet to communicate with a doctor or doctor’s office; (3) looked medical test results; and (4) made appointments with a health care provider?” Access within this context is the interpretation and appraisal of information aimed at completing an action. Additional covariates extracted include age in years, health insurance status, sex at birth, employment status, marital status, education, annual household income (in ranges), race and ethnicity, and US Census region. These variables have been used in previous studies to evaluate mHealth app use or can be relevant confounders regarding the associations between eHealth literacy skills and mHealth app use [49,50].

Due to low counts, the following categories were combined: employment status of unemployed across lengths of unemployment; employment status of students and others; marital status of separated and divorced; marital status of married and those living as married or with a romantic partner; education categories below 11 years of education; and race and ethnicity categories of non-Hispanic Native Hawaiian or Other Pacific Islander and American Indian or Alaska Native.

Statistical Analysis

In order to investigate the research objective and hypothesis, a comprehensive statistical analysis was performed on the collected data using univariate and multiple logistic regression modeling. The weights provided by HINTS were used to perform all analyses and adjust for sampling biases [51]. A weighted complete case analysis was performed on the data. Characteristics were summarized using means, SDs, counts, and percentages as appropriate. Weighted chi-square and 2-tailed t tests were used to explore univariate associations between each of the covariates and mHealth app use, with test statistics and corresponding P values tabulated. Visualizations were created to explore associations, including (1) Ising model

https://mhealth.jmir.org/2024/1/e46656

Shaw Jr et al JMIR MHEALTH AND UHEALTH 2024 | vol. 12 | e46656 | p.721 https://mhealth.jmir.org/2024/1/e46656 (page number not for citation purposes)
The primary study aim is to assess associations between mHealth app use (binary outcome) and each of the eHealth literacy skills dimensions of access and application (main covariates). Univariate analysis is included to provide a comprehensive description of the individual variables in the study and establish a foundation for more complex multivariable analyses. These were further examined using a multivariable weighted logistic regression adjusted for the aforementioned sociodemographic factors. Adjusted odds ratios (ORs), corresponding 95% CIs, and P values were reported across eHealth literacy skills dimensions and sociodemographic variables. Results were tabulated and highlighted using a significance level of 5%. A pseudo-\( R^2 \) was calculated. The receiver operating characteristic (ROC) is a common approach used to measure the sensitivity versus specificity of logistic models. Additionally, the area under the curve (AUC) is a single metric for that trade-off, with AUC=1 meaning that the model perfectly fits the data and AUC=0.5 indicating there is a split chance that the model fits the data. Both of these approaches are used to evaluate the performance of logistic models. The ROC curve was estimated, and the corresponding AUC value and Delong 95% CI were reported. R software (version 4.0.3; R Foundation for Statistical Computing) was used for statistical analyses.

**Ethical Considerations**

This research was approved by the institutional review board (IRB) of the University of North Carolina at Charlotte (study #IRB-22-0585). This data set consisted of deidentified, aggregated data. The IRB approval process did not require additional consent from the respondents representing the data.

**Results**

A total of 1079 participants were identified as obese and owners of a smartphone, tablet, or both. Fewer than 15% (156/1079) of the responses were removed due to incomplete or incoherent data, resulting in 923 complete observations, with mHealth app use (dependent variable), eHealth literacy skills dimensions (main independent variable), and additional covariates summarized in Table 1. The average age was 53.51 (SD 14.91) years, and most participants were female (550/923, 59.6%) and non-Hispanic White (543/923, 58.8%). A college degree or above was the highest level of education for 43.3% (400/923) of study participants, and they were mainly employed (with a single or multiple employer; 594/923, 64.4%) and covered by health insurance (872/923, 94.5%). The South contained the largest percentage of participants (436/923, 46.7%), which also corresponds to the nation’s most populous region [52]. Individuals in households earning less than US $50,000 of participants had an annual household income that fell within the range, containing the median annual household income in the United States of US $67,521 in 2020 [53].

The majority of participants (482/923, 52.2%) did not use mHealth apps, resulting in a balanced outcome variable. Within the eHealth literacy skills access dimension, 77.5% (715/923) of respondents used an electronic device to look for health or medical information for themselves within the past 12 months, and approximately half (468/923, 50.7%) used electronic means to look up medical test results, also within the past 12 months. Within the eHealth literacy skills application dimension, 55.6% (513/923) of respondents used email or the internet to communicate with a doctor or doctor’s office within the past 12 months, and 53.3% (429/923) made an appointment with a health care provider through electronic means in that same time period. We also examined the univariate association between mHealth app use, covariates, and main independent covariates.

Table 2 summarizes results from weighted chi-square and t tests for univariate associations between mHealth app use and each of the covariates. Most covariates and all the main covariates were found to be significant at the 5% level. All eHealth literacy skills dimensions were found to be significantly associated with mHealth app use based on univariate weighted chi-square tests (\( P<.001 \)). Similarly, all demographic factors were found to be significantly associated with mHealth app use except for employment status (\( P=.20 \)) and Census region (\( P=.16 \)). Figure 1 displays pairwise weighted 100% stacked bar charts for each of the eHealth literacy skills dimensions versus mHealth app use. Figure 2 portrays a joint network representation of the weighted associations between the eHealth literacy skills dimensions and mHealth app use, which demonstrate strong positive associations both between the skills dimensions as well as between those and the outcome (mHealth app use). Figures S1-S9 in Multimedia Appendix 1 include a weighted box plot (age) and weighted 100% stacked bar charts visualizing the univariate associations with mHealth app use.

Table 3 presents the results of the multivariable weighted logistic regression model. The adjusted odds of using an mHealth app are 3.13 (95% CI 1.69-5.80) times higher among those who responded with an access eHealth literacy skill of using an electronic device to look for health or medical information for themselves within the past 12 months. Similarly, those with an application eHealth literacy skill of using email or the internet to communicate with a doctor or doctor’s office within the past 12 months experience 2.99 (95% CI 1.67-5.37) times higher odds of using an mHealth app compared to those without this skill. Sociodemographic factors found to be significantly associated with mHealth app use include age, disabled or retired status, single or never married or widowed, education, and Hispanic ethnicity. Each additional year of age is associated with 4% lower odds of using mHealth apps (OR 0.96, 95% CI 0.94-0.98). Disabled and retired participants experienced 4.21 (95% CI 1.28-13.82) and 2.53 (95% CI 1.14-5.60) higher odds, respectively, of using mHealth apps compared to those who were employed. Single or never married and widowed participants experienced 49% and 81% lower odds of mHealth app use, respectively, than those who are married, living as married, or living with romantic partners. Previous work has...
indicated that surrogate seekers, those who may seek health information on behalf of others, were more likely to be married or have someone close to them with a chronic illness [54]. Those who received more than 11 years of formal education experienced higher odds of mHealth app use than those with 11 years or less, with OR estimates ranging from 7.77 to 17.24, though with substantially wide CIs. Hispanic participants experienced higher odds of using mHealth apps than non-Hispanic White participants (OR 2.61, 95% CI 1.28-5.33). Insurance status, sex at birth, annual household income, and Census region were not found statistically significant upon adjusting for the other covariates, though there is some level of collinearity present among sociodemographic covariates, as demonstrated in the univariate significance of some of these variables. The multivariable weighted logistic regression adjusted for sociodemographic characteristics showed relatively strong explanatory power with a pseudo-$R^2$ of 0.32 and AUC of 0.7957 (95% CI 0.7671-0.8243). The corresponding ROC is included in Multimedia Appendix 1.
Table 1. Unweighted characteristics of study participants (n=923) using the 2020 Health Information National Trends Survey data set.

<table>
<thead>
<tr>
<th>Sociodemographic variables</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>53.51 (14.91)</td>
</tr>
<tr>
<td><strong>Health insurance, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Insured</td>
<td>872 (94.5)</td>
</tr>
<tr>
<td>Uninsured</td>
<td>51 (5.5)</td>
</tr>
<tr>
<td><strong>Sex at birth, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>550 (59.6)</td>
</tr>
<tr>
<td>Male</td>
<td>373 (40.4)</td>
</tr>
<tr>
<td><strong>Employment status, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Disabled</td>
<td>67 (7.3)</td>
</tr>
<tr>
<td>Employed</td>
<td>494 (53.5)</td>
</tr>
<tr>
<td>Homemaker</td>
<td>25 (2.7)</td>
</tr>
<tr>
<td>Multiple</td>
<td>100 (10.8)</td>
</tr>
<tr>
<td>Retired</td>
<td>183 (19.8)</td>
</tr>
<tr>
<td>Unemployed</td>
<td>42 (4.6)</td>
</tr>
<tr>
<td>Other</td>
<td>12 (1.3)</td>
</tr>
<tr>
<td><strong>Marital status, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Married or living as married or with a romantic partner</td>
<td>518 (56.1)</td>
</tr>
<tr>
<td>Separated or divorced</td>
<td>180 (19.5)</td>
</tr>
<tr>
<td>Single or never married</td>
<td>159 (17.2)</td>
</tr>
<tr>
<td>Widowed</td>
<td>66 (7.2)</td>
</tr>
<tr>
<td><strong>Education, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>≤11 years</td>
<td>49 (5.3)</td>
</tr>
<tr>
<td>12 years or completed high school</td>
<td>174 (18.9)</td>
</tr>
<tr>
<td>Post–high school training other than college (vocational or technical)</td>
<td>68 (7.4)</td>
</tr>
<tr>
<td>Some college</td>
<td>232 (25.1)</td>
</tr>
<tr>
<td>College graduate</td>
<td>241 (26.1)</td>
</tr>
<tr>
<td>Postgraduate</td>
<td>159 (17.2)</td>
</tr>
<tr>
<td><strong>Annual household income (US $), n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>0-9999</td>
<td>51 (5.5)</td>
</tr>
<tr>
<td>10,000-14,999</td>
<td>49 (5.3)</td>
</tr>
<tr>
<td>15,000-19,999</td>
<td>37 (4)</td>
</tr>
<tr>
<td>20,000-34,999</td>
<td>113 (12.2)</td>
</tr>
<tr>
<td>35,000-49,999</td>
<td>132 (14.3)</td>
</tr>
<tr>
<td>50,000-74,999</td>
<td>174 (18.9)</td>
</tr>
<tr>
<td>75,000-99,999</td>
<td>128 (13.9)</td>
</tr>
<tr>
<td>100,000-199,999</td>
<td>198 (21.5)</td>
</tr>
<tr>
<td>≥200,000</td>
<td>41 (4.4)</td>
</tr>
<tr>
<td><strong>Race and ethnicity, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Black or African American</td>
<td>148 (16)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>171 (18.5)</td>
</tr>
<tr>
<td>Non-Hispanic Asian</td>
<td>18 (2)</td>
</tr>
<tr>
<td>Non-Hispanic Native Hawaiian or Other Pacific Islander or American Indian or Alaska Native</td>
<td>9 (1)</td>
</tr>
<tr>
<td>Sociodemographic variables</td>
<td>Values</td>
</tr>
<tr>
<td>--------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>543 (58.8)</td>
</tr>
<tr>
<td>Non-Hispanic multiple races</td>
<td>34 (3.7)</td>
</tr>
<tr>
<td>Census region, n (%)</td>
<td></td>
</tr>
<tr>
<td>Midwest</td>
<td>155 (16.8)</td>
</tr>
<tr>
<td>Northeast</td>
<td>141 (15.3)</td>
</tr>
<tr>
<td>South</td>
<td>431 (46.7)</td>
</tr>
<tr>
<td>West</td>
<td>196 (21.2)</td>
</tr>
<tr>
<td>Outcome variable, n (%)</td>
<td></td>
</tr>
<tr>
<td>mHealth app use</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>482 (52.2)</td>
</tr>
<tr>
<td>Yes</td>
<td>441 (47.8)</td>
</tr>
<tr>
<td>Main covariates, n (%)</td>
<td></td>
</tr>
<tr>
<td>eHealth literacy skills access dimension</td>
<td></td>
</tr>
<tr>
<td>Electronic health information for self</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>208 (22.5)</td>
</tr>
<tr>
<td>Yes</td>
<td>715 (77.5)</td>
</tr>
<tr>
<td>Electronic test results</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>455 (49.3)</td>
</tr>
<tr>
<td>Yes</td>
<td>468 (50.7)</td>
</tr>
<tr>
<td>eHealth literacy skills application dimension</td>
<td></td>
</tr>
<tr>
<td>Electronic communication with doctor or doctor’s office</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>410 (44.4)</td>
</tr>
<tr>
<td>Yes</td>
<td>513 (55.6)</td>
</tr>
<tr>
<td>Made provider appointments electronically</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>431 (46.7)</td>
</tr>
<tr>
<td>Yes</td>
<td>492 (53.3)</td>
</tr>
</tbody>
</table>
Table 2. Weighted chi-square and 2-tailed \( t \) tests (test statistics and \( P \) values) for univariate associations between mHealth app use (dependent variable) and each of the covariates.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Chi-square (df)</th>
<th>( P ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sociodemographic factors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td>3.77 (921)(^{a})</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Health insurance</td>
<td>6.28 (1)</td>
<td>.01</td>
</tr>
<tr>
<td>Sex at birth</td>
<td>6.25 (1)</td>
<td>.01</td>
</tr>
<tr>
<td>Employment status</td>
<td>8.54 (6)</td>
<td>.20</td>
</tr>
<tr>
<td>Marital status</td>
<td>18.55 (3)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Education</td>
<td>78.02 (5)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Annual household income</td>
<td>34.38 (8)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Race and ethnicity</td>
<td>11.31 (5)</td>
<td>.046</td>
</tr>
<tr>
<td>Census region</td>
<td>5.18 (3)</td>
<td>.16</td>
</tr>
<tr>
<td><strong>Main covariates</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eHealth literacy skills: access dimension</td>
<td></td>
<td></td>
</tr>
<tr>
<td>eHealth information for self</td>
<td>95.60 (1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Electronic test results</td>
<td>97.48 (1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>eHealth literacy skills: application dimension</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electronic communication with doctor or doctor’s office</td>
<td>127.87 (1)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Made provider appointments electronically</td>
<td>81.48 (1)</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

\(^{a}\) \( t \) test was used for the univariate analysis.
Figure 1. Visualization of weighted 100% stacked bar charts for each of the eHealth literacy skills dimensions (main covariates) against mHealth app use (outcome). A: Electronic health information for self; B: Electronic test results; C: Electronic communication with doctor or doctor’s office; D: Made appointments electronically.

Figure 2. Ising model-weighted network visualization of eLASSO associations (unadjusted by other covariates and with 0.25 penalization factor) between the eHealth literacy skills dimensions (SHI: self-health information; TD: talk to a doctor or doctor’s office; TR: test results; MA: made appointments; and mH: mHealth app use). Thicker edges (lines) between nodes (circles) represent stronger associations.
Table 3. Adjusted odds ratios (ORs), corresponding 95% CIs, and P values for the multivariable weighted logistic regression model assessing mHealth app use (n=923). The regression model included adjustments for eHealth literacy, age, insurance status, sex, employment and marital status, education, income, race and ethnicity, and census region.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>OR (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Explanatory demographic variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td>0.96 (0.94-0.98)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Insured</td>
<td>2.25 (0.94-5.38)</td>
<td>.07</td>
</tr>
<tr>
<td>Male</td>
<td>0.75 (0.46-1.20)</td>
<td>.22</td>
</tr>
<tr>
<td><strong>Employment status</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employed (reference)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Disabled</td>
<td>4.21 (1.28-13.82)</td>
<td>.02</td>
</tr>
<tr>
<td>Homemaker</td>
<td>2.10 (0.66-6.70)</td>
<td>.21</td>
</tr>
<tr>
<td>Multiple</td>
<td>2.16 (0.92-5.09)</td>
<td>.08</td>
</tr>
<tr>
<td>Retired</td>
<td>2.53 (1.14-5.60)</td>
<td>.02</td>
</tr>
<tr>
<td>Unemployed</td>
<td>1.12 (0.44-2.87)</td>
<td>.81</td>
</tr>
<tr>
<td>Other</td>
<td>0.21 (0.04-1.12)</td>
<td>.07</td>
</tr>
<tr>
<td><strong>Marital status</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married or living as married or with a romantic partner (reference)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Separated or divorced</td>
<td>0.67 (0.34-1.33)</td>
<td>.25</td>
</tr>
<tr>
<td>Single or never married</td>
<td>0.51 (0.27-0.96)</td>
<td>.04</td>
</tr>
<tr>
<td>Widowed</td>
<td>0.19 (0.06-0.57)</td>
<td>.003</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤11 years (reference)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>12 years or completed high school</td>
<td>7.77 (2.08-29.01)</td>
<td>.002</td>
</tr>
<tr>
<td>Post–high school training other than college</td>
<td>12.75 (3.18-51.17)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Some college</td>
<td>9.25 (2.60-32.98)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>College graduate</td>
<td>14.01 (3.68-53.26)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Postgraduate</td>
<td>17.24 (4.09-72.64)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Annual household income (US $)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;10,000 (reference)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>10,000-14,999</td>
<td>1.67 (0.38-7.45)</td>
<td>.50</td>
</tr>
<tr>
<td>15,000-19,999</td>
<td>0.81 (0.20-3.18)</td>
<td>.76</td>
</tr>
<tr>
<td>20,000-34,999</td>
<td>1.31 (0.41-4.26)</td>
<td>.65</td>
</tr>
<tr>
<td>35,000-49,999</td>
<td>2.47 (0.73-8.37)</td>
<td>.15</td>
</tr>
<tr>
<td>50,000-74,999</td>
<td>2.27 (0.71-7.27)</td>
<td>.17</td>
</tr>
<tr>
<td>75,000-99,999</td>
<td>3.16 (0.90-11.04)</td>
<td>.07</td>
</tr>
<tr>
<td>100,000-199,999</td>
<td>2.47 (0.72-8.40)</td>
<td>.15</td>
</tr>
<tr>
<td>≥200,000</td>
<td>1.81 (0.37-8.93)</td>
<td>.47</td>
</tr>
<tr>
<td><strong>Race and ethnicity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black or African American</td>
<td>1.05 (0.51-2.15)</td>
<td>.90</td>
</tr>
<tr>
<td>Hispanic</td>
<td>2.61 (1.28-5.33)</td>
<td>.008</td>
</tr>
<tr>
<td>Non-Hispanic Asian</td>
<td>0.30 (0.06-1.50)</td>
<td>.14</td>
</tr>
<tr>
<td>Non-Hispanic Native Hawaiian or Other Pacific Islander or American Indian or Alaska Native</td>
<td>1.64 (0.17-16.19)</td>
<td>.67</td>
</tr>
<tr>
<td>Non-Hispanic White (reference)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Characteristics</td>
<td>OR (95% CI)</td>
<td>P value</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>-------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Non-Hispanic multiple races</td>
<td>1.21 (0.34-4.22)</td>
<td>.77</td>
</tr>
<tr>
<td><strong>Census region</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>South (reference)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Midwest</td>
<td>1.42 (0.73-2.75)</td>
<td>.30</td>
</tr>
<tr>
<td>Northeast</td>
<td>0.89 (0.47-1.69)</td>
<td>.72</td>
</tr>
<tr>
<td>West</td>
<td>1.09 (0.57-2.08)</td>
<td>.80</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.03 (0.00-0.19)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>eHealth literacy skills: access dimension</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electronic health information for self (reference: yes)</td>
<td>3.13 (1.69-5.80)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Electronic test results (reference: yes)</td>
<td>1.55 (0.87-2.73)</td>
<td>.13</td>
</tr>
<tr>
<td><strong>eHealth literacy skills: application dimension</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electronic communication with a doctor or doctor’s office (reference: yes)</td>
<td>2.99 (1.67-5.37)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Made appointments electronically (reference: yes)</td>
<td>1.53 (0.91-2.58)</td>
<td>.11</td>
</tr>
</tbody>
</table>

aN/A: not applicable.

**Discussion**

**Overview**

The purpose of this study was to explore the associations between the 2 dimensions, access and application, of eHealth literacy skills and mHealth app use among US adults with obesity. We used the HINTS 2020 data to explore this potential association with a sample of 923 respondents with complete information represented in the data set. We found that the majority of the respondents had health insurance, were female, and were non-Hispanic White, with an average age of 54 years. Also, more than half of the respondents had some level of college or were college graduates. This study highlights the association between eHealth literacy skills for accessing and the application of health information using mHealth apps among people with obesity.

The weighted univariate analyses demonstrated associations between all of the covariates and mHealth app use except employment status and census region. Socioeconomic factors of education and income have been found to be important in the general use of content within digital environments (ie, internet) [44]. However, more specifically to this study, these factors are important in showing the potential relationship they have with mHealth app use among people with obesity. When considering weight management or physical activity interventions using mHealth apps, future interventions should attempt to improve the eHealth literacy of participants by targeting segments of people with obesity identified to be more at risk, such as older individuals with obesity or those in lower income brackets. These initial metrics can be collected through a variety of eHealth literacy assessment tools. The eHealth Literacy Scale, for example, has been studied in diverse languages and populations, and it was designed to convey an estimate of people’s eHealth-related skills. Other instruments, such as the eHealth Literacy Questionnaire, were established to support researchers, designers, and the government in evaluating, developing, and applying effective digital health interventions [54]. Previous research identified that patients with adequate eHealth literacy had more ability to seek health information on the internet and find reliable and high-quality information than patients with inadequate eHealth literacy [55,56].

Accessing health information requires active information-seeking skills. Additionally, context and behaviors to gain information are intertwined in this process. Respondents’ access to web-based health information (seeking health information for themselves) can be informed by the Lenz search behavior stage. The respondents’ access to information requires a search for information from impersonally related sources. There is no indication of the familiarity that respondents have with these resources based on the survey questions. People with obesity who use mHealth devices may exhibit multiple factors in their search for health information, and recent eHealth literacy work supports the nuances involved in seeking health information [57,58].

The information acquired through the information-seeking process impacts an information seeker’s decision-making process. Electronic communication with the doctor’s office can be the result of gathering enough information to move forward based on their original goal or a stop in the information acquisition process that prompts information seeking through a personal connection through digital communication. The results from this work are not intended to model these variables but demonstrate that Lenz’s model, enhanced with recent theories, may help inform studies aimed at understanding active information seeking at the intersection of digital health devices such as mHealth and eHealth literacy skills. Recent models and theories commonly demonstrate that health information-seeking behavior involves the action of seeking out information, irrespective of how or why it is sought [36].

Consistent with Mahmood et al [59], education and age are important sociodemographic factors associated with mHealth app use among people with obesity. As access to health services
issues through the use of telehealth technology embedded within mHealth apps. It is imperative that this population group be able to benefit from this type of health service [60]. People with lower education levels and older individuals experience more limited eHealth literacy skills and lower mHealth app use, further widening the digital divide gap [18,24]. In addition, when we examine other sociodemographic factors, marginalized ethnic groups such as Hispanic populations may have access to mHealth apps but experience digital divide issues [30,61].

Issues such as use and knowledge as they relate to using mHealth apps can also contribute to the digital divide [62]. Additional attention is needed to focus on these vulnerable populations. Interventions that can attempt to address this issue are the development of apps and health promotion campaigns that are designed to be culturally relevant [63]. Within the realm of health promotion and wellness, mHealth mindfulness approaches have been used for African American populations [64]. Moreover, studies have described the importance of mHealth interventions with phone features that are familiar to the target population group [62]. There should be consideration of acceptability and efficacy during the developmental phases to support the use of mHealth apps. When considering efficacy, simpler solutions in app design and use should be evaluated. For the older population, features such as 1-click access to a dashboard within health apps that are appropriately displayed in size may be appropriate. Future mHealth apps should also consider health education–related features to support users with low eHealth literacy skills [30,65]. The recent COVID-19 pandemic highlighted the continued digital divide and the disparity in health care services for those who lack sufficient digital literacy skills [66]. mHealth apps benefit people with various chronic conditions, including obesity. People with obesity are less likely to benefit from these mHealth apps if they have low eHealth literacy skills.

mHealth apps facilitate access to health information that has increasingly migrated to web-based spaces [67]. More importantly, mHealth devices assist individuals with seeking health information and decision-making regarding their health [68]. mHealth apps are also advantageous to improve access to health information for personal health data management [15,69]. Since we found that the eHealth literacy skills dimension of access for people with obesity is associated with higher odds of using mHealth apps for seeking health information for themselves, health services should reconsider how they disseminate health information to reach higher proportions of the population. Inevitably, the accessibility of web-based health information has changed the way people engage in health decision-making [70]. This is also evident from our network analysis results, which demonstrate the interconnectivity among all elements related to eHealth literacy skills and mHealth app use among people with obesity, resulting in the need for holistic solutions to enhance mHealth app use and access to health information. Lenz’s model primarily focuses on the search process and use of the information; however, future studies should consider the nuanced contextual factors for people with obesity and their use of mHealth-related devices.

Accessing health information through mHealth apps streamlines the application of health information for decision-making. Many people with obesity have additional chronic diseases that can benefit from timely communication with their health care provider [71]. Effective communication is important for reported satisfaction and perceived health management outcomes. Face-to-face communication has been the standard for communication among patients and health care providers. However, there are mixed results on the perceived effectiveness of face-to-face communication versus IT-aided communication such as mHealth devices [72]. Recent studies have found that mHealth apps are viewed as useful by patients for improving communication and the accessibility of health data [73]. Therefore, this constant communication creates the potential for a bidirectional channel of communication among people with obesity and their health care providers. An in-depth content analysis of vaccination apps showed that few apps provide the capability for bidirectional communication among users and health care providers [74]. The challenges of bidirectional communication can be attributed to barriers to data integration. Given the numerous mHealth apps available for download, this creates interoperability challenges for electronic health care record systems [75]. For mHealth apps that are designed to improve physician and consumer communication, transdisciplinary scholarship is necessary to overcome these barriers. More importantly, technical and networking policies must be developed to support and incentivize the ability to improve this type of communication.

This study benefits from the use of a nationally representative sample of noninstitutionalized US adults. This study provides an adjusted analysis of the associations between mHealth app use and eHealth literacy skills among people with obesity. New technologies that require eHealth literacy skills are transforming how we receive health care and access health information, but they also highlight new disparities as they relate to digital health services [30]. However, to address the rise of chronic conditions such as obesity, it is essential to empower patients to engage in their own health management. One promising strategy is using mHealth apps as a complementary tool to manage weight loss and track physical activity [26]. We provide evidence of several significant factors that can be informative when designing inclusive mHealth app-based health intervention studies. Our research also has implications for studies aimed at managing weight loss or tracking the physical activity of people with obesity to assist with mHealth app development and uptake.

Concerning limitations, first, there could be additional confounding variables that are not included in the study, which is limited by the survey design questionnaire. Some of these confounding variables may be related to self-care behaviors or use patterns with mHealth apps [25]. Furthermore, a bias in the survey design includes the assumption that apps are used only on tablets or smartphones, such that only individuals who indicated having a tablet or smartphone were asked within the survey about having or using health or wellness apps. Second, respondents were only asked about access to information within the previous 12 months. There is a possibility that users do not access or seek health information between visits with their doctors on a yearly basis. Nevertheless, many patients with low health literacy are often left dissatisfied and unsure of the information shared by their doctor and seek third-party sources.
such as web-based health communities to fill those gaps [76,77]. Also, respondents who report ownership of a tablet, smartphone, or both may also use a computer, but they did not indicate that as part of their response. Third, results from the Ising model visualization show a strong relationship between respondents seeking health information for themselves and mHealth app use. As a result of seeking health information, this may also explain the strong relationship between mHealth app use and talking with one's doctor. Therefore, these correlations may exist because they are measuring the same events. Also, some sociodemographic variables used in this study are correlated (eg, age and retired status), so some multicollinearity may be present. Fourth, with a small sample of uninsured people represented in our sample, the statistical significance for health insurance in our model may have been different with a larger sample of uninsured people. Additionally, the data did not provide a distinction between private and public health insurance, though the information content of such a factor may already be embedded in the income variable. A study aimed at self-monitoring of diet, physical activity, and weight among patients who were underinsured or uninsured demonstrated higher adherence through the use of 2 mHealth-related apps in comparison to a paper group [6]. Also, the sample in this study covers the COVID-19 peak period in early 2020, which may have represented a crest (and potentially a permanent shift in behaviors) in electronic access to health information among people with obesity compared to previous time periods. Lastly, there is a limitation in the HINTS survey questions as they were not designed using a web-based health information–seeking behavior framework, though we were able to detect relevant associations even with this design limitation.

Future directions of this work should consider this model structure for people without obesity. A comparative analysis may identify whether eHealth literacy relevance differs between individuals with obesity and those without obesity. This work also considered mHealth app use, but it did not examine cognitive motivational factors for mHealth use. The identification of motivational barriers and facilitators can be analyzed within the context of psychological motivation frameworks to identify potential intervention targets to leverage in mHealth intervention–based studies. Also, since the COVID-19 pandemic may have brought behavioral changes in the overall population regarding eHealth literacy, a dynamic study that explores those changes over time could highlight whether segments of people with obesity may now experience heightened needs compared to prepandemic stages.

Conclusion
This study estimated the associations between mHealth app use and eHealth literacy skills. Our findings are consistent with previous literature, showing that eHealth literacy skills are associated with accessing digital health information and the application of digital health services. For example, age is negatively associated with mHealth app use among people with obesity, with other sociodemographic factors also showing strong associations. This highlights substantial uneven access to eHealth information among people with obesity, potentially leading to disparities in health outcomes among sociodemographic groups. It is imperative that this phenomenon be further investigated as digital health–related services that involve the use of mHealth apps become more integrated into health care services and aim to reach wider segments of the population. A continued challenge is to engage people with chronic conditions such as obesity to use mHealth apps, especially older individuals with obesity and those with lower educational backgrounds. Our work provides evidence of factors associated with mHealth app use in relation to access and application. This work provides an initial understanding of mHealth app use and eHealth literacy skills among people with obesity, and future studies should identify equitable solutions for people with obesity (as well as other groups) and their use of mHealth apps.

Acknowledgments
This project was funded by the National Institute of Health National’s Heart, Lung, and Blood Institute (grant R25HL126145). Dr Thorpe was funded by U54MD000214.

Authors' Contributions
The review concept was designed by GSJ with support from KN and RJTJ. The literature search, screening, analysis, data collection, and data cleaning were conducted by GSJ with support from BDC. LHG provided the data analysis. GSJ, LHG, and BDC wrote the original draft, with review and edits by GSJ, LHG, BDC, RJTJ, and KN. The figures were designed by LHG. Additional review was carried out by GSJ, LHG, RJTJ, and KN.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Supplementary figures and table. [DOCX File, 212 KB - mhealth_v12i1e46656_app1.docx]

References


20. Apps can help keep older folks healthy—but most don't use them. Michigan Medicine-University of Michigan. URL: https://medicine.umich.edu/dept/dgpm/news/archive/2020203/apps-can-help-keep-older-folks-healthy-most-dont-use-them [accessed 2023-10-27]


46. Shaw Jr et al. JMIR MHEALTH AND UHEALTH 2024; vol. 12 | e46656 | p.733

https://mhealth.jmir.org/2024/1/e46656


Abbreviations

AUC: area under the curve
HINTS: Health Information National Trends Survey
IRB: institutional review board
mHealth: mobile health
OR: odds ratio
ROC: receiver operating characteristic

©George Shaw Jr, Bianca A Castro, Laura H Gunn, Keith Norris, Roland J Thorpe Jr. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 10.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Barriers and Implications of 5G Technology Adoption for Hospitals in Western China: Integrated Interpretive Structural Modeling and Decision-Making Trial and Evaluation Laboratory Analysis

Linyun Zhou1,2, MPH; Minghuan Jiang3, PhD; Ran Duan4, PhD; Feng Zuo5, MSc; Zongfang Li2,6, PhD; Songhua Xu2, PhD

1Department of Health Management, The Second Affiliated Hospital of Xi’an Jiaotong University, Xi’an, China
2Institute of Medical Artificial Intelligence, The Second Affiliated Hospital of Xi’an Jiaotong University, Xi’an, China
3Department of Pharmacy Administration and Clinical Pharmacy, School of Pharmacy, Xi’an Jiaotong University, Xi’an, China
4Information Department, Chongqing General Hospital, Chongqing, China
5Information Center, The First Affiliated Hospital of Army Medical University (Southwest Hospital), Chongqing, China
6Clinical Research Center for Hepatic & Splenic Diseases of Shaanxi Province, The Second Affiliated Hospital of Xi’an Jiaotong University, Xi’an, China

Corresponding Author:
Songhua Xu, PhD
Institute of Medical Artificial Intelligence
The Second Affiliated Hospital of Xi’an Jiaotong University
No 5 Jianqiang Road
Xi’an, 710016
China
Phone: 86 029 86320798
Email: songhuaxu@126.com

Abstract

Background: 5G technology is gaining traction in Chinese hospitals for its potential to enhance patient care and internal management. However, various barriers hinder its implementation in clinical settings, and studies on their relevance and importance are scarce.

Objective: This study aimed to identify critical barriers hampering the effective implementation of 5G in hospitals in Western China, to identify interaction relationships and priorities of the above-identified barriers, and to assess the intensity of the relationships and cause-and-effect relations between the adoption barriers.

Methods: This paper uses the Delphi expert consultation method to determine key barriers to 5G adoption in Western China hospitals, the interpretive structural modeling to uncover interaction relationships and priorities, and the decision-making trial and evaluation laboratory method to reveal cause-and-effect relationships and their intensity levels.

Results: In total, 14 barriers were determined by literature review and the Delphi method. Among these, “lack of policies on ethics, rights, and responsibilities in core health care scenarios” emerged as the fundamental influencing factor in the entire system, as it was the only factor at the bottom level of the interpretive structural model. Overall, 8 barriers were classified as the “cause group,” and 6 as the “effect group” by the decision-making trial and evaluation laboratory method. “High expense” and “organizational barriers within hospitals” were determined as the most significant driving barrier (the highest R–C value of 1.361) and the most critical barrier (the highest R+C value of 4.317), respectively.

Conclusions: Promoting the integration of 5G in hospitals in Western China faces multiple complex and interrelated barriers. The study provides valuable quantitative evidence and a comprehensive approach for regulatory authorities, hospitals, and telecom operators, helping them develop strategic pathways for promoting widespread 5G adoption in health care. It is suggested that the stakeholders cooperate to explore and solve the problems in the 5G medical care era, aiming to achieve the coverage of 5G medical care across the country. To our best knowledge, this study is the first academic exploration systematically analyzing factors resisting 5G integration in Chinese hospitals, and it may give subsequent researchers a solid foundation for further studying the application and development of 5G in health care.

(JMIR Mhealth Uhealth 2024;12:e48842) doi:10.2196/48842
KEYWORDS
5G health care; 5G adoption barriers; 5G adoption strategy; smart health care; Western China hospitals

Introduction

Background
With the advancement of information and communication technology, along with the gradual improvement of China's medical information system construction, China's medical industry is moving away from 1.0 medical informatization to 3.0 medical intelligence [1]. Leveraging the advantages such as ultralow latency, high capacity, high speed, seamless connectivity, high reliability, and low power consumption [2], 5G technology plays an essential role in realizing the interconnection and remote monitoring of medical equipment, patient monitoring, remote consultation, and other telemedicine scenarios. At the same time, 5G technology accelerates the data collection, circulation, analysis, and feedback of various applications in the broad medical and health field. With the advent of 5G technology, medical information can now flow and be shared seamlessly among patients, medical equipment, and hospital information systems. This has paved the way for hospitals to simplify the entire medical treatment and service process, right from prediagnosis to diagnosis and postdiagnosis stages [3].

In recent years, the Chinese government has invested considerably in developing innovative 5G smart hospitals to offer better health care to patients and improve their internal management. For example, in July 2021, a total of 10 departments including the Ministry of Industry and Information Technology, Office of the Central Cyberspace Affairs Commission, National Development and Reform Commission, Ministry of Education, Ministry of Finance, Ministry of Housing and Urban-Rural Development, Ministry of Culture and Tourism, National Health Commission, State-Owned Assets Supervision and Administration Commission of the State Council, and National Energy Administration, jointly released the Sailing Action Plan for 5G Applications (2021–2023). The plan aims to encourage the development of various 5G medical products such as robots, emergency vehicles, medical access gateways, and intelligent medical equipment across the country. The plan also emphasizes the need to strengthen the deployment of 5G medical and health network infrastructure, focusing on improving the coverage of 5G in top-tier national hospitals, disease prevention and control centers, elderly care institutions, and other critical locations. Additionally, the plan aims to build 5G networks and 5G medical edge clouds to enhance in-hospital medical care and telemedicine [4]. In September 2021, the Ministry of Industry and Information Technology, in collaboration with the National Health Commission, released the “Notice on Announcing Pilot Projects for 5G+Medical and Healthcare Applications.” This announcement identified 988 pilot projects aimed at advancing the application of 5G technology in various health care domains, including first aid, telediagnosis, teletreatment, tele-intensive care, traditional Chinese medicine diagnosis and treatment, hospital management, intelligent disease control, health management, and other directions. Among the pilot projects, as many as 611 5G smart medical projects are led by hospitals (general hospitals, emergency centers, and specialized hospitals) [5].

More and more hospitals in China are investing in 5G construction. For instance, Guangdong Provincial People’s Hospital put into use the 5G hospital in July 2021, aimed to integrate 5G, big data, artificial intelligence, and other new technologies into various medical scenarios such as treatment, teaching, research, management, and service [6]. Shanghai aims to realize 100% 5G in-depth coverage and 5G typical services for all tertiary hospitals and at least 50% 5G in-depth coverage and 5G typical services for other hospitals by 2023 [7]. Sir Run Run Shaw Hospital Affiliated with Zhejiang University School of Medicine successfully performed a cholecystectomy for a patient from Xinjiang Corps Alar Hospital thousands of miles away by leveraging the robotic arm, which achieved a breakthrough in China’s 5G ultraromote robot human liver and gallbladder surgery [8]. The Second Affiliated Hospital of Xi'an Jiaotong University has piloted several scenarios, such as 5G+emergency rescue, 5G+mobile computed tomography, 5G+unmanned aerial vehicle medical delivery, and 5G+integrated remote diagnosis.

However, though 5G technology undoubtedly introduces enormous benefits for hospitals if adequately applied, it has yet to be widely used in many health care scenarios. Hospitals are experiencing various challenges during the actual 5G application process. Different problems are met in the implementation process, including expertise, operation, resource, regulation, and market access factors, as described in the innovation resistance theory (IRT) [9]. At the same time, there are still no systematic studies that have explored the barriers to the adoption of 5G applications in hospitals in Western China. This is particularly important given that technological development in the Eastern region of China is more advanced compared to the Western region. The lag in 5G development in Western China may become another factor that increases the economic imbalance between these 2 regions [10]. Hence, more research is essential for Western China to provide a better understanding of the barriers hindering the adoption of 5G in health care.

Objectives
This study addresses the critical research question below: what are the barriers to implementing 5G in hospitals in Western China? Based on the question, the following research objectives have been formulated: (1) to identify critical barriers hampering the effective implementation of 5G in hospitals in Western China, (2) to identify interaction relationships and priorities of the above-identified barriers, and (3) to assess the intensity of the relationships and cause-and-effect relations between the adoption barriers.

Methods

Ethical Considerations
The data were collected through literature review and anonymous questionnaires, which posed no harm to individuals.
and did not involve sensitive personal information or commercial interests. Based on the Regulation for Ethical Review of Life Sciences and Medical Research involving human beings issued by the Chinese National Health Commission, Ministry of Education, Ministry of Science and Technology, and State Administration of Traditional Chinese Medicine (Chapter III, Article 32) [11], ethical review was exempted for this study.

Study Design

To address objectives 1-3, a 3-stage hybrid research methodology was proposed by the authors. As shown in Figure 1, a flowchart of the research procedure is conducted. In the first stage, barriers are identified from the existing literature and discussed with experts for further modification and addition using the Delphi technique. The second stage incorporated the interpretive structural modeling (ISM) to obtain a hierarchical structure and interrelationship between the barriers. ISM has a significant advantage because it displays conclusions in the form of a hierarchical topology diagram that is highly intuitive. The hierarchical topology diagram clearly explains the causal relationship and ladder structure among system factors. However, more is needed to determine the intensity of the relationship between factors. It needs to provide the cause-and-effect relation among barriers, which limits the ISM approach [12]. The decision-making trial and evaluation laboratory (DEMATEL) method, on the other hand, can precisely overcome the limitations of the ISM approach. It can determine the strength of influence between variables within the identified structure, providing a deeper understanding of the causal relationship between influencing factors [13]. Therefore, this paper intends to combine Delphi expert consultation, ISM, and DEMATEL to study the hierarchical structure of driving factors and the causal relationship between them. The procedure is explained in more detail in the following section.

Figure 1. Flowchart of research methodology. DEMATEL: decision-making trial and evaluation laboratory; ISM: interpretive structural modeling; MICMAC: Matriced Impacts Corises-multiplication Appliance Classement.

Stage I: Determination of the Barriers Using Literature Review and Delphi Expert Consultation

Overview

This study conducted a literature review from November 20 to 30, 2022, to gather the resistance factors toward 5G development in health care. After that, a 2-round Delphi expert consultation was implemented to refine the factors identified from the above literature review on December 30, 2022, and February 9, 2023. The Delphi technique was developed by the Rand Corporation in 1953 and used as a multistage self-completed questionnaire with individual feedback [14]. It was initially developed as a method for forecasting but has since been widely applied in other areas, including health research [15,16].

Step 1: Literature Review

Web of Science, PubMed, Google Scholar, Chinese government’s official websites, and gray literature, including industry reports, were searched by referencing keywords including “5G healthcare,” “5G smart healthcare,” “5G in hospitals,” “5G applications in healthcare,” and “digital health in China.”

Step 2: Selection of the Delphi Panel

To ensure the authoritative scoring results of the consulting expert group on the evaluation indicators, experienced professionals with intermediate or senior titles who are willing to cooperate actively and who are interested in this research were selected. Leaders responsible for informatization work in health authority, heads and frontline employees from the
hospital’s information management department, 5G communication suppliers, and scholars in 5G and hospital informatization field were invited to this panel.

**Step 3: Development of Questionnaire**

To determine the importance of the barriers selected from the literature review, consolidated criteria have been designed in questionnaire format following a 5-point Likert-type scale (5=very important, 4=relatively important, 3=intermediate, 2=unimportant, and 1=very unimportant). To determine the degree of expert authority (Cr), the expert’s familiarity with the indicator (Cs) and the judgment basis (Ca) were collected in the questionnaire. The quantitative values for Cs are divided into 5 levels (0.9=very familiar, 0.7=relatively familiar, 0.5=intermediate, 0.3=unfamiliar, and 0.1=very unfamiliar), and the quantitative values for Ca are present in Table 1. The questionnaire is also designed to allow the experts to offer their judgments, with space provided for them to add, remove, and justify their responses.

**Table 1.** Quantitative values for judgment basis.

<table>
<thead>
<tr>
<th>Judgment basis (Ca)</th>
<th>Quantitative value of influence degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theoretical analysis</td>
<td>High (0.3)</td>
</tr>
<tr>
<td>Practical experience</td>
<td>Intermediate (0.2)</td>
</tr>
<tr>
<td>Learn from domestic and foreign peers</td>
<td>0.1</td>
</tr>
<tr>
<td>Intuition</td>
<td>Low (0.1)</td>
</tr>
</tbody>
</table>

**Step 4: Data Collection**

The questionnaires were distributed and collected via the WeChat platform, the most widely and frequently used mobile social media in China, which is profoundly integrated into the daily life of Chinese people [17]. It is often used for distributing and collecting questionnaires.

**Step 5: Data Analysis**

An analytical stage followed each round of the Delphi questionnaires. The questionnaire recall rate expresses the degree of positivity of the experts. The degree of expert authority Cr can be calculated from the values of Cs and Ca as follows: Cr=(Cs+Ca)/2. The degree of coordination of expert opinions is judged by the coefficients of variation (CVs) and Kendall coefficient of concordance (W). In this study, the barrier screening standard is CV≤0.250. Barriers whose CVs are higher than 0.250 will be modified or deleted. CV is calculated by the mean value and SD. For Kendall coefficient of concordance, the larger the value, the better the coordination of expert opinions. After the analysis, experts’ feedback and perspectives will be presented to all participants.

**Stage II: Development of Research Framework Using ISM**

The ISM method originated from structural modeling and was introduced by Warfield [18] for better decision-making when too many factors or constructs exist. It is a qualitative and interpretive method that involves a mutual learning process that uses the experience of experts to identify the relationship between factors, variables, enablers, and barriers [19,20]. Based on the relationship, an overall multilevel structure is extracted from the complex items. It is very suitable for interdisciplinary research of natural science and social science. The ISM method has been widely used in management and new technology research in different industries.

Referring to the above studies, the basic steps of the ISM method in this study are as follows.

Step 1 involved constructing a “structural self-interaction matrix (SSIM)” for the barriers determined by the Delphi expert consultation. In this step, the symbols “L, M, N, and O” are used to develop a link between the proposed barriers, where L indicates that barrier i has an impact on barrier j, M indicates that barrier j has an impact on barrier i, N indicates that barriers i and j interact with each other, and O indicates that barriers i and j have no interaction with each other.

Step 2 involved converting the SSIM into an “initial reachability matrix (IRM).” In this step, the symbols “L, M, N, and O” are converted into binary elements 0 and 1, and the conversion rules are shown in Table 2.

Step 3 involved checking the transitivity of the IRM to obtain the “final reachability matrix (FRM).” Some new interrelationships between barriers can be established during this step. Transitivity was tested as if barrier A influences barrier B, barrier B influences barrier C, and then, barrier A indirectly influences barrier C.

Step 4 involved performing the level partition through the FRM to get the hierarchy of barriers to plot the ISM. Based on the FRM, a “reachability set,” an “antecedent set,” and an “intersection set” for each barrier were developed.

Step 5 involved building the ISM, checking the model for conceptual inconsistencies, and modifying it accordingly.

Step 6 involved performing the Matriced Impacts Corises-multiplication Appliance Classement (MICMAC) analysis. The driving power (DP) and dependence power (DEP) of the identified barriers based on the FRM were calculated, and the barriers were classified into 4 clusters, known as an autonomous cluster, dependent cluster, linkage cluster, and independent cluster. The details of these 4 clusters are the following:

- Autonomous clusters: the barriers within the autonomous cluster have low DEP and DP. These barriers have no direct
relation with other barriers and can be considered almost isolated from the system.

- Dependent clusters: the barriers in this group do not have robust DP, but their DEP is strong.
- Linkage clusters: the barriers in this cluster are categorized by high DP and DEP. These factors are unstable, so making any changes to them will significantly affect other barriers and may influence them.
- Independent clusters: the barriers within this cluster have high DP and low DEP. These barriers affect other barriers but are less affected.

<table>
<thead>
<tr>
<th>(i,j) in SSIM&lt;sup&gt;b&lt;/sup&gt;</th>
<th>(i,j) in IRM</th>
<th>(j,i) in IRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>L&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>M&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>N&lt;sup&gt;e&lt;/sup&gt;</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>O&lt;sup&gt;f&lt;/sup&gt;</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<sup>a</sup>IRM: initial reachability matrix.
<sup>b</sup>SSIM: structural self-interaction matrix.
<sup>c</sup>L indicates that barrier i has an impact on barrier j.
<sup>d</sup>M indicates that barrier j has an impact on barrier i.
<sup>e</sup>N indicates that barriers i and j interact with each other.
<sup>f</sup>O indicates that barriers i and j have no interaction with each other.

### Stage III: Identification of Cause-and-Effect Group Using DEMATEL

The DEMATEL approach is a system analysis method based on graph theory and matrix tools. It is used to analyze the cause-effect relationship between factors in complex systems and identify the interaction’s intensity [19,21]. The basic steps to carry out DEMATEL analysis are as follows:

**Step 1** establishes the average “direct relation matrix.” In this step, experts are invited to evaluate each barrier’s influence on another using an integer scale. The designed scale has 5 levels, including integers from 0 to 4, where 0 means no impact, 1 means slight impact, 2 means moderate impact, 3 means high impact, and 4 means extremely high impact. Accordingly, the direct influence matrix of each expert is obtained. Then, the average direct relation matrix is obtained by summarizing and averaging all feedback expert data. Given that $k$ is the index of experts from a total of $p$ experts, $q$ is the index of the barriers, and $i$ and $j$ are the indices for 2 barriers, the decision matrix of each expert is given by $D_{ij}$, and then, the direct impact matrix $D$ is given by equation (1): 

$$D = S / x \quad (1)$$

**Step 2** normalizes the average direct relation matrix. The row and maximum value methods are used. The elements of each row in the average direct influence matrix are summed, and then, the maximum value is obtained by comparison. Finally, each element in the average direct relation matrix is divided by the maximum value. The calculation process can be expressed as equation (2):

$$D = S / x \quad (2)$$

**Step 3** calculates the “total influence matrix T” by adding all the direct and indirect effects using equation (3):

$$T = D + F \quad (3)$$

**Step 4** develops the “cause-and-effect diagram” by adding elements of vector $R$ (row) and vector $C$ (column) using equations (4) and (5), where, $R_i$ is the sum of the row and $C_j$ is the sum of the column of the “total influence matrix.” $(R_i - C_j)$ is called the degree of centrality, and $(R_i - C_j)$ is called the degree of cause.

The horizontal and vertical coordinates can be established according to the values of the degrees of centrality and cause. Among them, the degree of centrality is taken as the abscissa, and the degree of cause is taken as the ordinate. Meanwhile, the cause group and effect group are divided according to the positive and negative values of $(R_i - C_j)$. If the value of $(R_i - C_j)$ is greater than 0, it indicates that this factor has a more significant influence on other factors in the system, and it is classified as a causal factor. If the value of $(R_i - C_j)$ is less than 0, it indicates that other barriers influence this factor greater and attribute it to the outcome factor.

### Results

#### Literature Review

The literature review identified 15 factors influencing the adoption of 5G in health care. Based on the IRT, we divided the barriers to adopting innovation into 5 primary dimensions: expertise, operation, resource, regulation, and market access.
Compared with the unified theory of acceptance and use of technology, technology acceptance model or technology acceptance model 2, and theory of reasoned action, IRT mentioned above has been verified as an effective and significant alternative for researchers who aim to uncover resistance factors in the health care context [22]. The details are listed in Table 3.

### Table 3. Barriers influencing adoption of 5G in health care: review of literature.

<table>
<thead>
<tr>
<th>Barriers</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A. Expertise barrier</strong></td>
<td></td>
</tr>
<tr>
<td>A1 Lack of 5G technical talents</td>
<td>5G experts and 5G equipment operators within hospitals are understaffed [23-27].</td>
</tr>
<tr>
<td>A2 Insufficient informatization level</td>
<td>The level of informatization construction of different hospitals is uneven. Significant gaps in equipment networking capabilities, medical data collection, and information integration make it challenging to implement and replicate 5G solutions [28,29].</td>
</tr>
<tr>
<td>A3 Insufficient security verification</td>
<td>Most of the data in the medical field adopt cross-level and multichannel data collection and analysis methods, including hospital management data and private data such as patient physiology, psychology, and behavior data. The security of the data transmitted through 5G network still needs to be verified [28-35].</td>
</tr>
<tr>
<td><strong>B. Operation barrier</strong></td>
<td></td>
</tr>
<tr>
<td>B1 Organizational barriers within hospitals</td>
<td>Not a lot of people understand what 5G is and how it works, and the willingness of traditional hospitals to upgrade and transform 5G networks is relatively low, considering the fact that mature 5G application is mainly concentrated in peripheral medical scenarios such as outpatient guidance and remote consultation [25,28,32,36,37].</td>
</tr>
<tr>
<td>B2 Communication obstacles among hospitals</td>
<td>Communication obstacles exist among hospitals, especially among the higher- and lower-level hospitals [24,28,29,38,39].</td>
</tr>
<tr>
<td><strong>C. Resource barrier</strong></td>
<td></td>
</tr>
<tr>
<td>C1 High expense</td>
<td>Related equipment and communication costs are high, making it difficult for hospitals to afford [25,30,32,40].</td>
</tr>
<tr>
<td>C2 Huge time cost</td>
<td>Installing appropriate equipment and training relevant personnel demand significant time investment [28,40].</td>
</tr>
<tr>
<td>C3 Lack of well-trained medical and technical personnel</td>
<td>Existing medical care and technical personnel are insufficient for 5G integration in medical scenarios [33].</td>
</tr>
<tr>
<td>C4 Lack of mature compatible equipment and systems</td>
<td>It is difficult for 5G network to integrate with existing equipment and systems [31].</td>
</tr>
<tr>
<td><strong>D. Regulation barrier</strong></td>
<td></td>
</tr>
<tr>
<td>D1 Lack of policies related to 5G smart medical integration</td>
<td>Currently, there is no established policy for the integration of 5G smart medical applications [24,28,34].</td>
</tr>
<tr>
<td>D2 Lack of ethics, rights, and responsibilities policies</td>
<td>Lack of policies on ethical controversies, rights, and responsibilities related to the application of 5G [28,31,33,34].</td>
</tr>
<tr>
<td>D3 Lack of standards for corresponding scenarios</td>
<td>There are many 5G smart medical application scenarios; different scenarios have different requirements for network and technical architecture. At present, there is a lack of 5G application standards corresponding to many medical scenarios [24,29,41-43].</td>
</tr>
<tr>
<td><strong>E. Market access barrier</strong></td>
<td></td>
</tr>
<tr>
<td>E1 Lack of unified 5G product standards and listing standards</td>
<td>Emerging 5G smart medical products (such as wearable intelligent terminal equipment and medical instruments) still need unified and perfect listing standards [28,41].</td>
</tr>
<tr>
<td>E2 Lack of complete 5G smart medical product system</td>
<td>5G private network equipment and terminal equipment that meet the customized services of smart medical care still need to be further improved [28].</td>
</tr>
<tr>
<td>E3 Lack of mature business model</td>
<td>There need to be more mechanisms for cross-field cooperation and mature business models [28].</td>
</tr>
</tbody>
</table>

### Delphi Expert Consultation

The Delphi panel in this study comprises 15 members, including practitioners from the health authority, academia, information management departments of the primary, secondary, and tertiary hospitals, and 5G network operators (see Table 4 for panel composition). A 2-round Delphi expert consultation was conducted to explore the views of different experts on the resistance factors toward 5G development in hospitals. The questionnaire was developed based on the literature review. In the first and second round, we distributed 15 questionnaires each time. In the first round, all the distributed questionnaires were retrieved, while in the second round, 12 questionnaires were collected. The positive coefficient of experts=number of questionnaires returned/number of questionnaires distributed, which can reflect the degree of concern of experts to this study. Thus, in the first round, the positive coefficient of experts was 100%, while in the second round, this coefficient was reduced to 80%.
In the first and second rounds, we obtained data related to the degree of expert authority as follows:

- Degree of familiarity (Cs): 0.670 in the first round and 0.680 in the second round. It shows that the authority of experts in the 2 rounds of consultation is relatively high, and the opinions given are representative to a certain extent.
- Judgment basis (Ca): 0.930 in the first round and 0.920 in the second round.
- Authority coefficient (Cr): 0.800 for both the first and second rounds.

After the first round of expert consultation, the indicator adjustments are as follows, the selection results can be seen in Table 5:

1. Deleted indicators: original C3 (lack of well-trained medical and technical personnel), original C4 (lack of mature compatible equipment and systems), and original E3 (lack of mature business model).
2. Modified indicators: A1 (lack of personnel familiar with 5G within the hospitals) and D2 (lack of policies on ethics, rights, and responsibilities in core health care scenarios).
3. Newly added indicators: B3 (lack of cross-unit resource integration channels) and new C3 (lack of means for hospitals to manage their own 5G networks).

The consultation process of the second round is consistent with the first round. According to the expert’s scoring, the CV was calculated, and the W test was carried out. As can be seen in Table 6, none of the CVs for the second round of barriers were higher than 0.250.

In the first and second rounds, we obtained the following data on Kendall coefficient of concordance W test: Kendall coefficient of concordance was 0.195 in the first round and 0.258 in the second round, $\chi^2_{14}=40.854$ in the first round and $\chi^2_{13}=40.320$ in the second round, and P value was <.001 for both first and second rounds, which is statistically significant, indicating that the coordination of expert opinions is good. The experts’ opinions tended to be unanimous in the second round of consultation, with no modification and new indicators. The final determined barriers can be seen in Figure 2.

---

Table 4. Basic information of Delphi panelist.

<table>
<thead>
<tr>
<th>Category</th>
<th>Experts, n</th>
<th>Constituent ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Major</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communication technology</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>Computer science and technology</td>
<td>12</td>
<td>80</td>
</tr>
<tr>
<td>Health management</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td><strong>Work experience (years)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;10</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>10-19</td>
<td>10</td>
<td>67</td>
</tr>
<tr>
<td>20-29</td>
<td>4</td>
<td>26</td>
</tr>
<tr>
<td><strong>Professional title</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intermediate</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Vice senior</td>
<td>11</td>
<td>73</td>
</tr>
<tr>
<td>Senior</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td><strong>Job description</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information technology operations management</td>
<td>10</td>
<td>67</td>
</tr>
<tr>
<td>Academia</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>Administrative management</td>
<td>3</td>
<td>20</td>
</tr>
</tbody>
</table>
Table 5. Selection results of the first round of expert consultation.

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>3.867 (1.024)</td>
<td>4.267 (0.573)</td>
<td>3.867 (0.806)</td>
<td>4.333 (0.789)</td>
<td>3.200 (0.748)</td>
<td>4.133 (0.884)</td>
<td>3.533 (0.718)</td>
<td>3.267 (1.062)</td>
<td>3.467 (0.957)</td>
<td>3.467 (0.573)</td>
<td>4.267 (0.806)</td>
<td>4.000 (0.234)</td>
<td>4.133 (0.182)</td>
<td>3.667 (0.158)</td>
<td></td>
</tr>
<tr>
<td>CV</td>
<td>0.265</td>
<td>0.134</td>
<td>0.208</td>
<td>0.182</td>
<td>0.234</td>
<td>0.214</td>
<td>0.203</td>
<td>0.325</td>
<td>0.314</td>
<td>0.247</td>
<td>0.276</td>
<td>0.134</td>
<td>0.158</td>
<td>0.195</td>
<td>0.340</td>
</tr>
</tbody>
</table>

- CV: coefficient of variation.
- The symbol "✓" indicates that the indicator is deleted.
- The symbol "V" indicates that the indicator is retained.
- The symbol "V*" indicates that the indicator is modified.

Table 6. Selection results of the second round of expert consultation.

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>3.917 (0.640)</td>
<td>4.167 (0.553)</td>
<td>3.917 (0.862)</td>
<td>4.750 (0.433)</td>
<td>3.333 (0.745)</td>
<td>4.250 (0.595)</td>
<td>4.167 (0.799)</td>
<td>3.500 (0.866)</td>
<td>4.167 (0.745)</td>
<td>3.333 (0.808)</td>
<td>4.083 (0.759)</td>
<td>4.000 (0.707)</td>
<td>4.417 (0.759)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CV</td>
<td>0.163</td>
<td>0.133</td>
<td>0.220</td>
<td>0.091</td>
<td>0.224</td>
<td>0.140</td>
<td>0.192</td>
<td>0.247</td>
<td>0.224</td>
<td>0.215</td>
<td>0.232</td>
<td>0.186</td>
<td>0.177</td>
<td>0.172</td>
<td></td>
</tr>
</tbody>
</table>

- CV: coefficient of variation.
- The symbol "✓" indicates that the indicator is retained.

Figure 2. Adoption barriers of 5G for hospitals in Western China.
Results of ISM

Establish SSIM

In total, 10 experts from the abovementioned 15 were invited to develop the relationships between different barriers using L, M, N, and O. Due to the nonexistence of a special criterion in the literature of ISM concerning what establishes the majority and as in the methodology of ISM [44], this study adopted a 50% criterion to determine the final relationships between different barriers, such as in a relation, if 5 from 10 experts ticked L, the corresponding column is defined as L. However, to be unbiased, for this kind of relationship, the number of specialists answering by M or N or O should be lesser than or equivalent to 3. Fulfilling both requirements, 47 of 91 cases of relations were included to obtain final results, and the remaining cases were all taken as biased and to be taken for more discussion with experts. To analyze the relations of the remaining barriers, this study proceeded for more discussion with 2 senior experts. To eliminate biases, the selected experts were taken out of the given pool of 10 experts. The 2 senior experts were requested to analyze and discuss the remaining barriers until a consensus was reached. Doing so determined a final relationship between different barriers called SSIM, as shown in Table S1 in Multimedia Appendix 1.

Formation of IRM

Transformation of the SSIM into IRM uses the binary rule shown in Table 2. For example, the entry of A1 and D2 in the SSIM is “O,” which is replaced by “0” for A1 and D2 and “0” for D2 and A1 in the IRM (Table S2 in Multimedia Appendix 1).

Formation of FRM

FRM is formed after checking IRM for transitivity. It was done to represent all indirect connections to maintain the consistency of relationships among the barriers. For example, there is a direct relation between A1 and C2 and C2 and D1, but there is no relation between A1 and D1, as shown in the SSIM. Hence, according to the transitivity rule, there is an indirect relation between A1 and D1 corrected during the formation of FRM. It can be observed in Table S3 in Multimedia Appendix 1, where the relation of A1 and D1 is represented by 1*.

Level Partition

To have a clearer understanding of the relationship between the barriers, a hierarchical structure of the factors is required. Based on FRM, the reachability set, antecedent set, and intersection set for each barrier were developed. Suppose the reachability and intersection set for a specific barrier are identical. In that case, that barrier is deemed at level 1 and assigned the highest position in the ISM hierarchy. After the first iteration, the barriers constituting level 1 are removed, and the previously mentioned procedure is repeated with the remaining barriers until the levels of all barriers have been determined. The results of the different sets and the level iterations are shown in Table S4 in Multimedia Appendix 1.

Formation of ISM

ISM is formulated based on the partition level of barriers. In the first iteration, A2 (insufficient informatization level), A3 (insufficient security verification), B1 (organizational barriers within the hospitals), B2 (communication obstacles among hospitals), B3 (lack of cross-unit resource integration channels), C2 (huge time cost), C3 (lack of means for hospitals to manage their own 5G networks), and E2 (lack of complete 5G smart medical product system) were placed at the top of the ISM. The second iteration resulted in second-level barriers involving A1 (lack of personnel familiar with 5G within hospitals), C1 (high expense), D1 (lack of policies related to 5G smart medical integration), D3 (lack of standards for corresponding scenarios), and E1 (lack of unified 5G product standards and listing standards) placed below the first level. Similarly, in the third iteration, D2 (lack of policies on ethics, rights, and responsibilities in core health care scenarios) was placed below the second level. The developed framework or ISM of barrier adoption is shown in Figure 3.
**MICMAC Analysis**

The MICMAC analysis is performed to identify barriers’ DP and DEP and classify them accordingly. As shown in Figures 3 and 4, A2 (insufficient informatization level), A3 (insufficient security verification), B1 (organizational barriers within hospitals), B3 (lack of cross-unit resource integration channels), and C3 (lack of means for hospitals to manage their own 5G networks) were placed at the top of the ISM and fell under the “dependent” cluster. B2 (communication obstacles among hospitals), C2 (huge time cost), and E2 (lack of complete 5G smart medical product system) were categorized under the “linkage” cluster. The barriers under the linkage cluster were volatile due to high DP and DEP. C1 (high expense), D1 (lack of policies related to 5G smart medical integration), D2 (lack of policies on ethics, rights, and responsibilities in core health care scenarios), D3 (lack of standards for corresponding scenarios), and E1 (lack of unified 5G product standards and listing standards) were placed in the independent cluster. Considering these barriers as drivers of other barriers in the system, hospitals should prioritize them in their decision-making processes. In addition, D2 (lack of policies on ethics, rights, and responsibilities in core health care scenarios) has a relatively high driving force and low dependence force, which reveals that it strongly impacts the whole system as displayed in the ISM.

**Figure 4.** Driving and dependence diagram.

**Results of DEMATEL**

The abovementioned 10 experts in the ISM scoring process were also invited to participate in the data collection for the DEMATEL analysis. Experts were invited to evaluate each barrier’s influence on another using a scale of 0-4. After collecting the direct relation matrix of each expert, the average direct relation matrix of Table S5 in Multimedia Appendix 1 was obtained by summarizing and averaging all feedback expert data. Then, the direct relation matrix was converted into a normalized direct relation matrix (Table S6 in Multimedia Appendix 1) using equation (2). Furthermore, the normalized matrix was converted into a total influence matrix (Table S7 in Multimedia Appendix 1) using equation (3). Finally, the degree of influence was calculated using equations (4) and (5). The cause-effect matrix is shown in Table S8 in Multimedia Appendix 1.

The barriers with an R–C value less than 0 were identified as the effect group, while barriers with an R–C value greater than 0 fell under the cause group. As shown in Figure 5, a total of 8 barriers could be classified in the “cause group,” and 6 as the “effect group,” in which C1 (high expense), E1 (lack of unified 5G product standards and listing standards), D1 (lack of policies related to 5G smart medical integration), and D2 (lack of policies on ethics, rights, and responsibilities in core health care scenarios) took high priority in the causal group, B1 (organizational barriers within hospitals) and A2 (insufficient informatization level) were the most influenced barriers.
Sensitivity Analysis

The sensitivity analysis was performed to verify the consistency of the calculated value and validate the stability of professional judgment. As shown in Table 7, a different weighting was assigned to one expert's feedback while keeping equal weightings for the other. Four different total relationship matrices and comparable matrices were created by multiplying each weight assigned to the experts, the average relationship matrices were then computed, and the cause-effect relationships among the different barriers were established. As shown in Table 8, the same rank order for cause-effect barriers for each expert was obtained, accepting minor rank order variation. Based on Figure 6, the plots of all barriers during 4 iterations of sensitivity analysis are the same as the base rank. Therefore, it is clear that there was no major change in barrier rankings. The sensitivity analysis confirms the robustness of obtained results.

Table 7. Results of sensitivity analysis.

<table>
<thead>
<tr>
<th></th>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
<th>Scenario 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R+C</td>
<td>R−C</td>
<td>R+C</td>
<td>R−C</td>
</tr>
<tr>
<td>A1</td>
<td>1.939</td>
<td>0.678</td>
<td>1.939</td>
<td>0.678</td>
</tr>
<tr>
<td>A2</td>
<td>2.862</td>
<td>−1.448</td>
<td>2.862</td>
<td>−1.448</td>
</tr>
<tr>
<td>A3</td>
<td>3.062</td>
<td>−0.864</td>
<td>3.062</td>
<td>−0.864</td>
</tr>
<tr>
<td>B1</td>
<td>4.401</td>
<td>−2.207</td>
<td>4.401</td>
<td>−2.207</td>
</tr>
<tr>
<td>B2</td>
<td>3.982</td>
<td>−0.510</td>
<td>3.982</td>
<td>−0.510</td>
</tr>
<tr>
<td>B3</td>
<td>2.792</td>
<td>−1.266</td>
<td>2.792</td>
<td>−1.266</td>
</tr>
<tr>
<td>C2</td>
<td>3.358</td>
<td>0.287</td>
<td>3.358</td>
<td>0.287</td>
</tr>
<tr>
<td>C3</td>
<td>1.780</td>
<td>−0.720</td>
<td>1.780</td>
<td>−0.720</td>
</tr>
<tr>
<td>D1</td>
<td>3.379</td>
<td>1.171</td>
<td>3.379</td>
<td>1.171</td>
</tr>
<tr>
<td>D2</td>
<td>1.962</td>
<td>0.955</td>
<td>1.962</td>
<td>0.955</td>
</tr>
<tr>
<td>D3</td>
<td>3.843</td>
<td>0.863</td>
<td>3.843</td>
<td>0.863</td>
</tr>
<tr>
<td>E1</td>
<td>3.474</td>
<td>1.373</td>
<td>3.474</td>
<td>1.373</td>
</tr>
<tr>
<td>E2</td>
<td>3.701</td>
<td>0.316</td>
<td>3.701</td>
<td>0.316</td>
</tr>
</tbody>
</table>
Table 8. Ranking obtained after sensitivity analysis.

<table>
<thead>
<tr>
<th></th>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
<th>Scenario 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert 1</td>
<td>0.4</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>Expert 2</td>
<td>0.2</td>
<td>0.4</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>Expert 3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>Expert 4</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Figure 6. Sensitivity analysis plots.

Discussion

Principal Results

As shown in Figures 3 and 5, all the barriers in the bottom and middle levels fell into the cause group, indicating those barriers are the primary reasons why 5G technology cannot be adopted in hospitals. In particular, D2 (lack of policies on ethics, rights, and responsibilities in core health care scenarios) was the only factor at the bottom level, implying it is the underlying influencing factor of the whole system. It also has a relatively high driving force and low dependence force based on the MICMAC analysis, which is consistent with the analysis results of DEMATEL as it took high priority in the causal group with an R–C value of 0.900 and an R+C value of 1.875. The finding aligns with that of [28], which highlights that medical and health care fields are closely linked to people's safety. Additionally, ethical considerations and definitions of responsibilities and risks are the most significant obstacles to the evolution and development of 5G in the core areas of health care. It is suggested that the government should strengthen policy supervision to ensure the safe and ethical use of 5G in the medical field. Specific measures need to be implemented to supervise and regulate the application of 5G in the medical field. In addition, further research on relevant policies and norms is required. At the same time, clarification should be made in policy documents, laws, and regulations.

Another significant obstacle in promoting 5G medical applications is the massive capital investment required, as stated in C1, with the highest R–C value of 1.361, indicating its most significant driving force over other barriers. This finding reflects the facts stated by CN-HEALTHCARE [45]. The overall cost of 5G medical treatment includes the construction of 5G communication networks by operators, procurement and maintenance costs for 5G medical information systems and related medical equipment, purchase of medical care equipment, and services for users. The application of 5G in hospitals also
requires installing indoor base stations, software support, and computer room upgrades, all of which incur significant expenses. The high cost of 5G applications limits the financing capabilities of hospitals, especially in the Western China region, where hospitals are generally smaller and have limited funding. As a result, only a few large hospitals with telecom operator support have been able to implement 5G medical care. Meanwhile, smaller hospitals are expected to undertake 5G telemedicine with larger tertiary hospitals. Thus, at this stage, it is recommended that smaller hospitals increase their collaboration with larger hospitals to accelerate the adoption of 5G. At the same time, the government is essential to develop affordable solutions and provide financial support for 5G adoption in health care.

With the highest R+C value of 4.317, B1 (organizational barriers within hospitals) is the most closely related barrier to other factors, indicating that it is the most important barrier to adopting 5G in hospitals in Western China. It also has the lowest R–C value of −2.107, which means it is in the effect group and the most influenced barriers by other factors. These results echo the findings of CN-HEALTHCARE [46], which identified a lack of mature 5G applications in medical scenarios, varying levels of acceptance of new technology, concerns about 5G security, and limitations in human, financial, and material resources as factors influencing attitudes toward 5G adoption among different hospitals. Organizational barriers may be addressed through effective management and leadership, clear communication, and collaboration among different departments.

E1 (lack of unified 5G product standards and listing standards) has the second-highest R–C value of 1.316 and a relatively high R+C value of 3.300. The high R–C value specifies its driving force over the other barriers, while the high R+C value indicates its strong impact on the adoption process. This outcome is akin to the findings of Bruer and Doug [47], who mentioned that 5G standards play a crucial role in hardware infrastructure to software running on top of components, and unified 5G standards help to ensure that a range of devices and equipment can operate in a shared system. Therefore, developing unified 5G product standards and listing standards for corresponding scenarios are essential. It is recommended that the Chinese government, 5G network operators, and hospital administrators accelerate cooperation to establish unified 5G product standards and list standards to jointly promote the large-scale development of 5G in the health care sector.

B2 (communication obstacles among hospitals) is a part of the linkage cluster with the second-highest R+C value of 3.782, indicating it is one of the crucial factors in the whole system. This finding is consistent with that of Wang et al [39], who found that effective communication is critical for the success of 5G adoption in health care. Communication obstacles can lead to misunderstandings and a lack of trust. It is urgent to break through the communication obstacles among hospitals, especially between the higher- and lower-level hospitals, which need to accelerate top-level design, formulate policy documents, and improve relevant legislation and management mechanisms to promote the opening and sharing of 5G medical data and ensure the deep integration of the 5G medical industry.

Concluding Remarks
The paper comprehensively analyzes barriers to 5G adoption in hospitals in Western China. Experts from different stakeholders validated 14 resistance factors. Next, an integrated ISM-DEMATEL approach was applied to model the barriers as a network of factors and alternatives categorized into clusters. All barriers were related to human expertise, resource allocation, operational procedures, laws and regulations, and market access capability. Overall, the study shows that promoting the integration of 5G in hospitals in Western China faces multiple complex and interrelated barriers. It constructs a framework for the main barriers to 5G adoption in the hospital context and provides regulatory authorities, hospital managers, and telecom operators with theoretical and managerial insights into the strategic pathways.

Theoretical and Managerial Implications
- The barrier at the bottom level of the ISM should be emphasized for short-term strategy. The middle-level barriers can be considered for medium- and long-term strategies. The barriers at the top of the ISM can be a long-term strategic focus.
- The effect group can easily be influenced by the cause group, and therefore, managers should give the most attention to causal barriers when implementing 5G practices in hospitals.
- The ranking of cause-effect group barriers can assist regulatory authorities, hospital managers, and telecom operators in developing strategic policy during 5G implementation.
- To overcome these barriers, hospital managers should formulate a 5G adoption strategy that considers the specific needs of the institution and the local market. Specific measures include increasing investment in information infrastructure and human resources, establishing supplier communication channels, and promoting cross-unit resource integration.
- To expand the use of 5G in health care scenarios, it is recommended that the government accelerates the construction of an innovation system consisting of regulators, hospitals, telecom operators, academic researchers, and patient representatives.

Outlook of 5G Health Care
As the infrastructure of intelligent medical care, 5G allows the transmission of vast amounts of data and information, making the information superhighway a reality. Furthermore, with the full deployment of 5G medical care, especially the integration with big data, artificial intelligence, internet, internet of things, and blockchain technology, 5G is expected to bring significant changes to the current medical and health system and promote the evolution of the entire medical ecology, including hospital operation and management. In the long run, 5G health care promotes the sinking of high-quality medical resources and the development of China's “primary health care” and “family doctor” systems. It can improve the population's overall health, reduce medical expenses for ordinary people, and relieve medical insurance burdens. Therefore, no matter how barriers are faced, the benefits of 5G medical care are expected to
outweigh the costs, making it a worthy investment. In the development of 5G medical care, various technical, economic, institutional, interest, and ethical problems will inevitably be encountered. The regulatory authorities, hospitals, telecom operators, and the public must cooperate to explore and solve the problems in the 5G medical care era, aiming to achieve the coverage of 5G medical care across the country.

Limitations

Although this study suggests some useful implications, there are some limitations that could be considered for future research. First, due to the challenges in reaching out to health care professionals and telecom operators from all regions of China, most of the experts invited for this research were from representative cities in Northwest and Southwest China, including Xi'an, Chongqing, and Chengdu. This may not fully represent the entire country, as Eastern China is generally more developed than Western China. Therefore, there is a lack of balance among the groups of participants in this research. In future research, we plan to invite experts from Eastern China as a complement study. Second, the relations established among barriers might be biased because they are selected and analyzed based on expert opinions that are context-dependent and depend on their organization’s culture and experience. Third, the outcome of this study is valid for the Chinese health care field and cannot be generalized for other sectors without modifications. It can be extended from the Chinese context to a broader coverage by selecting experts from different countries for benchmarking studies. Finally, only 4 groups of experts, namely, government information department staff, managers from the hospital information technology departments, telecom operators, and scholars, are involved in the research process. Other vital stakeholders, such as patients, can also provide crucial information and insights related to the development of 5G health care.

Acknowledgments

The authors acknowledged the support of the Key Research and Development Projects of Shaanxi Province, China (grant 2021SF-188), and the Youth Fund of the Second Affiliated Hospital of Xi'an Jiaotong University, China (grant YJ(QN)202018). The authors also acknowledged the support of Mr Duwei Dai, Ms Caixia Dong, and Ms Chunyan Zhang during the investigation and data analysis process.

Authors’ Contributions

LZ was responsible for the initial concept development, formal analysis, and writing the first draft of the study. MJ contributed to the development of the research methods and edited the writing. RD and FZ were involved in data collection and curation. ZL and SX provided supervision and reviewed the writing. All authors contributed to the final version of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Interpretive structural modeling model and decision-making trial and evaluation laboratory method operation process. [DOCX File, 45 KB - mhealth_v12i1e48842_app1.docx ]

References

7. Shanghai: By 2023, 100% of all tertiary hospitals will achieve 5G deep coverage and 5G typical services online. Shanghai Communications Administration. URL: https://www.cn-healthcare.com/articlewm/20220331/content-1332938.html [accessed 2022-11-22]


45. 5G healthcare: pain points, problems, applications and prospects. CN-HEALTHCARE. URL: https://www.sohu.com/a/426797130_139908 [accessed 2023-03-22]

46. Bottom line “5G+medical care” craze: just a need for hospitals or a commercial gimmick? CN-HEALTHCARE. URL: http://tinyurl.com/3v5r72uy [accessed 2023-03-25]


Abbreviations

CV: coefficient of variation
DEMETATEL: decision-making trial and evaluation laboratory
DEP: dependence power
DP: driving power
FRM: final reachability matrix
IRM: initial reachability matrix
IRT: innovation resistance theory
ISM: interpretive structural modeling
MICMAC: Matriced Impacts Corises-multiplication Appliance Classement
SSIM: structural self-interaction matrix

©Linyun Zhou, Minghuan Jiang, Ran Duan, Feng Zuo, Zongfang Li, Songhua Xu. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 23.01.2024. This is an open-access article distributed under the terms of the Creative Commons License.
Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Views of Service Users, Their Family or Carers, and Health Care Professionals on Telerehabilitation for People With Neurological Conditions in Ghana: Qualitative Study

Lorna Paul1*, PhD; Katie Thomson1*, PhD; Shadrack Osei Asibey2*, PhD; Marian Brady1*, PhD; Frederike van Wijck1*, PhD; Derrick Antwi2*, MSc; Eric Nkansah Opoku1*, MSc; Fred Stephen Sarfo2*, MD

1School of Health and Life Sciences, Glasgow Caledonian University, Glasgow, United Kingdom
2Department of Psychiatry, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

* all authors contributed equally

Corresponding Author:
Lorna Paul, PhD
School of Health and Life Sciences
Glasgow Caledonian University
Cowcaddens Road
Glasgow, G4 0BA
United Kingdom
Phone: 44 01413318108
Email: Lorna.Paul@gcu.ac.uk

Abstract

Background: Up to 50% of people in low- and middle-income countries do not receive the rehabilitation they require. Telerehabilitation has the potential to improve access to neurorehabilitation services especially in low- and middle-income countries. Although there are reports of the barriers and facilitators to telerehabilitation in such settings, almost all are anecdotal. Furthermore, family or carers have a significant influence on the adoption and success of telerehabilitation, but their views have not been reported.

Objective: This study aimed to investigate the views of service users, their family or carers, and health care professionals (HCPs) on telerehabilitation for people with neurological conditions in Ghana.

Methods: Two focus groups were held at Komfo Anokye Hospital in Kumasi, Ghana: one in person for service users (n=11) and their family or carers (n=9), conducted in the Ghanaian language of Twi, and one hybrid for HCPs (n=18) conducted in English. The mean (SD) age of the service users was 59.8 (8.6) years; 5 users had a stroke and 6 had Parkinson disease. The HCP group consisted of 7 speech and language therapists, 3 physiotherapists, 3 occupational therapists, 3 medical staff, 1 nurse, and 1 industry representative. Focus groups were semi-structured and explored previous experiences of telerehabilitation, perceived benefits and challenges, and solutions to overcome these challenges. Focus groups were audio transcribed, and the service user transcript was translated into English. The resulting transcripts were analyzed using thematic analysis.

Results: Overall, participants were positive about the role of telerehabilitation but recommended hybrid delivery, with in-person rehabilitation in the early stages and telerehabilitation in the later stages. In relation to telerehabilitation in Ghana, there were 3 main themes: benefits, challenges or barriers, and implementation. Benefits included the convenience and lower cost for service users, the higher dose of therapy possible, and increased access for people in remote areas. However, challenges included lack of a stable internet connection, cost of phones and data packages, and low levels of literacy. Implementation issues included cultural relevance, information governance, and the platform used to deliver telerehabilitation, with most participants being familiar with WhatsApp.

Conclusions: Telerehabilitation has the potential to be a useful method of delivering rehabilitation to people with neurological conditions in Ghana, especially in a hybrid rehabilitation model with telerehabilitation augmenting in-person sessions. However, many people were unaware of telerehabilitation, and challenges such as a reliable internet connection, cultural relevance, and costs need to be addressed. Clinical trials of low-cost telerehabilitation interventions contextualized to the specific user group are required.

(JMIR Mhealth Uhealth 2024;12:e49501) doi:10.2196/49501
KEYWORDS
telerehabilitation; low- to middle-income country; LMIC; service user; health care professional; qualitative study; caregiver; neurorehabilitation; barriers; facilitators; eHealth; focus group; thematic analysis

Introduction

The World Health Organization defines rehabilitation as “a set of interventions designed to optimize functioning and reduce disability in individuals with health conditions in interaction with their environment” [1]. They also highlight the significant unmet need for rehabilitation services that is most evident in some low- and middle-income countries (LMICs), where up to 50% of people do not receive the rehabilitation they require [1].

Rehabilitation in Ghana, an LMIC, is provided by the public and private sector and costs between US $5 and $20 per session. Rehabilitation sessions are paid for by the individual or through various national insurance schemes. Even where individuals have insurance, the coverage may be limited and “top-up payments” are required. The duration of neurological rehabilitation can vary between 3 and 12 months depending on the condition, and often, there is a waiting list of 1-2 months.

Numerous barriers to rehabilitation have been cited in LMICs, such as Ghana, including low numbers of therapists, especially in rural locations, services being concentrated in main towns or cities with the majority of the population living rurally, transport issues, costs for appointments, and lack of specialist staff and equipment [2-5].

As well as experiencing barriers to rehabilitation, LMICs are experiencing an increase in life expectancy and greater numbers of people with noncommunicable diseases including neurological conditions [2,6,7]. Although in its early stages, telemedicine is increasingly being explored in LMICs to deliver care to challenging or remote areas. Telerehabilitation, a branch of telemedicine, may provide part of the solution to the increased patient demand coupled with restricted service especially in neurological rehabilitation. Telerehabilitation is defined as the provision of rehabilitation, including physiotherapy, speech and language therapy, and occupational therapy, over distance and, oftentimes, using communication technology [8]. Potential telerehabilitation technologies include telephone and video calls, apps, virtual reality, and robotics [4]. A recent review suggests that the telephone or video is the main media through which telerehabilitation is delivered worldwide [9].

The evidence base for the acceptability, feasibility, and clinical and cost-effectiveness of telerehabilitation for people with neurological conditions is increasing. Several recent systematic and scoping reviews suggest that telerehabilitation improves access to services and is generally well received by patients and therapists, with high adherence to telerehabilitation programs and low adverse events, supporting its safety in practice [10,11]. Evidence of the clinical effectiveness of telerehabilitation for people with neurological conditions is mixed, but overall it is reported to be at least equivalent to standard care [12,13]. There is, however, limited evidence on the cost-effectiveness of telerehabilitation [6,9]. Much of the research in telerehabilitation has been undertaken in high-income countries, with notably fewer studies in LMICs [6,14,15], where the rehabilitation context, as well as the barriers and facilitators to the feasibility, adoption, scalability, and sustainability of telerehabilitation, may be quite different. In terms of neurological conditions, most telerehabilitation research in LMICs has been conducted in stroke [3,14,16].

To influence rehabilitation and improve patient outcomes, technology needs to be adopted into services. There are various models and theories of technology adoption including the Technology Acceptance Model and Self-Determination Theory. Central to these models are the beliefs and attitudes of the users of the technology. In the context of telerehabilitation, only a few previous studies exist on the views and beliefs of therapists and patients in LMICs and beyond [7]; however, these studies tend to be process evaluations of specific telerehabilitation interventions that are being researched and thus are open to selection bias as participants are exposed to the technology under study conditions. Furthermore, although telerehabilitation generally happens at home, there is a recognition that almost no research has sought the views of the patient, carer, or family member—although they have a strong and significant influence on the adoption and success, or otherwise, of telerehabilitation interventions [17].

Telerehabilitation is in its infancy in LMICs, including Ghana; however, to ensure successful adoption, implementation to routine practice, and scalability, it is important to understand the views of health care professionals (HCPs), service users (patients), and their carers. Therefore, the aims of this study were to investigate the views of HCPs involved in neurological rehabilitation, service users with a long-term neurological condition, and the carer or family member of someone with a long-term neurological condition in terms of previous experience of telerehabilitation, the perceived potential benefits, the potential challenges, and possible solutions to overcome these challenges in Ghana.

Methods

Study Design

Focus groups were used as the method of data collection for this qualitative study, and this paper is presented in line with the Consolidated Criteria for Reporting Qualitative Research guidelines. Separate focus groups were conducted: one for service users and carers or family members and another for HCPs and IT industry representatives. Both focus groups were conducted in person in Komfo Anokye Hospital in Kumasi, Ghana, with HCPs given the option of joining remotely via teleconference. The service user focus group was conducted in person to ensure that those who did not have technology or technology skills could contribute, to encourage engagement, and to support those with communication problems. For service users and carers traveling, expenses were covered and both groups were provided with lunch at the end of the focus group. The service user and carer focus group was conducted in Twi,
the local language, and the HCP focus group was conducted in English.

**Inclusion Criteria**

Inclusion criteria for the stakeholder group were HCPs (physiotherapists, occupational therapists or speech and language therapists, or medical staff) with experience of working in neurological rehabilitation in Ghana or staff with expertise in commissioning or delivering health services (commissioner or industry expert) and able to speak and understand English and attend the focus group either in person or via teleconference (Zoom).

Inclusion criteria for the service users were people with a neurological condition such as stroke, Parkinson disease (PD), or spinal cord injury and able to travel to the venue. Carers or family members had to have experience of caring for someone with a neurological condition and able to travel to the venue.

**Recruitment**

A convenience sample of service users, who fulfilled the inclusion and exclusion criteria, was identified from two sources: (1) members of a support group for people with PD were telephoned by the research coordinator (SOA) and (2) stroke survivors attending a neurology clinic at Komfo Anokye Hospital 2 weeks before the focus groups were approached by the research coordinator. The study was explained to potential participants, and if they were happy to take part, they were given the details of the focus group date, time, and venue. They were given the option of bringing a carer or family member to the focus group although that was not a requirement. A convenience sample of HCPs was recruited from the professional networks of the research team. They were contacted through email or WhatsApp groups and asked to take part. If they were interested, study information was emailed to them with details of the date, time, and venue of the focus group, or if attending via Zoom, a link was shared. Consent was taken from all participants at the start of the focus group.

**Focus Groups**

A focus group schedule was prepared for each of the 2 focus groups including main questions and prompts. For both groups, the questions were related to their experiences of tele rehabilitation, perceived potential benefits of delivering or receiving rehabilitation via tele rehabilitation, perceived potential challenges or difficulties, and suggestions for overcoming these difficulties.

The service user and carer focus group was facilitated by the male research coordinator in the room (SOA). This facilitator had been involved in recruitment of participants, had experience of conducting focus groups, and spoke the local dialect (Twi). The HCP focus group was facilitated by a senior female researcher from the United Kingdom (LP) via Zoom. This facilitator was a physiotherapist, with experience in tele rehabilitation research and facilitating focus groups. She did not know any of the participants except those within the research team. The research team was introduced, and participants were provided with background information on the purpose of the study. The service user focus group lasted approximately 1 hour and the HCP group 1.5 hours. Both were audio-recorded and transcribed verbatim in the language in which they were conducted. The transcript of the service user and carer focus group was translated into English by professional transcribers from the University of Energy Sunyani, Ghana, and checked by the research team for accuracy. In addition, research team members made field notes during and after the focus groups.

**Analysis**

Thematic analysis was the method of data analysis using the 6 phases outlined by Braun and Clark [18], including data familiarization, generation of initial codes, generation of themes, review of themes, definition of themes, and writing the report. Initial coding of the transcripts was completed by one researcher (KT) who had no previous relationship with study participants and checked by a second researcher (LP) who had conducted the HCP focus groups. Themes were derived from the data, and initial themes were presented to the research team (in verbal, written, and diagrammatic form), which were further refined after feedback and discussion. Thematic analysis was undertaken using Microsoft Word (Microsoft Corporation) with documentation shared at each stage ensuring transparent recording of the data analysis process with the research team. Thematic analysis of each focus group took place separately before a final round of analysis integrated these together. Further feedback was provided by the research team before the final themes were confirmed.

**Ethical Considerations**

Ethical approval was received from the Ethics Committee of Kwame Nkrumah University of Science and Technology, Kumasi, in December 2022 (reference CHRPE/AP/822/22), and all participant data were deidentified. Informed consent was provided by all participants. Travel expenses were covered and lunch was provided for participants who attended in person focus groups.

**Results**

**Characteristics of Participants**

A total of 13 service users agreed to take part in the focus group; however, 2 were unable to attend on the day of the focus group, so overall 11 service users (9 male users) took part (Table 1). Service users had a mean (SD) age of 59.8 (8.6) years; 6 lived in an urban location, and 5 in a semiurban location. Six service users had PD, and 5 had had a stroke. The mean (SD) age of those with stroke was 53 (8.3) years, and the mean (SD) time since stroke was 4 (1.6) years. In contrast, the mean (SD) age of those with PD was 65.5 (2.5) years, and they had had the condition for a mean (SD) of 3.8 (2.1) years. In addition, 9 carers took part (4 male carers), and their mean (SD) age was 38.4 (8.6) years. They had a variety of occupations and had been caring for people with PD (n=6) and stroke (n=3) for a mean (SD) of 3.4 (2.2) years (Table 1).

Service users had different prior exposure to tele rehabilitation, both synchronous and asynchronous. One had participated in a previous trial of an app that delivered an individualized rehabilitation program remotely supervised by a therapist [14],...
some had received rehabilitation plans via WhatsApp to undertake without supervision, and others had no previous experience of telerehabilitation.

Table 1. Demographic details of service users (n=11) and carers (n=9).

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Service users (n=11)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Sex, n</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>9</td>
</tr>
<tr>
<td>Female</td>
<td>2</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>59.8 (8.6)</td>
</tr>
<tr>
<td><strong>Diagnosis of service user, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Stroke</td>
<td>6 (55)</td>
</tr>
<tr>
<td>Parkinson disease</td>
<td>5 (45)</td>
</tr>
<tr>
<td>Time since diagnosis (years), mean (SD)</td>
<td>6.0 (3.9)</td>
</tr>
<tr>
<td><strong>Residence, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Urban</td>
<td>6 (55)</td>
</tr>
<tr>
<td>Semiurban</td>
<td>5 (45)</td>
</tr>
<tr>
<td>Rural</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>Carer or significant other (n=9)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Sex, n</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>4</td>
</tr>
<tr>
<td>Female</td>
<td>5</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>34.8 (8.6)</td>
</tr>
<tr>
<td><strong>Diagnosis of service user, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Stroke</td>
<td>3 (33)</td>
</tr>
<tr>
<td>Parkinson disease</td>
<td>6 (67)</td>
</tr>
<tr>
<td>Length of care (years), mean (SD)</td>
<td>3.3 (2.2)</td>
</tr>
<tr>
<td><strong>Relationship to service user, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Daughter</td>
<td>4 (44)</td>
</tr>
<tr>
<td>Brother</td>
<td>2 (22)</td>
</tr>
<tr>
<td>Spouse</td>
<td>1 (11)</td>
</tr>
<tr>
<td>Son</td>
<td>1 (11)</td>
</tr>
<tr>
<td>Unknown</td>
<td>1 (11)</td>
</tr>
<tr>
<td><strong>Occupation, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Trader</td>
<td>2 (22)</td>
</tr>
<tr>
<td>Driver</td>
<td>2 (22)</td>
</tr>
<tr>
<td>Hairdresser</td>
<td>1 (11)</td>
</tr>
<tr>
<td>Caterer</td>
<td>1 (11)</td>
</tr>
<tr>
<td>Shoemaker</td>
<td>1 (11)</td>
</tr>
<tr>
<td>Seamstress</td>
<td>1 (11)</td>
</tr>
<tr>
<td>Student</td>
<td>1 (11)</td>
</tr>
<tr>
<td><strong>Residence, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Urban</td>
<td>5 (56)</td>
</tr>
<tr>
<td>Semiurban</td>
<td>4 (44)</td>
</tr>
<tr>
<td>Rural</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>
Eighteen stakeholders took part in the focus groups (10 female and 8 male). In total, 7 speech and language therapists, 3 physiotherapists, 3 occupational therapists, 2 neurologists, and 1 specialist physician, 1 nurse, and 1 person from industry participated. Eleven of the participants were aged 30-39 years, 3 were aged 20-29 years, and 3 were aged over 40 years. They had been in their current post for a mean of 4.8 (SD 3.3) years, and all worked in urban locations, most in the 2 main cities of Accra and Kumasi.

Analysis of the focus groups resulted in the following themes and subthemes (Figure 1).

**Rehabilitation Services**

Service users identified different ways of accessing rehabilitation services, in-person services, telerehabilitation, and hybrid (a combination of in-person and remote services). Service users also expressed a need for increased awareness of the differing ways to access rehabilitation services and a need for reliable health information.

**In-Person Rehabilitation**

Service users valued in-person rehabilitation for the purpose of monitoring their rehabilitation progress and receiving feedback on this.

*If there is an improvement, the doctor or the health facility that will tell you to either reduce or increase some aspect of your rehabilitation so that the body will recover.* [Service user]

It was however acknowledged that in-person rehabilitation had limitations particularly when there was a large demand for...
services, which could result in long queues at the health care facility, or the feedback received was not as hoped.

I stopped the rehab because of waking up early to join a queue. [Service user]

One physio teasingly told me that my rehab has taken too long which downed my emotion and that made me stop coming for physio. [Service user]

Hybrid Rehabilitation

Some service users and carers expressed a preference for a hybrid model of rehabilitation (in-person and telerehabilitation), which they had previously found useful.

I feel like if we combine the two it will help, both should go hand in hand. [Service user and carer]

Both, because sometimes you may miss or do certain procedures wrongly without knowing but going to the hospital you will be able to know the exact procedure, then you can go ahead and practice at home. [Service user]

Access to Health Information

A need for access to reliable health information was highlighted by both patients and carers with the potential for technology to be used as means of distributing information. Concern was expressed at misinformation, which hindered treatment such as medication use.

There should be associations to broadcast to people and to serve as guidance for others to follow. [Service user and carer]

I have seen that everything is going on well here, but the information given is very low and because of that most people depend on herbal medicines. [Service user and carer]

Telerehabilitation

Service users and carers who had experienced telerehabilitation found this to be a helpful way to access services and were happy to recommend this method of service delivery to others.

Yes, (I would recommend it) because I have tried and tested and it really helped me. [Service user and carer]

We the caregivers also go through the same stress as the patients, so our experiences shows that the tele-rehab is very good. [Carer]

For others, this was a novel way to access rehabilitation services, which, they thought, would be of benefit, indicating that they would be happy to try this in the future.

Gives me hope that even if a person gets any neurological disorder, the telerehab is there to help them get back their strength. [Service user and carer]

However, there was a lack of awareness of telerehabilitation with service users, suggesting that further awareness raising was needed as well as ensuring that access was available throughout Ghana.

Many Ghanaians are not aware of the telerehab and so it should be publicised through TV and other media groups. [Service user and carer]

Telerehabilitation should be extended to clinics to enhance easy access. It should not be in the big hospitals alone. [Service user and carer]

A range of telerehabilitation services were described by service users who had experienced stroke (including those with aphasia), those with PD, and those with tinnitus or balance issues. These services included checking rehabilitation progress or receiving speech and language therapy sessions. In addition, service users sought further follow-up and reminders through telerehabilitation.

HCPs also had a range of experiences using telerehabilitation ranging from no involvement to using telerehabilitation for a range of rehabilitation purposes including teaching exercises, promoting engagement in activities, conducting hearing assessments, sending intervention messages and information, balance re-education, and reviewing videos to direct parent-led rehabilitation.

Benefits of Telerehabilitation

Reduction in Travel, Time, and Associated Costs

A range of benefits were identified by service users, carers, and HCPs, including a reduction in travel, which made rehabilitation more convenient, easier to access, less stressful, accommodated other caring responsibilities, and was more affordable. Less time was also spent waiting for rehabilitation once patients and carers arrived at the rehabilitation unit.

It helped because having her come all the way to the clinic and the child who didn’t sit and fussy two hours was also time that was solved with tele-therapy. [HCP]

It reduces financial costs to the hospital and the stress involved in sitting in a trotro (car). [Service user and carer]

Increased Access to Rehabilitation

HCPs identified that they were able to access service users who lived some distance away (including visual access to the home environment) increasing access and relevance of the rehabilitation services delivered. This provided the opportunity for further rehabilitation, increasing the intensity, consistency, and adherence to rehabilitation offered.

with the in-person you’re scheduled to come let’s say two or three days a week but with telerehabilitation, even if you want to do it every day as far as the resources are there, you can do it every day. It also allows you to do it more frequently and over a more sustained period of time. That’s one of the benefits. [HCP]

It can help with intensity because the more we meet them online we can achieve our goals. It is easy to find your therapy with more therapy sessions without having to move from one location to another. You sit in the comfort of your home and have more therapy...
sessions within the scheduled period but with a cost-effective system. [HCP]

HCPs also saw the benefits of using devices such as mobile phones as a repository for information. This provided service users with a reminder of their rehabilitation exercises.

Most patients forget the therapy exercise we do for them so probably videoing it and saving for patients can help them do it at their leisure time. [HCP]

HCPs reported that developing new ways of delivering rehabilitation services encouraged them to be creative in their approach including delivering content in a range of local dialects with the potential of developing new rehabilitation services or using technology to measure changes in service user knowledge for example.

I would like to add it makes you creative as a professional because you look for other means of making it fun with telerehabilitation in order to suit your client. It brings out creativity in you. [HCP]

I think it will give the opportunity to assess the clients’ environment for recommendations for possible adaptations to enhance function because when the person comes in-person, they may describe the home environment and we don’t know exactly how it looks but through telerehab with video conference or picture you can see how the environment really looks like for recommendations for the adaptations. [HCP]

Rehabilitation Progress Using Telerehabilitation

Service users highlighted the progress that they had made with their rehabilitation delivered using telerehabilitation, which included improvements in speech, arm movements, activity levels, and independence in activities.

It is through the tele-rehab that I am able to lift my hand today. It really helped me. [Service user]

I’m forever grateful for the physios because up till date, they still call me and they helped in so many ways. [Service user]

Challenges of Telerehabilitation

Challenges of delivering telerehabilitation were identified by service users, carers, and HCPs including access to equipment and availability of a reliable internet connection, the cost of data packages, and challenges to rehabilitation delivery.

Equipment and Network Access

Availability of equipment and the requirement of a reliable network needed for telerehabilitation were highlighted as a barrier. Most service users in the focus group had access to mobile phones (including smartphones that access the internet), but not all did. Challenges with unreliable internet connections meant therapists often prepared alternatives such as printed exercise sheets, sending SMS text messages to promote engagement, watching videos offline, use of images (rather than videos), or switching platforms seeking a better connection. At times, however, the connection was so poor that therapy sessions were abandoned and rescheduled or therapists advised returning to in-person rehabilitation. For some service users, this meant finding other solutions including referral to local therapists to access rehabilitation.

Only 35% of our patients said that they own a smartphone and even of that 35% it’s not necessarily their own but there is a smartphone in the house. It’s not as if the smartphone belongs to them. [HCP]

The network decided to fail all of us. When we call again, we can’t hear anything. We used other platforms FaceTime, other things, but the network just wouldn’t channel us until we had to put the whole therapy to an ending try to refer to someone closer to him. [HCP]

In terms of connectivity, we can record the videos, store them, and send them so that when the connectivity improves the person can use the stored videos to be able look back on their exercise. [HCP]

Financial Cost

Those who did have access to smartphones acknowledged that there was a financial burden of purchasing packages or “bundles” for calls and network access. Some rehabilitation services that required use of additional equipment provided this, but not all did.

Most of us are retired so money is hard to come by so it will really help us (if the equipment and data package are provided). [Service user]

Rehabilitation Delivery

Use of telerehabilitation meant that therapists had to find new ways of delivering therapy services for an individual rather than their usual in-person service that included group therapy, acknowledging that not all service users had the digital literacy skills or ability needed to use technology.

because if I am at the physiotherapy department I can supervise maybe 4 or 5 patients simultaneously; this one is doing this, this one is doing that, I can just observe them but in telerehab they have to do one on one supervision so that may also eventually reduce the number of participants or patient they can attend to at a time if it is ongoing supervision they have to do. [HCP]

Sometimes they are not very tech savvy so we need to see them in-person. And we have had some patients try it but most are not tech savvy so we haven’t been able to expand this to all our clients. [HCP]

Another barrier I was thinking about was that if the client has multiple deficits so maybe visual needs and other possible deficits, I think that may impact teletherapy. [HCP]

Other challenges with delivery of rehabilitation, such as location or timing of therapy sessions, the move toward carer involvement, and the type of therapy session being delivered, were highlighted.

Mum and I decided to have teletherapy instead [for the child] because they were coming from far. But each time I book them, I give them a time. But each
time I am ready to have the teletherapy the child might be asleep or would be at a place where it is uncomfortable place to have teletherapy. [HCP]

So let me add one more barrier, with the issue of adaptive devices especially if there are no caregivers who really understand or who can be trained on how to retrain the patient on how to use the assistive device it becomes difficult. Then unless the patient comes in-person for you to maybe fabricate or measure their assistive device and then train the patient in how to use it. [HCP]

Implementation of Telerehabilitation

Stage in the Patient Journey

Service users had a range of opinions on what stage in their rehabilitation journey they preferred to receive telerehabilitation, from the initial stages of diagnosis to using telerehabilitation following a period of in-person rehabilitation.

Initial stages of my condition. [Service user]

At the initial stages, you should visit the hospital for physio then later, join the telerehab. [Service user and carer]

Platform of Choice

WhatsApp was the preferred platform although Microsoft Teams, Zoom, and FaceTime were also used. Ease of use was important with WhatsApp reported to be familiar with options for low data consumption.

Some find zoom cumbersome it’s difficult for them to manage their way through zoom and then having meetings with them, but WhatsApp is just like having a call. I just call you; you see me on the video and then whatever we need to, it’s easier using that platform compared to the other platforms. I used WhatsApp because I think it is easier over here and with data consumption, you know it has the option for you to select low data mode where the streaming is easier for the patient. [HCP]

Information Governance

Integrating use of personal technology (such as mobile phones) into therapy did cause concern for information security particularly when sharing media such as videos.

I have been thinking about ever since I started working with the clients that I work with. Recently I lost the password to my laptop, and I was thinking there might be a case where a third party has to come in. My phone gets called or any other thing, it’s just confidentiality how is it handled? because a third party will have to come in and help me unlock or do something to my phone Videos, I have a lot of videos of the same person doing mostly I delete them, but I have been thinking about situations where I might forget, or something would happen that another person had access to the videos. So sometimes you know a client may feel uncomfortable sharing videos across, because they don’t know where that would end so that’s one of the barriers I think. [HCP]

Instructions for Use

Participants felt that further instructions or information was needed to take full advantage of rehabilitation delivered via technology with access to videos requested.

I only had to follow the instructions on the phone. [Service user and carer]

I feel like the physiotherapists should be involved in the video demonstrations. There should be visual demonstration videos so the patient can see the physio demonstrating the exercise and imitate it correctly. [Service user and carer]

Cultural Relevance

It was highlighted that many of the current telerehabilitation exercise videos featured White individuals with instructions in English. To increase the cultural relevance of the materials, videos should be in a local dialect featuring Black individuals completing exercises.

It should be conducted in our local dialect Twi to be specific. [Service user and carer]

Language shouldn’t be an issue if we stick to the local dialect. But using English it could be a challenge...another one has to do with the videos where a white person is involved, it makes it difficult for the most clients to understand but it can solve by introducing blacks in such videos. [HCP]

Discussion

Principal Results

Participants from both focus groups were overall positive in terms of telerehabilitation for people with neurological conditions in Ghana but identified a number of challenges. The conceptual framework for sustainable eHealth in resource-limited countries proposed by Fanta and Pretorius [19], comprising technological, social, economic, and organizational factors, will be used to discuss the findings of the focus groups.

In terms of technological factors, one of the main challenges was the lack of a stable internet connection to conduct a telerehabilitation session. This has been identified as one of the main barriers to successful implementation of telerehabilitation in many others studies in LMICs [4,14,15,20], especially in relation to the transfer of images or videos. There was also an appreciation that not all service users had access to a smartphone, as previously reported [5,16]. Interestingly, only 1 therapist raised concerns about the security of service user data particularly when videos of service users completing tasks are sent. In terms of the technology used, this is one of the first studies to report that WhatsApp was the preferred platform for delivering telerehabilitation as it is commonly used in Ghana by both HCPs and service users, so people are familiar with its use and it also has an option for a low data mode. Previous studies of telerehabilitation in LMICs have provided patients
with videos to play in a video player [21] or have used Skype [15].

Social factors were most commonly raised within the focus groups. Participants overall had positive views on telerehabilitation but also discussed the advantages or preferences of in-person rehabilitation especially with regard to the HCP monitoring their progress. Previous research particularly in relation to patients receiving speech and language therapy also reported good satisfaction with telerehabilitation interventions, but many preferred in-person therapy where there was better eye contact between the patient and the therapist and it was easier to understand facial expression [15,22].

Participants in this study however recognized that there were many barriers to service users accessing in-person rehabilitation. Service users also highlighted the lack of awareness of telerehabilitation services in Ghana.

Like previous studies of telerehabilitation generally [11,13,20] and specifically in relation to LMICs [3,23], participants reported a number of benefits to telerehabilitation especially the convenience, requiring less travel time and reduced travel cost. Some HCPs raised that telerehabilitation improved access to rehabilitation services for patients, especially those who stayed a distance from the clinic. In Ghana, up to 43% of people after stroke access herbal medicines [24], as was raised by 1 person in the focus groups, for which there is little or no efficacy data. Improving access to rehabilitation may reduce the reliance on herbal medicine for some patients.

However, there were also some negative aspects. Not all service users had a smartphone, and furthermore, low levels of literacy and digital skills were barriers to implementing telerehabilitation in line with previous papers [3,4,25]. Although there were examples of some apps being used by therapists, they were not felt to be culturally relevant for the Ghanaian context as they were in English rather than the local dialect and tended to have White people (generally Americans or Europeans) demonstrating the activity, which had the potential to reduce engagement of service users. Odetunde et al [21] developed a telerehabilitation video solution for patients with stroke in Nigeria, delivered both in the local language Yoruba and in English, and this was positively received by participants. To promote uptake and adherence, future development of telerehabilitation interventions should consider the local language and other contextual issues [4,10].

In terms of economic factors, as discussed above, not all service users had access to a smartphone, and the financial implications of requiring such a phone and data package were a barrier for many people. Sarfo et al [16] reported that only 35% of their respondents owned a smartphone although 80% had a family member who did. A 2020 household survey of information and communication technology use in Ghana reported that 47.9% of the population had a basic phone and 46.1% had a smartphone; however, in rural areas, 61.3% of people had a basic phone and only 28.1% had a smartphone [26]. However, the figures suggest that family or carers may need to be actively involved for telerehabilitation to be implemented, especially in rural areas of Ghana and in other LMICs. On the positive side, telerehabilitation reduced costs and time required to travel to in-person appointments. In terms of clinical effectiveness, telerehabilitation is variable with some evidence in support of telerehabilitation and other evidence suggesting that it is not superior to conventional care [13,27]. Although there are associated costs, even if it is not superior, the convenience, reach of services, and time-saving aspects would support its further development; however, cost-effectiveness analyses of telerehabilitation for neurological conditions generally and specifically related to resource-limited settings are required [6,9].

Organizational factors were generally related to the delivery of the telerehabilitation interventions. Telerehabilitation was felt to be a positive development that could address long waiting times and high demand on services. A number of service users with stroke and PD had used telerehabilitation with examples of telerehabilitation delivery across different allied HCP groups: physiotherapy, occupational therapy, and speech and language therapy. Service users however felt that they would have liked more instructions on how to use telerehabilitation.

HCPs reported using telerehabilitation for undertaking patient assessments, sending information to service users, and receiving videos of patient progress. Neurological rehabilitation often involves intensive therapy with highly repetitive, task-specific exercise to optimize neuroplastic changes in the central nervous system [12]. A novel finding of this study was that therapists reported that telerehabilitation, in this resource-limited setting, allowed more intensive and consistent therapy than would be possible in person and which importantly facilitated increased adherence and improved outcomes. An additional novel finding was that therapists felt that new ways of delivering therapy encouraged them to be more creative in their approach.

The negative aspects considered under organizational factors were the location and timing of telerehabilitation sessions that were arranged in advance; however, when the appointment time came, the service user was not available, or it was not appropriate to complete the session. This often meant that therapists had to have alternative plans in place should that occur. There was also an appreciation that some activities, such as providing assistive devices, required to be done in person.

Strengths and Limitations
This research had a number of strengths. The views of service users and carers, HCPs, and other stakeholders on telerehabilitation in Ghana were sought directly. Aljabri et al [5] recommended that future research should explore the views of HCPs from different disciplines, which we did, including a range of HCPs such as occupational therapy and speech and language therapy, professions seldom included in the telerehabilitation literature in LMICs—perhaps due to their relatively small numbers compared with, for example, physiotherapy. This is also the first study to include participants with PD from LMICs, although there are previous reviews of telerehabilitation in PD but not in an LMIC context [27]. To be as inclusive as possible, the service user and carer focus group was conducted in the local language. The use of teleconferencing for the stakeholder focus group allowed a wide geographical spread of participants from across Ghana. However, this research also had a number of limitations. Although traveling expenses
and refreshments were provided, service users and their carers had to be able to travel to the hospital to take part in the focus group, possibly biasing the sample to a local, urban dwelling and less disabled group. Also, many had had experience of using telerehabilitation, so perhaps they did not represent the views of most people with neurological conditions in Ghana; however, it was important that they were able to share their experiences. Also, none of the study participants resided in a rural setting, thus limiting the transferability of the views captured in this study.

Conclusions
This is the first study to elicit the views of service users, carers, and HCPs of telerehabilitation for people with neurological conditions in a resource-limited setting of Ghana. The focus group findings overall demonstrated that service users, carers, and HCPs had positive views and experiences of telerehabilitation, especially the convenience and lower cost for service users and the consistency and higher intensity of therapy possible, with some negative aspects including lack of a stable internet connection, cost of phones and data packages, and low levels of literacy. Overall, the findings suggest the need for future research of the clinical and cost-effectiveness of lost cost telerehabilitation interventions for people with neurological conditions, taking into account the local context in Ghana and other LMICs. Telerehabilitation in Ghana is currently not covered within the National Insurance system; however, these findings support the development of telerehabilitation in Ghana with suggestions for future implementation and scale.

Acknowledgments
We would like to thank all the participants for their active contribution and professional transcribers from the University of Energy Sunyani Ghana transcribing the focus group recordings. This work was supported by the Academy of Medical Sciences Global Challenges Research Network Grant (Ref GCRFNGR51278).

Conflicts of Interest
None declared.

References


**Abbreviations**

- **HCP**: health care professional
- **LMIC**: low- to middle-income country
- **PD**: Parkinson disease

Edited by L Buis; submitted 31.05.23; peer-reviewed by E Nizeyimana, S Quaglini; comments to author 18.12.23; revised version received 29.01.24; accepted 20.02.24; published 27.03.24.

Please cite as:

Paul L, Thomson K, Asibey SO, Brady M, van Wijck F, Antwi D, Opoku EN, Sarfo FS

Views of Service Users, Their Family or Carers, and Health Care Professionals on Telerehabilitation for People With Neurological Conditions in Ghana: Qualitative Study

JMIR Mhealth Uhealth 2024;12:e49501

URL: https://mhealth.jmir.org/2024/1/e49501
doi:10.2196/49501
PMID:38536218
Reflective Engagement With a Digital Physical Activity Intervention Among People Living With and Beyond Breast Cancer: Mixed Methods Study

Michael C Robertson¹,², MPH, PhD; Emily Cox-Martin³, PhD; Karen Basen-Engquist⁴, MPH, PhD; Elizabeth J Lyons², MPH, PhD

¹Department of Family and Preventive Medicine, TSET Health Promotion Research Center, University of Oklahoma Health Sciences Center, Oklahoma City, OK, United States
²Department of Nutrition, Metabolism & Rehabilitation Sciences, University of Texas Medical Branch at Galveston, Galveston, TX, United States
³VA Puget Sound Health Care System, Tacoma, WA, United States
⁴Department of Health Disparities Research, University of Texas MD Anderson Cancer Center, Houston, TX, United States

Corresponding Author:
Michael C Robertson, MPH, PhD
Department of Family and Preventive Medicine
TSET Health Promotion Research Center
University of Oklahoma Health Sciences Center
655 Research Pkwy #400
Oklahoma City, OK, 73104
United States
Phone: 1 405 271 6872
Email: michael-robertson@ouhsc.edu

Abstract

Background: People living with and beyond breast cancer can face internal barriers to physical activity (eg, fatigue and pain). Digital interventions that promote psychological acceptance and motivation may help this population navigate these barriers. The degree to which individuals (1) adhere to intervention protocols and (2) reflect on and internalize intervention content may predict intervention efficacy.

Objective: The objective of this study was to characterize the nature of reflective processes brought about by an 8-week acceptance- and mindfulness-based physical activity intervention for insufficiently active survivors of breast cancer (n=75). Furthermore, we explored the potential utility of a metric of reflective processes for predicting study outcomes.

Methods: Of the intervention’s 8 weekly modules, 7 (88%) included an item that asked participants to reflect on what they found to be most useful. Two coders conducted directed content analysis on participants’ written responses. They assessed each comment’s depth of reflection using an existing framework (ranging from 0 to 4, with 0=simple description and 4=fundamental change with consideration of social and ethical issues). The coders identified themes within the various levels of reflection. We fit multiple linear regression models to evaluate whether participants’ (1) intervention adherence (ie, number of modules completed) and (2) the mean level of the depth of reflection predicted study outcomes.

Results: Participants were aged on average 57.2 (SD 11.2) years, mostly non-Hispanic White (58/75, 77%), and mostly overweight or obese (54/75, 72%). Of the 407 responses to the item prompting personal reflection, 70 (17.2%) were rated as reflection level 0 (ie, description), 247 (60.7%) were level 1 (ie, reflective description), 74 (18.2%) were level 2 (ie, dialogic reflection), 14 (3.4%) were level 3 (ie, transformative reflection), and 2 (0.5%) were level 4 (ie, critical reflection). Lower levels of reflection were characterized by the acquisition of knowledge or expressing intentions. Higher levels were characterized by personal insight, commentary on behavior change processes, and a change of perspective. Intervention adherence was associated with increases in self-reported weekly bouts of muscle-strengthening exercise (B=0.26, SE 0.12, 95% CI 0.02-0.50) and decreases in sleep disturbance (B=−1.04, SE 0.50, 95% CI −0.06 to −2.02). The mean level of reflection was associated with increases in psychological acceptance (B=3.42, SE 1.70, 95% CI 0.09-6.75) and motivation for physical activity (ie, integrated regulation: B=0.55, SE 0.25, 95% CI 0.06-1.04).

Conclusions: We identified a useful method for understanding the reflective processes that can occur during digital behavior change interventions serving people living with and beyond breast cancer. Intervention adherence and the depth of reflection...
each predicted changes in study outcomes. Deeper reflection on intervention content was associated with beneficial changes in the determinants of sustained behavior change. More research is needed to investigate the relations among digital behavior change intervention use, psychological processes, and intervention efficacy.

(JMIR Mhealth Uhealth 2024;12:e51057) doi:10.2196/51057

KEYWORDS

survivors of cancer; exercise; acceptance and commitment therapy; fatigue; mindfulness; motivation; behavioral sciences

Introduction

Background

There were an estimated 4.1 million people living with and beyond breast cancer in 2022 [1]. This number is estimated to reach nearly 5 million by 2030 [2]. This population can encounter long-term challenges related to health and quality of life, including fatigue, anxiety and depressive symptoms, breast cancer–related lymphedema, metabolic dysregulation, bone loss and osteoporosis, and cancer recurrence [3,4]. Physical activity may protect people living with and beyond breast cancer from these problems [5-7]. However, most people who have been diagnosed with breast cancer do not meet nationally recommended physical activity guidelines [8,9].

Psychotherapy-informed digitally delivered acceptance- and mindfulness-based approaches may facilitate physical activity promotion for people living with and beyond breast cancer. People living with and beyond breast cancer commonly cite uncertainty, frustration, cancer-related fatigue, and pain to be barriers to physical activity [10-13]. Interventions that can help people to better navigate the uncomfortable thoughts and sensations that can act as impediments to physical activity may support their efforts to be more physically active. Authoritative entities recommend acceptance- and mindfulness-based interventions for people living with and beyond cancer; strong evidence supports their efficacy for reducing anxiety and depressive symptoms in this population, and they have been shown to reduce cancer-related fatigue [14-17]. Furthermore, physical activity interventions that simultaneously target both health-related behaviors and quality-of-life issues are more effective at achieving sustained physical activity outcomes than interventions that only promote physical activity for this population [18]. Emerging evidence suggests that acceptance- and mindfulness-based physical activity interventions can be feasible and effective at promoting physical activity [19-21] and that digitally delivered approaches can be acceptable and potentially effective for people living with and beyond cancer [22,23].

The degree to which participants meaningfully engage with and internalize health promotion content may determine the efficacy of digital behavior change interventions (DBCIs). However, achieving high levels of engagement can be a marked challenge for DBCIs [24]. Participants often exhibit low adherence to digital physical activity interventions and poor study retention [25]. Accordingly, researchers have tended to prioritize measuring and optimizing intervention use (eg, optimizing metrics such as the number of modules completed and time spent in an app) [26]. However, using system use data alone as a proxy for engagement is problematic. First, the relationship between these metrics and intervention efficacy is not always straightforward. Although there is sometimes a positive linear relationship between DBCI use and intervention efficacy, this is not always the case [25,27,28]. DBCIs may tend to have certain thresholds of engagement that confer the majority of the benefits, or intervention components may interact with one another. Furthermore, engagement with a DBCI may be expected to taper over time and even be supplanted by engaging in the desired behavior itself [29]. Second, focusing exclusively on quantitative use metrics affords only limited insight into the psychological aspects underpinning participants’ experiences with DBCIs. This precludes achieving a full understanding of the DBCI’s mechanisms of action. Thus, researchers have called for broadening our conceptualization of DBCI engagement [28].

A growing literature highlights the importance of understanding the psychological processes that occur when participants engage with DBCI content; this includes supplementing system use metrics with additional aspects of DBCI engagement that pertain to affect, attention, interest, immersion, flow, and reflection [26,29,30]. Researchers of human-computer interaction have discussed engagement in similar terms, parsing engagement according to notions of behavioral adherence, behavioral effort (eg, discussing emotions, thoughts, and behaviors in intervention), cognition, and affect [31]. Researchers with an educational perspective have provided similar definitions of engagement with digital systems, noting that the cognitive and affective aspects of engagement are largely neglected [32]. Yang et al [25] have contributed a framework specific to physical activity promotion in mobile health in which they conceptualize engagement as being determined by breadth, depth, interaction, and length of engagement. These broad conceptualizations are useful for helping to orient researchers toward the many and varied facets of engagement. Rather than attempting to establish a universal conceptualization, it may be advantageous to tailor the conceptualizations of engagement based on context [33]. Using qualitative and mixed methods to investigate the nature of participants’ engagement with DBCIs may be particularly useful to this end [24,26,29].

Reflection is an essential process for self-improvement and making lasting behavior changes in the context of DBCIs [34]. As applied to health-related education, reflection and reflective processes connotate deliberate critical analysis of knowledge and experiences to achieve a fuller understanding [35]. The process of reflecting on the outcomes of past and ongoing efforts or newly acquired information helps individuals integrate knowledge and skills into practice and overcome persistent barriers to behavior change [35,36]. Reflection may be particularly relevant in the context of DBCIs because it may allow people to gain meaningful insights from their personal...
health-related data that can support lasting behavior change [34]. In the education literature, reflective thinking and writing are commonly used strategies to prompt personal reflection, and it is a common practice to evaluate written responses to determine to what degree an individual has considered and applied didactic content in the context of their own lived experiences [37-39]. Fleck and Fitzpatrick [40] present a framework that operationally defines 5 levels of reflection. This framework has been shown to be particularly useful for evaluating the levels of reflection promoted by the design features of DBCIs [34]. In this study, we sought to extend this literature by characterizing the reflective processes engaged in by people living with and beyond cancer who experienced a digital acceptance- and mindfulness-based physical activity intervention.

Objectives

It may be possible to help people living with and beyond breast cancer reframe some of the unpleasant internal sensations that can act as barriers to physical activity (eg, pain, fatigue, and frustration). Understanding the degree to which participants reflect on and internalize digital intervention content targeting these psychological processes may be useful for understanding how best to support this population. This study answers calls to investigate the deeper psychological aspects involved in engagement with DBCIs to provide insight into the interplay among intervention use, individual experiences, and intervention efficacy. The aims of this study were to (1) characterize the nature of reflective processes brought about by a DBCI designed to increase physical activity in insufficiently active people living with and beyond breast cancer and (2) explore the potential utility of a metric of reflective processes in DBCIs within an acceptance- and mindfulness-based physical activity intervention.

Methods

Study Design

We conducted a secondary analysis on data obtained from a 1-group pilot study. The purpose of the parent study was to evaluate the acceptability of an acceptance- and mindfulness-based intervention to increase physical activity in survivors of breast cancer [23]. This study is an investigation of the use of the intervention. Participants were female adults (aged ≥18 y) who had been diagnosed with breast cancer but were not actively preparing for surgery or undergoing chemotherapy or irradiation treatment (n=75). Eligibility criteria included that participants reported engaging in <150 minutes of moderate-intensity aerobic physical activity per week. We recruited participants using a large listserve of individuals who were interested in receiving information about breast cancer reframe some of the unpleasant internal sensations that can act as barriers to physical activity (eg, pain, fatigue, and frustration). Understanding the degree to which participants reflect on and internalize digital intervention content targeting these psychological processes may be useful for understanding how best to support this population. This study answers calls to investigate the deeper psychological aspects involved in engagement with DBCIs to provide insight into the interplay among intervention use, individual experiences, and intervention efficacy. The aims of this study were to (1) characterize the nature of reflective processes brought about by a DBCI designed to increase physical activity in insufficiently active people living with and beyond breast cancer and (2) explore the potential utility of a metric of reflective processes in DBCIs within an acceptance- and mindfulness-based physical activity intervention.

Intervention Description

The intervention has been described in detail elsewhere [23]. Briefly, the ACTive program was an 8-week DBCI designed to help insufficiently active survivors of breast cancer increase moderate-intensity aerobic and muscle-strengthening physical activity. It was centered on increasing physical activity acceptance (ie, cognitive acceptance and behavioral commitment) and autonomous motivation for physical activity (ie, enjoyment, values, interest, and identification). It was grounded in acceptance and commitment therapy (ACT) principles (ie, values, committed action, acceptance, defusion, and contacting the present moment). It consisted of 8 modules (along with a brief introductory module), each of which was delivered electronically via a weekly REDCap (Research Electronic Data Capture; Vanderbilt University) survey. The modules included didactic content and experiential exercises targeting core ACT principles. They were presented in brief videos, audio files, images, and other documents. Participants were prompted to input information periodically (eg, physical activity levels and reactions and thoughts on experiential exercises). Branching logic within REDCap was used to remind participants of their responses to items from previous modules as well as provide feedback on their responses within modules (eg, an emoji appeared if responses indicated that the participant had met her physical activity goal for the week). The modules contained a repository of additional optional content (eg, muscle-strengthening physical activity videos for survivors of cancer, videos with yoga classes for survivors of breast cancer, and a video on proper walking posture for survivors of cancer).

At the end of modules 1 to 7, an item asked participants, “What is one important, personal take-away point from this session?” This item was included to identify specific aspects of the intervention that were perceived as the most useful by participants because usefulness is an important component of acceptability. Participants were presented with an open-text box to provide a reply. Providing participants with the opportunity to reply to open-ended items such as this can help contribute to a more in-depth understanding of participants’ engagement and experience with DBCI content [26].

Measures

Physical Activity Behaviors

The Godin Leisure-Time Exercise Questionnaire was administered to obtain pre- and postintervention estimates of participants’ average weekly leisure time aerobic physical activity over the past month [41,42]. To obtain estimates of participants’ muscle-strengthening physical activity levels, we included an item derived from the Godin Leisure-Time Exercise Questionnaire [43,44]. This item asked, “In a typical week, outside of your job or work around the house, how many days do you do leisure-time physical activities specifically designed to strengthen your muscles such as lifting weights, circuit training, or resistance bands? (Do not include cardio/aerobic types of exercise).” Response options ranged from 0 to 7.

Physical Activity Acceptance

The ACTive program was centered on increasing physical activity acceptance. This was defined as one’s willingness to
experience the negative internal experiences that can sometimes be associated with physical activity, rather than avoiding them. This was operationalized by the Physical Activity Acceptance Questionnaire (PAAQ) [45]. The PAAQ consists of two 5-item subscales. The cognitive acceptance subscale measures one’s propensity to accept the reality of unpleasant sensations associated with physical activity, whereas the behavioral commitment subscale pertains to persisting in committed action despite the occurrence of challenging thoughts or sensations. Responses on items of the PAAQ range from 1=never true to 7=always true. For scoring, the items on the cognitive acceptance subscale are reverse coded, and the items of each subscale are summed (range 5-35). This questionnaire was administered before and after the intervention.

Motivation for Physical Activity

The ACTive program also aimed to increase motivation for physical activity as conceptualized by self-determination theory (SDT). SDT parses motivation conceptually based on the degree to which it is autonomous in nature. SDT posits that changes in more autonomous motivations (eg, integrating a behavior into relevant self-narratives) will yield longer-lasting behavior changes than changes in less autonomous motivations (eg, receiving a badge as reinforcement for performance). We operationalized motivation for physical activity before and after the intervention using the Behavioral Regulation for Exercise Questionnaire-3 (BREQ-3). The BREQ-3 consists of six 4-item subscales (amotivation, external regulation, introjected regulation, identified regulation, integrated regulation, and intrinsic regulation). The intrinsic regulation subscale captures a highly autonomous form of motivation, defined by the degree to which one engages in a behavior because one finds it inherently interesting or enjoyable. The integrated regulation subscale captures another highly autonomous form of motivation, an intrinsic form of motivation defined by the degree to which an individual has fully internalized the reason for action owing to finding it concordant with their values. The identified regulation subscale captures a somewhat less autonomous form of motivation (although still relatively autonomous overall), defined by the degree to which an individual consciously values a reason for action. For scoring, the mean scores for each set of items are calculated (range 0-4).

Health-Related Outcomes

For exploratory purposes, we administered measures of quality of life and physical functioning before and after the intervention. To operationalize these constructs, we used the National Institutes of Health–funded Patient-Reported Outcomes Measurement Information System (PROMIS)-29 profile measure (version 2.1) [42]. The PROMIS-29 consists of 8 subscales: physical function (eg, “Are you able to run errands and shop?”), anxiety (eg, “In the past 7 days...I felt fearful”), depressive symptoms (eg, “In the past 7 days...I felt worthless”), fatigue (eg, “In the past 7 days...I felt fatigued”), sleep disturbance (eg, “In the past 7 days...I had a problem with my sleep”), ability to participate in social roles and activities (eg, “I have trouble doing all of my regular leisure activities with others”), pain interference (eg, “In the past 7 days...How much did pain interfere with your day to day activities?”), and pain intensity (eg, “How would you rate your pain on average?”). All subscales except the pain intensity subscale have 4 items and 5-point Likert-type responses ranging from 0 (not at all) to 4 (very much). The pain intensity subscale has 1 item and an 11-point scale ranging from 0 (no pain) to 10 (worst pain imaginable). Scores are coded, summed, and converted to T-scores such that higher scores indicate more of the concept being measured (eg, range for physical functioning subscale: 22.5-57.0).

Reflection Framework

Reflection levels were based on the reflection framework presented by Fleck and Fitzpatrick [40]. This framework emerged from research conducted in the context of the design of digital technologies and human-computer interaction. It defines 5 levels of reflection intended to serve as a resource for thinking about, and designing for, reflection. Reflection level 0 (revisiting) is defined as “Description or statement about events without further elaboration or explanation. Not reflective.” Reflection level 1 (reflective description) is defined as “Description including justification or reasons for action or interpretation, but in a reporting or descriptive way. No alternate explanations explored, limited analysis and no change of perspective.” Reflection level 2 (dialogic reflection: exploring relationships) is defined as “A different level of thinking about intervention content. Identifying or exploring relationships between relevant concepts. Applying experience or knowledge, providing evidence of cycles of interpreting and questioning, consideration of different explanations, hypotheses and/or other points of view.” Reflection level 3 (transformative reflection: fundamental change) is defined as “Revisiting an event or knowledge with intent to re-organize and/or do something differently. Asking fundamental questions and challenging personal assumptions leading to a change in practice or understanding.” Finally, reflection level 4 (critical reflection: wider implications) is defined as “Social and ethical issues are taken into consideration. Generally considering the (much wider) picture.”

Data Analysis

Qualitative Data Analysis

MCR and EJL conducted 2 phases of directed content analysis on participants’ written responses to the open-ended item asking, “What is one important, personal take-away point from this session?” [46]. First, the coders independently rated the reflection level of each individual response by evaluating the response against the reflection framework presented by Fleck and Fitzpatrick [40]. Throughout this process, they produced descriptors to extend and apply the definitions presented by Fleck and Fitzpatrick [40] of the various reflection levels to our study context. The 2 coders first coded all the responses from 1 module independently. Next, they met to discuss the functional definitions and descriptors of reflection levels in our study context and resolve coding discrepancies. The coders then evaluated the rest of the responses independently. After doing so, they met to reconcile discrepancies. The coders created a table, based on the reflection framework presented by Fleck and Fitzpatrick [40] as well as supplemental descriptors and illustrative examples from this study, to further clarify their conception of the various reflective levels as they might be
applied in the context of DBCIs. For the second phase of directed content analysis, MCR sorted responses by reflection level and MCR and EJL coded responses using inductive codes that were informed by principles of behavioral science (eg, behavior change techniques) and ACT [46]. MCR coded all responses first and provided a list of inductive codes to EJL. Next, EJL coded all responses using the list of inductive codes provided by MCR and adding additional codes as needed. The 2 coders then met to discuss codes and reconcile differences.

**Quantitative Data Analysis**

We computed descriptive statistics for participant demographics, study outcome variables, intervention adherence, and each participant’s mean level of reflection (ie, each individual’s mean score of the items scored via the qualitative procedures detailed in the previous subsection). We defined intervention adherence as the number of modules that each participant completed; this is a commonly used measure of engagement with DBCIs [26]. We then conducted multiple linear regression analyses with maximum likelihood estimation. The independent variables were (1) the number of modules completed and (2) participants’ mean level of reflection. The dependent variables (in separate models) were the follow-up measures of self-reported aerobic and muscle-strengthening physical activity, the PAAQ, the BREQ-3, and the PROMIS-29 subscales described previously. All analyses adjusted for the baseline value of the dependent variable and, given our interest in parsing the possible effects of the breadth versus the depth of engagement, the other independent variable of interest (ie, both intervention adherence and the mean level of reflection were included in all models). All analyses also adjusted for sociodemographic and cancer-related factors that we identified a priori as potentially confounding variables. Specifically, we adjusted for age (years), education level (no bachelor’s degree, bachelor’s degree, or graduate school), ethnicity (Hispanic or not Hispanic), race (American Indian or Alaska Native or other, Asian, Black or African American, or White), BMI category (underweight, normal, overweight, or obese), time since cancer diagnosis (years), and stage at diagnosis (1, 2, or 3/4). Missing data were handled using full information maximum likelihood. We set our nominal α level to .05 for all analyses. All statistical analyses were performed using R (version 4.0.3; R Foundation for Statistical Computing).

**Results**

**Participant Demographic Characteristics**

The average age of the participants was 57.2 (SD 11.2; range 31-78) years. The median time since breast cancer diagnosis was 8.2 (IQR 3-12) years. The participants were relatively well educated, mostly non-Hispanic White (58/75, 77%), and mostly overweight or obese (54/75, 72%; Table 1).
Table 1. Participant characteristics.

<table>
<thead>
<tr>
<th>Characteristic and category</th>
<th>Participants, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Education level (n=75)</strong></td>
<td></td>
</tr>
<tr>
<td>High school diploma or GED(^a)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Some college</td>
<td>16 (21)</td>
</tr>
<tr>
<td>Bachelor’s degree</td>
<td>34 (45)</td>
</tr>
<tr>
<td>Graduate school degree</td>
<td>25 (33)</td>
</tr>
<tr>
<td><strong>Stage of breast cancer at diagnosis (n=71)</strong></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>32 (45)</td>
</tr>
<tr>
<td>2</td>
<td>28 (39)</td>
</tr>
<tr>
<td>3</td>
<td>9 (13)</td>
</tr>
<tr>
<td>4</td>
<td>2 (3)</td>
</tr>
<tr>
<td><strong>Race (n=75)</strong></td>
<td></td>
</tr>
<tr>
<td>American Indian, Alaska Native, or other</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Asian</td>
<td>4 (5)</td>
</tr>
<tr>
<td>Black or African American</td>
<td>7 (9)</td>
</tr>
<tr>
<td>White</td>
<td>63 (84)</td>
</tr>
<tr>
<td><strong>Ethnicity (n=74)</strong></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>7 (10)</td>
</tr>
<tr>
<td>Non-Hispanic</td>
<td>67 (91)</td>
</tr>
<tr>
<td><strong>Marital status (n=74)</strong></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>11 (15)</td>
</tr>
<tr>
<td>Married</td>
<td>54 (73)</td>
</tr>
<tr>
<td>Living with significant other</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Divorced</td>
<td>5 (7)</td>
</tr>
<tr>
<td>Widowed</td>
<td>3 (4)</td>
</tr>
<tr>
<td><strong>Employment status (n=68)</strong></td>
<td></td>
</tr>
<tr>
<td>Employed full-time</td>
<td>38 (56)</td>
</tr>
<tr>
<td>Employed part-time</td>
<td>10 (15)</td>
</tr>
<tr>
<td>Retired</td>
<td>20 (29)</td>
</tr>
<tr>
<td><strong>BMI status (n=74)</strong></td>
<td></td>
</tr>
<tr>
<td>Underweight</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Normal</td>
<td>19 (26)</td>
</tr>
<tr>
<td>Overweight</td>
<td>32 (43)</td>
</tr>
<tr>
<td>Obese</td>
<td>22 (30)</td>
</tr>
</tbody>
</table>

\(^a\)GED: General Educational Development Test.

**Intervention Adherence**

The median number of the 8 modules completed was 8 and the first quartile value was 4 modules (IQR 4). The minimum number of modules completed was 1 (7/75, 9%).

**Coding the Levels of Reflection**

The coders created a table while going through the process of coding the reflections based on the reflection framework presented by Fleck and Fitzpatrick [40]. We added descriptors and illustrative examples from this study to further clarify our conception of the various reflective levels as they might be applied in the context of DBCIs (Table 2).
<table>
<thead>
<tr>
<th>Reflection level and additional descriptors</th>
<th>Illustrative examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reflection level 0. Description: revisiting (description or statement about events without further elaboration or explanation; not reflective)</strong></td>
<td></td>
</tr>
<tr>
<td>Repeating or paraphrasing intervention content</td>
<td>“That exercise makes you less tired but also helps you sleep better.” [SID (^b) 157, 47 years]</td>
</tr>
<tr>
<td>Superficial imperative statements</td>
<td>“[G]et up and move.” [SID 95, 49 years]</td>
</tr>
<tr>
<td>Surface-level comments</td>
<td>“I thought the extended mindfulness video was helpful.” [SID 100, 37 years]</td>
</tr>
<tr>
<td>Platitudes</td>
<td>“Life is hard sometimes.” [SID 99, 68 years]</td>
</tr>
<tr>
<td>Not responsive to the prompt</td>
<td>“Bad weather.” [SID 117, 64 years]</td>
</tr>
<tr>
<td><strong>Reflection level 1. Reflective description: revisiting with explanation (description including justification or reasons for action or interpretation but in a reporting or descriptive way; no alternate explanations explored, limited analysis, and no change of perspective)</strong></td>
<td></td>
</tr>
<tr>
<td>Elaborating upon intervention content</td>
<td>“On our sickest days, there is more going right in our bodies than going wrong. I need to appreciate this. And celebrate this.” [SID 113, 50 years]</td>
</tr>
<tr>
<td></td>
<td>“[T]hat the uncomfortable feel of exercise is actually good for me. I just need to embrace it.” [SID 140, 60 years]</td>
</tr>
<tr>
<td>Imperative statements (including justification or reasons or descriptive strategy)</td>
<td>“I need to get moving, so I will feel better.” [SID 79, age not given]</td>
</tr>
<tr>
<td>Personal insight</td>
<td>“[T]hat I’m good at putting things off.” [SID 134, 60 years]</td>
</tr>
<tr>
<td></td>
<td>“I am totally the kid looking out the window—I try to make everything an adventure and to look at the positive mental attitude.” [SID 103, 45 years]</td>
</tr>
<tr>
<td>Skill building or learning a technique</td>
<td>“I’ve learned how to focus away from the chatter in my brain.” [SID 99, 68 years]</td>
</tr>
<tr>
<td></td>
<td>“[T]hat I can ‘pick up’ internal barriers which relieves some of the feeling of frustration and feeling like a failure.” [SID 98, 53 years]</td>
</tr>
<tr>
<td></td>
<td>“I liked the shear stress explanation. I can picture that while I exercise.” [SID 100, 37 years]</td>
</tr>
<tr>
<td><strong>Reflection level 2. Dialogic reflection: exploring relationships (a different level of thinking about intervention content; identifying or exploring relationships between relevant concepts; applying experience or knowledge, providing evidence of cycles of interpreting and questioning, and consideration of different explanations, hypotheses, and other points of view)</strong></td>
<td></td>
</tr>
<tr>
<td>Applying intervention content to one’s own life</td>
<td>“My old excuse of saying ‘I just don’t want to exercise,’ is not a good enough reason. I know the reasons are that it’s uncomfortable and inconvenient, but that’s really not true. Also, I know I can adapt and it won’t cause pain in my hip with bursitis.” [SID 145, 48 years]</td>
</tr>
<tr>
<td>Commentary on the nature of the relationships between disparate concepts</td>
<td>“If I am going to change my fitness habits, I must see how they relate to my values.” [SID 124, 42 years]</td>
</tr>
<tr>
<td>Taking a different perspective</td>
<td>“[T]hat I’m being invited, not required to experiment with activity and that I can choose how to do it. I liked the line about experimenting with what we’re being told the benefits are in our own bodies. That makes me feel more in control and interested.” [SID 137, 53 years]</td>
</tr>
<tr>
<td>Applying new skills or knowledge and reflecting on this</td>
<td>“In the almost 14-minute extended mindfulness exercise, I found that it did relax me even though that wasn’t the objective. I kept my eyes closed during the entire exercise. At first, my other thoughts included anxiety over today’s election. One of the suggestions was to look at yourself from outside yourself. I find I’m usually able to do that anyway...as it enables me to be more compassionate and understanding of others’ pain because of what I’ve experienced. For me, the only distraction, as I kept my eyes closed, was hearing the/your voice telling me things. I had no problem with pushing away thoughts because the only thing I was seeing was inside my eyes as they were closed. Once opened, everything becomes a distraction.” [SID 133, 70 years]</td>
</tr>
</tbody>
</table>

**Reflection level 3. Transformative reflection: fundamental change (revisiting an event or knowledge with intent to reorganize and do something differently; asking fundamental questions and challenging personal assumptions leading to a change in practice or understanding)**
Levels of Reflection

**Overall**

There were 407 total responses from the 75 participants over the course of the 7 modules that featured the open-ended item for reflection (the item was not included in module 8). Participants submitted an average of 5.4 (SD 2.2) out of 7 possible responses. Of the 407 responses, we rated 70 (17.2%) as reflection level 0 (ie, *description*), 247 (60.7%) as level 1 (ie, *reflective description*), 74 (18.2%) as level 2 (ie, *dialogic reflection*), 14 (3.4%) as level 3 (ie, *transformative reflection*), and 2 (0.5%) as level 4 (ie, *critical reflection*).

**Reflection Level 0 (Description: Revisiting)**

Responses that were rated as reflection level 0 were judged to be not reflective and often emphasized one’s desire or intention to increase one’s physical activity levels, simply repeated or paraphrased subject matter content, or provided commentary on the delivery of the subject matter itself. The themes we identified for reflection level 0 responses related to (1) making a resolution, (2) knowledge of subject matter, and (3) appreciation or distaste for intervention content.

Participants often made short imperative statements concerning a general need to increase physical activity or mindfulness (ie, *making a resolution*), but at this level of reflection, they did not demonstrate sufficient explanation to extend their response beyond merely revisiting intervention content (eg, “Just start!” [study identification number (SID) 148, 64 years]).

The theme related to knowledge of subject matter characterized a subset of responses that were descriptive paraphrases of didactic intervention content related to physical activity or ACT principles; for example, a participant wrote, “Even 10 minutes of activity is better than none” (SID 152, 54 years).

Responses that spoke to an appreciation or distaste for intervention content generally provided commentary on the intervention content without evidencing deeper reflection:

- “I am stronger than the spoiled, damaged, hurt child inside me. I don’t have to listen to her. Just put her plump pouty face in my backpack and carry on.” [SID 82, 64 years]

**Reflection Level 1 (Reflective Description: Revisiting With Explanation)**

Responses that were rated as reflection level 1 were judged to provide more justification than reflection level 0 comments, but this additional substance was generally descriptive in nature and without evidence of a deeper change in perspective. This reflection level made up the majority of responses (247/407, 60.7%). These comments commonly emphasized one’s desire or intention to increase one’s physical activity levels with some action-oriented or attitudinal elaboration. The themes we identified for reflection level 1 responses related to (1) making a resolution, (2) personal application or action planning, and (3) kindness.

Making a resolution was also a common theme for the level 1 responses. These responses were generally centered on the importance of increasing physical activity levels and provided a more nuanced explanation than the level 0 responses (eg, “[T]hat fitness in itself is a value that I should prioritize for overall happiness and well-being, not just when I want to lose weight and achieve a summer body” [SID 84, 32 years]).

Responses related to personal application or action planning tended to apply information from the intervention to the participant’s own life. This was often related to engaging in physical activity (eg, “One of the important takeaways for me is the proper way to stand and walk which allows our back to be straight...and assist with our posture” [SID 87, 64 years]). This theme also applied to using acceptance- and mindfulness-based techniques (eg, “Practical Action idea: Pausing to ‘unpack’ the negative thoughts, physically remove them” [SID 110, 74 years]).

Finally, the kindness theme captured responses that included an emphasis on being kind to oneself in the course of gradually
increasing physical activity levels. This theme was also evident in comments that spoke to an appreciation of intervention content that normalized the reality of challenges experienced in the course of engaging in physical activity and cancer survivorship:

Life is hard and challenging and that is NORMAL. A rich full life is one that consists of a variety of emotions. [SID 93, 56 years]

Reflection Level 2 (Dialogic Reflection: Exploring Relationships)

Responses that were rated as reflection level 2 were judged to exhibit reflection on more complex conceptual relationships than responses rated as reflection level 1. These responses often evidenced a change of perspective or a more nuanced consideration of the intervention content and its application to one’s own life. Reflection level 2 responses commonly displayed many of the themes presented previously (eg, making a resolution and personal application or action planning) but were further characterized by the themes of (1) personal insight or (2) discussion of personal facilitators or mechanisms of behavior change.

The theme of personal insight characterized comments that involved elements of personal introspection and reflection on oneself or one’s own thought and behavioral patterns:

It got me to thinking about what the negative thought might be. Why do I not want to exercise if I know it’s good for me? Somewhere way, way deep down I don’t think I believe it. That’s helpful to realize and good to see in print. Now to figure it out. [SID 120, 78 years]

Other participants shared candid observations about the difficulties involved with satisfying perceived needs:

Need to refocus on what has been important, can be again but with physical limitations like a nagging injury, isolation due to the pandemic, restrictions where I live, too comfortable with aloneness now. [SID 139, 76 years]

The theme of discussion of personal facilitators or mechanisms of behavior change applied to responses that evidenced reflection on what might help increase physical activity or how being more physically active might in turn support a valued aspect of the participant’s life:

I need to learn how to separate the thought from who I really am. I realize that I am overwhelmed by the “starting point,” that place where I am required to overcome the law of inertia in all the things I want or need to do. Once I get started I am generally able to complete the task. [SID 98, 53 years]

Reflection Level 3 (Transformative Reflection: Fundamental Change)

This level of reflection was relatively rare (14/407, 3.4%). Responses determined to represent reflection level 3 were characterized by substantively deeper reflection than responses determined to represent reflection level 2—additionally marked by evidence suggestive of more profound personal insight and a fundamental change in perspective. In reflections that attained this level of depth, individuals often intimated that they revised personally held beliefs in ways that were empowering and conducive to enduring change:

I think about “the body keeps the score” and how some of the problems and thoughts and feelings I have are programmed into my DNA just like my propensity towards cancer. Some, but not all. I need to look at those moments where I am able to distract myself and notice the little victories. I’ll never be perfect—never have a perfect body a perfect kid a perfect life, but there are moments every day where I feel I’m doing it right and I need to notice those moments. Last week i ran 2 miles without stopping. Not much—I used to run 5 without stopping—but that was a great feeling. Last week I carved pumpkins with my son and my neighbor came over and we talked for over an hour. Thoughts of being unlovable and depressed faded during those moments. Also I’m trying to stay away from situations where I feel this loneliness and unlovable-ness... staying off the dating site I paid so much for hurts in one way but since nobody was calling me anyways it’s better just to not open the app. [SID 104, 50 years]

Participants evidencing this level of reflection often reframed challenges that related to their cancer journey or spoke about appreciating life from a different perspective:

I love the point just made: “Even if we fail at our goals, if we act in a way that’s consistent with our values we are successful.” I didn’t do as much walking this week, but I did add strength training this week and it’s amazing how much more I felt. And that’s not a typo. I don’t know how else to describe it than I just felt “more.” [SID 113, 50 years]

Reflection Level 4 (Critical Reflection: Wider Implications)

Responses reaching reflection level 4 were very rare (2/407, 0.5%). These responses were judged to meet the level of intrapersonal depth characteristic of reflection level 3 as well as to incorporate some of the wider social and ethical contexts in which the participants’ experiences were situated. The responses spoke to a pressing tension the participants experienced as survivors of cancer in navigating social and community-level factors (1 quote is presented herein, and another quote is presented in Table 2):

Dr. Harris’ video was great to hear. When I was going through chemo some people would say, “Keep a positive attitude” and I didn’t always want to. I wanted to withdraw or rage sometimes. I think it’s easier for others to see you happy, it relieves their tension about what is happening to you. It does not mean you are not living a rich and fulfilling life and I was HAPPY to hear him say that. [SID 120, 78 years]
Intervention Adherence Predicting Outcomes

We observed associations between intervention adherence and change in 2 outcomes of interest (Table 3). Higher intervention adherence was associated with higher weekly bouts of muscle-strengthening physical activity at follow-up, adjusting for baseline levels, age, time since diagnosis, stage at diagnosis, race, ethnicity, education level, BMI category, and mean reflection level (B=0.26, SE 0.12, 95% CI 0.02-0.50; Cohen d=2.17). Participants in the fourth quartile of intervention adherence (21/75, 28%) averaged an increase of 0.38 bouts of muscle-strengthening physical activity in a typical week from baseline to follow-up, whereas those in the first quartile (40/75, 53%) averaged an increase of 1.43 bouts in a typical week. Similarly, higher intervention adherence was associated with less sleep disturbance at follow-up, adjusting for baseline levels and other covariates (B=−1.04, SE 0.50, 95% CI −2.02 to −0.06; Cohen d=2.08). Participants in the fourth quartile of intervention adherence averaged a decrease of 2.31 in the PROMIS-29 sleep disturbance score from baseline to follow-up, whereas those in the first quartile averaged a decrease of 4.68 in the PROMIS-29 sleep disturbance score. We did not observe statistically significant associations between intervention adherence and other outcomes of interest.

Mean Level of Reflection Predicting Outcomes

We observed associations between participants’ mean reflection level and change in 2 outcomes of interest (Table 3). Higher mean reflection level was associated with higher integrated regulation at follow-up, adjusting for baseline levels, age, time since diagnosis, stage at diagnosis, race, ethnicity, education level, BMI category, and intervention adherence (B=0.55, SE 0.25, 95% CI 0.06-1.04; Cohen d=2.20). Similarly, higher mean reflection level was associated with higher cognitive acceptance of physical activity at follow-up, adjusting for baseline levels and all other covariates (B=3.42, SE 1.70, 95% CI 0.09-6.75; Cohen d=2.01). We did not observe statistically significant

Table 3. Results from multiple linear regression analyses with independent variables regressed on intervention adherence and mean reflection level (n=75).

<table>
<thead>
<tr>
<th>Outcome variable</th>
<th>Intervention adherence</th>
<th>Mean reflection level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-reported weekly physical activity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aerobic moderate to vigorous exercise (min)</td>
<td>11.75 (8.08; −4.09 to 27.59)</td>
<td>−30.84 (31.50; −92.58 to 30.90)</td>
</tr>
<tr>
<td>Muscle-strengthening exercise (bouts)</td>
<td>0.26 (0.12; 0.02 to 0.50)</td>
<td>0.11 (0.49; −0.85 to 1.07)</td>
</tr>
<tr>
<td>BREQ-3c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identified regulation</td>
<td>0.01 (0.04; −0.07 to 0.09)</td>
<td>0.06 (0.16; −0.25 to 0.37)</td>
</tr>
<tr>
<td>Integrated regulation</td>
<td>0.00 (0.06; −0.12 to 0.12)</td>
<td>0.55 (0.25; 0.06 to 1.04)</td>
</tr>
<tr>
<td>Intrinsic regulation</td>
<td>0.01 (0.06; −0.11 to 0.13)</td>
<td>0.28 (0.25; −0.21 to 0.77)</td>
</tr>
<tr>
<td>PAAQd</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cognitive acceptance</td>
<td>−0.07 (0.41; −0.87 to 0.73)</td>
<td>3.42 (1.70; 0.09 to 6.75)</td>
</tr>
<tr>
<td>Behavioral commitment</td>
<td>0.07 (0.32; −0.56 to 0.70)</td>
<td>0.93 (1.33; 1.68 to 3.54)</td>
</tr>
<tr>
<td>PROMIS-29e</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical functioning</td>
<td>0.37 (0.41; −0.43 to 1.17)</td>
<td>−1.77 (1.65; −5.00 to 1.46)</td>
</tr>
<tr>
<td>Anxiety</td>
<td>0.82 (0.49; −0.14 to 1.78)</td>
<td>−1.19 (2.07; −5.25 to 2.87)</td>
</tr>
<tr>
<td>Depressive symptoms</td>
<td>−0.03 (0.41; −0.83 to 0.77)</td>
<td>2.45 (1.72; −0.92 to 5.82)</td>
</tr>
<tr>
<td>Fatigue</td>
<td>−0.18 (0.65; −1.45 to 1.09)</td>
<td>0.84 (2.66; −4.37 to 6.05)</td>
</tr>
<tr>
<td>Sleep disturbance</td>
<td>−1.04 (0.50; −2.02 to −0.06)</td>
<td>1.16 (2.07; −2.90 to 5.22)</td>
</tr>
<tr>
<td>Social roles</td>
<td>−0.28 (0.37; −1.01 to 0.45)</td>
<td>−0.83 (1.54; −3.85 to 2.19)</td>
</tr>
<tr>
<td>Pain interference</td>
<td>0.47 (0.60; −0.71 to 1.65)</td>
<td>0.12 (2.62; −5.26 to 5.02)</td>
</tr>
<tr>
<td>Pain intensity</td>
<td>0.11 (0.13; −0.14 to 0.36)</td>
<td>0.09 (0.58; −1.05 to 1.23)</td>
</tr>
</tbody>
</table>

a Adjusting for age, time since diagnosis, stage at diagnosis, race, ethnicity, education level, BMI category, mean reflection level, and baseline value of the construct.
b Adjusting for age, time since diagnosis, stage at diagnosis, race, ethnicity, education level, BMI category, intervention adherence, and baseline value of the construct.
c BREQ-3: Behavioral Regulation for Exercise Questionnaire-3.
d PAAQ: Physical Activity Acceptance Questionnaire.
e PROMIS: Patient-Reported Outcomes Measurement Information System.
associations between mean reflection level and other outcomes of interest.

Discussion

Principal Findings

This study was a secondary analysis of data from a digitally mediated intervention centered on applying principles and techniques from ACT to help people living with and beyond breast cancer overcome internal barriers to physical activity (eg, pain, fatigue, and frustration). We applied the framework presented by Fleck and Fitzpatrick [40] to gauge the depth of reflection evident in participants’ written responses to prompts encouraging reflection on the intervention’s weekly modules. We conducted content analysis to characterize participants’ responses and multiple linear regression analyses to explore to what extent intervention adherence and participants’ mean level of reflection were associated with study outcomes (ie, physical activity behaviors, motivation for physical activity, physical activity acceptance, and health-related quality of life). There were 407 written responses from the 75 participants over the course of 7 modules. Most of the responses were rated as either reflection level 0 (ie, description; 70/407, 17.2%) or level 1 (ie, reflective description; 247/407, 60.7%). Of the 407 responses, 90 (22.1%) demonstrated evidence of deeper levels of reflection. Intervention adherence was associated with more muscle-strengthening physical activity and better sleep outcomes. Mean reflection level was associated with more integrated motivation for physical activity and higher willingness to experience the full range of sensations that may accompany physical activity.

Comparison With Previous Literature

This study extends previous literature that has highlighted the importance of the cognitive and experiential aspects of DBCI engagement. To date, these aspects of DBCI engagement have largely been inferred from behavioral data (ie, system use metrics). In their review of methodologies for measuring engagement with DBCIs, Short et al [26] suggested that researchers should investigate how intervention content affects the cognitive and experiential aspects of DBCI engagement via the inclusion of open-ended items. We took this approach and found that participants’ responses to open-ended items could be evaluated using the reflection framework presented by Fleck and Fitzpatrick [40]. This approach is consistent with research in the education literature that has found reflective writing samples to be amenable to the quantitative assessment of the depth of reflection [35].

Our findings regarding the breakdown of reflection levels evident in this study were consistent with reviews that have evaluated reflective writing in the context of health-related education [39,47]. Even among graduate students, the great majority of reflective processes tend to occur at the descriptive level, and transformational and critical reflections are consistently rare [39,47]. There is limited literature applying these or similar techniques to better understand programs for health education and health promotion [39,47]. In this study, we observed considerable variation in the reflective depth of responses within and between individuals. This suggests that depth of reflection may be modifiable in this context. Indeed, reflection is a common goal of educational interventions, and it seems to be modifiable in other contexts [35]. Providing appropriate scaffolding, fostering collaboration in learning, and using varied exercises to stimulate reflection have been identified as techniques that may increase learners’ depth of reflection [35,39]. Time constraints, conflicting values, a lack of feedback, and a lack of trust have been identified as barriers to learning in the context of educational interventions promoting student reflection [35,39]. Future research is needed to identify how DBCIs might be designed to facilitate meaningful reflection and most effectively target the psychosocial determinants of lasting behavior change.

The themes identified in our qualitative analysis provided some insight into the variability of cognitive processes occurring in participants who experienced the digital physical activity intervention. Common themes in the lower levels of reflection generally suggested that participants had acquired new knowledge and endorsed intentions to change their physical activity behaviors (eg, making a resolution and knowledge of subject matter). These are important antecedents of successful behavior change [48]. However, they are generally not sufficient for realizing physical activity adherence goals; nearly half of the people who indicate that they intend to change their physical activity patterns do not do so [49]. The themes that characterized higher levels of reflective comments are concordant with processes that have been shown to moderate this intention-behavior gap [50]. Discussion of personal facilitators or mechanisms of behavior change and experiencing personal insight or a fundamental change in perspective conceptually align with the self-regulatory processes and notions of physical activity identity that tend to moderate the relationship between intention and physical activity adherence [50]. Encouraging personal reflection may be an autonomy-supportive approach to health promotion [34,51,52]. More research is needed to better understand if and how engaging in deep reflection may influence the conscious (eg, affective attitude) and automatic (eg, identity) processes that underlie successful behavior change.

Reflection may be a type of cognitive engagement uniquely suited to supporting knowledge transfer into other domains and contexts [36,53,54]. Schon [55] provides detailed commentary on reflective processes and distinguishes reflection-on-action (critical retrospective analysis) from reflection-in-action (conscious awareness of real-time behavioral modification). The author emphasizes the primacy of the latter as a determinant of sustained change and suggests that reflection-on-action may serve as a prerequisite for implementing change in real time. We found participants’ average depth of retrospective reflection to be associated with more integrated regulation for physical activity and cognitive acceptance in the context of physical activity but not with other outcomes of interest. It is encouraging that deeper levels of reflection on intervention content were associated with beneficial changes in motivation and physical activity acceptance, given that these were the theory-informed psychosocial constructs that the DBCI targeted [23]. The cognitive acceptance of physical activity has been shown to be associated with long-term changes in objectively measured physical activity, and integrated regulation is similarly predictive of change.

https://mhealth.jmir.org/2024/1/e51057

JMIR Mhealth Uhealth 2024 | vol. 12 | e51057 | p.775

(page number not for citation purposes)
of physical activity [45,56]. However, in this study, we did not observe evidence that the depth of reflection was associated with physical activity–related outcomes or other outcomes of interest. Simple intervention adherence was associated with increases in muscle-strengthening physical activity and reductions in sleep disturbance. This may be due to the intervention’s inclusion of practical resources for muscle-strengthening physical activity and mindfulness exercises. Our findings suggest that designing DBCIs to encourage reflective processes may help target some theory-supported mechanisms of action but may not always be necessary to engender changes in desired end points. More research is needed to investigate whether deep reflection mediates changes in key long-term behavioral and health outcomes and to what degree it should be prioritized in DBCIs.

Considering the depth of reflection may be a useful lens through which to evaluate participants’ cognitive engagement with the didactic components of DBCIs. At present, DBCIs are commonly oriented toward maximizing participant engagement as measured by system use. However, rather than simply attempting to maximize DBCI system use, it may be beneficial to target aspects of effective engagement [29]. In this study, participants’ depth of reflection was associated with 2 key study outcomes. It may be possible to design dynamic interventions that optimize psychological processes such as critical reflection. This endeavor would be supported by the ability to derive metrics that reliably reflect underlying psychological processes from participants’ verbal output. We demonstrated that applying the reflection framework presented by Fleck and Fitzpatrick [40] is a promising approach to quantifying participants’ qualitative data. Emerging technologies such as large language models may be applied to this end to make this process more expedient and conducive to just-in-time adaptive interventions. Digital health promotion efforts centered on optimizing psychological processes such as critical reflection may supplement or supplant approaches that are narrowly oriented toward maximizing system use.

Reflective writing can serve as a means of self-expression and have therapeutic effects [57]. However, it can also be perceived as burdensome by some individuals. Perski et al [30] emphasized the role played by participants’ subjective experiences, characterized by attention, interest, and affect, in DBCI engagement. There are likely important trade-offs that occur in optimizing for cognitive versus affective aspects of DBCI engagement. Participants’ comments in this study suggested that although many enjoyed the introspective aspects of the intervention, some did not. This observation is concordant with high ratings of acceptability in the parent study for those who completed the study but a less-than-ideal dropout rate (23.7%) [23]. Achieving deeper levels of reflection may have a dynamic and, in some contexts, diametrically opposed relationship with affective or subjective experiences. Particularly given the importance of affect in physical activity behavior [58], future research should investigate the interrelationships among these aspects of DBCI engagement and how to strike the right balance for different individuals in different contexts.

Limitations

The findings from this study must be considered in the context of its limitations. A small sample size, possible selection bias, and high attrition limit the generalizability of the findings of this study; the analytical sample was a convenience sample that was relatively well educated and had limited racial and ethnic diversity. Short et al [26] highlight that the use of qualitative methods to assess DBCI engagement is inherently limited by a lack of generalizability. The application of the reflection framework presented by Fleck and Fitzpatrick [40] along with directed content analytic methods may facilitate comparison in future studies; however, it is important to note that there may be other dimensions of reflection. We were primarily interested in assessing the vertical dimension of reflection (ie, the depth of reflection), but reflection has also been characterized as having an iterative process-oriented dimension that can be nonlinear and cyclic [39]. Given this prospect and the sometimes cyclical nature of behavior change itself [59], it may be beneficial to longitudinally investigate reflective processes and their bidirectional relationships with physical activity behaviors and determinants. Likewise, there are other ways to define intervention adherence. A limitation of this study is that we did not obtain other metrics that might characterize adherence, such as time spent in the modules. In future studies, obtaining more granular data concerning system use may be advantageous. Given that the study was conducted during the COVID-19 pandemic, history was a threat that may have influenced participants’ reflective processes and physical activity–related constructs. Reliance on the self-reported assessment of physical activity patterns also has well-documented limitations. Finally, we conducted multiple statistical tests and, although this investigation was explicitly exploratory in nature, the study findings are prone to an inflated chance of type I error.

Conclusions

In this study, we sought to explore a novel method for understanding critical reflection occurring in a DBCI designed to increase physical activity in insufficiently active people living with and beyond breast cancer. We found the application of qualitative content analysis based on the reflection framework presented by Fleck and Fitzpatrick [40] to be a useful tool for helping to gauge the extent to which participants engaged in reflective processes. Furthermore, we found that deeper reflection levels tended to be associated with changes in the targeted psychosocial constructs. Reflecting on newly acquired information is a critical process in integrating relevant insights for sustained behavior change. Encouraging personal reflection is an autonomy-supportive approach to promoting physical activity, and more research is warranted to investigate this approach in DBCIs serving people living with and beyond cancer.
Acknowledgments

MCR was supported by the National Cancer Institute of the National Institutes of Health (NIH; F31 CA236433). This study was supported in part by the Center for Energy Balance in Cancer Prevention and Survivorship. Participant recruitment was facilitated by the Love Research Army of the Dr Susan Love Research Foundation. This study used the MD Anderson Assessment, Intervention, and Measurement Core, a shared resource supported by the NIH through the MD Anderson Cancer Center support grant (P30 CA016672). The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Conflicts of Interest

None declared.

References


Abbreviations

ACT: acceptance and commitment therapy
BREQ-3: Behavioral Regulation for Exercise Questionnaire-3
DBCI: digital behavior change intervention
PAAQ: Physical Activity Acceptance Questionnaire
PROMIS: Patient-Reported Outcomes Measurement Information System
REDCap: Research Electronic Data Capture
SID: study identification number
SDT: self-determination theory

©Michael C Robertson, Emily Cox-Martin, Karen Basen-Engquist, Elizabeth J Lyons. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 09.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Using In-Shoe Inertial Measurement Unit Sensors to Understand Daily-Life Gait Characteristics in Patients With Distal Radius Fractures During 6 Months of Recovery: Cross-Sectional Study

Akiko Yamamoto¹, MD, PhD; Eriku Yamada¹, MD, PhD; Takuya Ibara², PhD; Fumiyuki Nihey³, PhD; Takuma Inai⁴, PhD; Kazuya Tsukamoto¹, MD; Tomohiko Waki¹, MD; Toshitaka Yoshii¹, MD, PhD; Yoshiyuki Kobayashi⁵, PhD; Kentaro Nakahara¹, PhD; Koji Fujita²,6, MD, PhD

¹Department of Orthopaedic and Spinal Surgery, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan
²Department of Functional Joint Anatomy, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan
³Biometrics Research Laboratories, NEC Corporation, Chiba, Japan
⁴Biomechanics and Exercise Physiology Research Group, Health and Medical Research Institute, Department of Life Science and Technology, National Institute of Advanced Industrial Science and Technology, Kagawa, Japan
⁵Human Augmentation Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan
⁶Division of Medical Design Innovations, Open Innovation Center, Institute of Research Innovation, Tokyo Medical and Dental University, Tokyo, Japan

Corresponding Author:
Koji Fujita, MD, PhD
Department of Functional Joint Anatomy
Graduate School of Medical and Dental Sciences
Tokyo Medical and Dental University
1-5-45, Yushima
Bunkyo-ku
Tokyo, 113-8519
Japan
Phone: 81 3 5803 5279
Email: fujioth@tmd.ac.jp

Abstract

Background: A distal radius fracture (DRF) is a common initial fragility fracture among women in their early postmenopausal period, which is associated with an increased risk of subsequent fractures. Gait assessments are valuable for evaluating fracture risk; inertial measurement units (IMUs) have been widely used to assess gait under free-living conditions. However, little is known about long-term changes in patients with DRF, especially concerning daily-life gait. We hypothesized that, in the long term, the daily-life gait parameters in patients with DRF could enable us to reveal future risk factors for falls and fractures.

Objective: This study assessed the spatiotemporal characteristics of patients with DRF at 4 weeks and 6 months of recovery.

Methods: We recruited 16 women in their postmenopausal period with DRF as their first fragility fracture (mean age 62.3, SD 7.0 years) and 28 matched healthy controls (mean age 65.6, SD 8.0 years). Daily-life gait assessments and physical assessments, such as hand grip strength (HGS), were performed using an in-shoe IMU sensor. Participants’ results were compared with those of the control group, and their recovery was assessed for 6 months after the fracture.

Results: In the fracture group, at 4 weeks after DRF, lower foot height in the swing phase (P=.049) and higher variability of stride length (P=.03) were observed, which improved gradually. However, the dorsiflexion angle in the fracture group tended to be lower consistently during 6 months (at 4 weeks: P=.06; during 6 months: P=.07). As for the physical assessments, the fracture group showed lower HGS at all time points (at 4 weeks: P<.001; during 6 months: P=.04), despite significant improvement at 6 months (P<.001).

Conclusions: With an in-shoe IMU sensor, we discovered the recovery of spatiotemporal gait characteristics 6 months after DRF surgery without the participants’ awareness. The consistently unchanged dorsiflexion angle in the swing phase and lower HGS could be associated with fracture risk, implying the high clinical importance of appropriate interventions for patients with...
DRF to prevent future fractures. These results could be applied to a screening tool for evaluating the risk of falls and fractures, which may contribute to constructing a new health care system using wearable devices in the near future.

(JMIR Mhealth Uhealth 2024;12:e55178) doi:10.2196/55178

KEYWORDS
distal radius fracture; gait analysis; daily life; long-term results; gait; sensor; sensors; walk; walking; fracture; fractures; wearable; wearables; recover; rehabilitation; spatiotemporal; inertial measurement; fragility; postmenopausal; menopause; grip; surgery; surgical; orthopedic; postoperative; orthopedics; fall; falls; bone; bones; wrist; radius; radial

Introduction

Gait analysis is useful for predicting future fall risk and reflecting various underlying physiological processes [1]. Quantitative gait characteristics, such as slower gait speed and shorter stride length, are associated with falls, resulting in fragility fractures [2-4]. Recently, inertial measurement units (IMUs) have been widely used to assess gait under free-living conditions owing to their convenience, low cost, small size, and high accuracy [5-7]. The shank and foot are the preferred related to falls and physical ability [4].

Among fragility fractures resulting from falls, distal radius fractures (DRFs) are one of the most frequent initial fractures in older adults [10]. Many DRFs occur in women in their postmenopausal period, aged <75 years, who are healthy, active, and functionally independent. More than half of these women do not meet the criteria for osteoporosis [11-13]. However, the initial DRF is associated with a greater risk of functional decline [14] and subsequent fractures [15] in all age groups. These changes are more than 5 times higher, even in those aged 50-59 years [16]. This could reflect early changes in frailty [17].

In patients with DRF, lower gait ability was observed in the laboratory, which slightly improved 6 months after the surgery [13,18]. However, these gait assessments were mainly performed for approximately 10 seconds, which may not accurately depict daily-life gait [19]. Further, most studies have only highlighted the consequences of wrist function and pain when investigating the long-term outcomes of DRF, and little is known about the effect of DRF on physical abilities, such as activities of daily living. Therefore, the long-term alterations of daily-life gait characteristics in patients with fractures remain unknown, and this study attempts to bridge this gap.

We previously found out that in-shoe IMU sensors were effective in the assessment of daily-life gait in patients with an initial DRF [7]. We hypothesized that, in the long term, it could enable us to identify future risk factors for secondary fractures by spatiotemporally following daily-life gait parameters using IMU sensors. We aimed to reveal the characteristics of spatiotemporal gait changes during 6 months following DRF.

Methods

Ethical Considerations

This study was approved by the Institutional Review Board of Tokyo Medical and Dental University (M2020-365) and followed the tenets of the Declaration of Helsinki. Written informed consent was provided by all participants. Participation in the study was voluntary, and no compensation was awarded for participation.

Recruitment

In the fracture group, we recruited 16 female patients with DRF who had undergone surgery for their first fragility fracture from 5 general hospitals. We compared their results to those of 28 healthy female volunteers. Women without a history of fragility fractures were recruited as the control group through local media advertisements. The inclusion criteria involved having the ability to walk without any support, no history of lower-extremity injury, and no known neuromuscular disorders or neurophysiological problems that may affect gait. Fragility fractures were defined as those that followed a fall from standing height or less. We excluded patients with DRF due to traffic or industrial accidents or multiorgan injuries. In the fracture group, 6 patients with DRF fell in the house without shoes, and the other 10 patients fell while wearing shoes. Owing to the lack of previous literature on the long-term data of gait in patients with DRF, the sample size estimations were based on the effect size of 0.78 (from the result of hand grip strength [HGS]) [13], with an assumed power of 0.8 and a type 1 error of 0.05. A sample of 16 participants with fractures was analyzed using G*Power (version 3.1; Heinrich Heine University Duesseldorf) [20].

Daily-Life Gait Assessments

We measured daily-life gait using in-shoe IMU sensors (A-RROWG, NEC Corporation; Figure 1). These sensors are small (40.0 mm $\times$ 30.5 mm $\times$ 7 mm) and lightweight (11 g), including a 3-axis accelerometer and gyroscope. The IMU sensor in the dedicated insole was placed at the foot arch, and the x-, y-, and z-axes of the IMUs were set along the mediolateral, anteroposterior, and vertical directions, respectively. When a person wearing these sensors walks in a stable straight line over 3 gait cycles between 5 AM and 10 PM, the in-shoe IMU sensor detects that the person is walking based on acceleration in the anteroposterior direction and saves the IMU signals of the next 3 gait cycles as 1 gait measurement [21]. The IMU signals were sampled at a rate of 100 Hz, transferred to a smartphone via Bluetooth, and stored in a specialized app if participants had one with them. If a person did not have a smartphone, the data were uploaded automatically via Bluetooth at 11 PM by keeping the smartphone near the IMU sensors.

From the saved IMU signals, the mean of 7 gait parameters from 3 gait cycles was calculated and stored on a smartphone, as previously described by Fukushi et al [21]. The following 7 parameters were calculated:
1. Gait speed: calculated as stride length (m) divided by stride time (s).
2. Stride length: the distance between the start and end points of the foot trajectory for 1 stride.
3. Dorsiflexion angle: the peak foot angle in the dorsal direction from the ground during the swing phase.
4. Plantarflexion angle: the peak foot angle in the plantar direction from the ground during the swing phase.
5. Foot height: the maximum height of the foot trajectory.
6. Toe-in or toe-out angle: the mean angle of foot adduction or abduction in the direction of the velocity vector during the swing phase.
7. Circumduction: the displacement in the medial-lateral direction during the swing phase.

In addition to these 7 gait parameters, the coefficient of variation (CV), calculated as SD divided by mean multiplied by 100, was used to evaluate the variability.

Figure 1. The in-shoe inertial measurement unit (IMU) sensor. The sensors with accelerometer and gyroscope were set into the dedicated insoles, measured the daily-life gait without the participants’ awareness, and saved the obtained data to their smartphone via Bluetooth.

Measurement Protocol

All participants completed a paper-based questionnaire on their general health status, which included fall history in the past year, frequency of stumbling, and fear of falling. Falls at the time of fracture in patients with DRFs were excluded based on the number of falls in the past year.

In the daily gait assessments, we inserted IMU sensors with dedicated insoles into the preferred shoes of both participants’ feet. Each participant was provided a smartphone with only the original app for storing gait data. We verified whether participants could walk using the sensors and whether the sensors worked with a smartphone. Participants were instructed to wear shoes with sensors for 2 to 6 weeks. We requested that participants spend their daily lives as usual; therefore, we did not establish a minimum time for wearing or walking with them. The measurements were recorded without the participants’ awareness, and they could view their most recent gait data by checking their smartphone. In the fracture group, daily gait assessments in the early postoperative period started 2 weeks after DRF surgery to enable the effects of casting or surgery.

To assess baseline functional ability and frailty, HGS [22] and the Timed Up and Go (TUG) test [23] were performed. HGS was measured in kilograms with a Jamer dynamometer (Sammons Preston). We assessed HGS on the nonfractured side of the fracture group and both sides of the control group. The mean values of 3 measurements were recorded. The time to complete the TUG test was recorded twice: once at the preferred speed and once at the fastest speed. These baseline physical tests were performed 4 weeks after DRF surgery in the fracture group and at the beginning of daily gait measurements in the control group.

To check for long-term functional ability, we asked the fracture group patients to perform the same physical and daily gait assessments again at an outpatient visit 6 months after the surgery.

Data Analysis

The median and CV of each participant were calculated using the obtained gait data. However, the automatically collected data included hills, turns, and straddling, and we used Smirnov-Grubbs analysis for every gait parameter to exclude any outliers from the data. After exclusion, approximately 20-1000 gait measurements were recorded for each participant during the measurement period, reflecting differences in their lifestyles. Even after the exclusion, gait data included various gait types, such as walking exercise and hurried walking; therefore, we excluded participants with less than 50 gait measurements in either foot, based on a previous report [24]. Since we computed the data obtained from both feet together,
we selected participants with 100 or more gait assessments for this study.

**Statistical Analysis**

Between-group differences were assessed using Student $t$ test (2-tailed) for continuous variables in the patient demographics, and the chi-square test for categorical variables. Since the participants’ gait data were not normally distributed, they were assessed using nonparametric (ie, Kruskal-Wallis and Steel) tests. The results of the fracture group in the early postoperative and long-term periods were compared using a paired $t$ test. Statistical significance was set at $P<.05$.

These analyses were performed using EZR (version 1.55; Division of Haematology, Saitama Medical Center, Jichi Medical University) [25].

**Results**

**Participants’ Demographics**

There were no significant differences in age, body characteristics, or lifestyle variables between the 2 groups. In the section on questions about falls, patients with fractures demonstrated a significantly higher incidence of falls (Table 1). None of the fracture group members experienced falls or subsequent fractures after the initial DRF.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Control (n=28)</th>
<th>Fracture (n=16)</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>62.3 (7.0)</td>
<td>65.6 (8.0)</td>
<td>.20(^a)</td>
</tr>
<tr>
<td>Height (cm), mean (SD)</td>
<td>155.2 (4.3)</td>
<td>154.8 (4.0)</td>
<td>.77(^a)</td>
</tr>
<tr>
<td>Body weight (kg), mean (SD)</td>
<td>54.4 (8.0)</td>
<td>51.6 (8.5)</td>
<td>.30(^a)</td>
</tr>
<tr>
<td>BMI (kg/m(^2)), mean (SD)</td>
<td>22.6 (3.2)</td>
<td>21.5 (3.2)</td>
<td>.28(^a)</td>
</tr>
<tr>
<td>Hand dominance (right), n (%)</td>
<td>27 (96)</td>
<td>15 (94)</td>
<td>.68(^b)</td>
</tr>
<tr>
<td>Foot dominance (right), n (%)</td>
<td>23 (82)</td>
<td>14 (88)</td>
<td>.64(^b)</td>
</tr>
<tr>
<td>Smoking (current and previous), n (%)</td>
<td>5 (18)</td>
<td>5 (31)</td>
<td>.31(^b)</td>
</tr>
<tr>
<td>Alcohol consumption, n (%)</td>
<td>11 (39)</td>
<td>5 (31)</td>
<td>.59(^b)</td>
</tr>
</tbody>
</table>

**Comorbidities, n (%)**

- Hypertension: 8 (29) vs 3 (19), $P=.47\(^b\)$
- Eye disease: 1 (4) vs 1 (6), $P=.68\(^b\)$
- Diabetes mellitus: 0 (0) vs 0 (0), $P=.48\(^b\)$
- Knee osteoarthritis: 2 (7) vs 0 (0), $P=.27\(^b\)$
- Hip osteoarthritis: 0 (0) vs 0 (0), $P=.48\(^b\)$
- Rheumatoid arthritis: 0 (0) vs 0 (0), $P=.48\(^b\)$
- Number of oral medications, mean (SD): 0.8 (0.03) vs 0.9 (0.06), $P=.27\(^a\)$
- The experience of fall in the past year, n (%): 0 (0) vs 4 (25), $P=.006\(^b\)$

**Number of falls, n**

- Once: N/A\(^c\) vs 1, N/A
- twice: N/A vs 2, N/A
- 3 times: N/A vs 1, N/A
- The experience of stumbling, n (%): 17 (61) vs 9 (56), $P=.77\(^b\)$

\(^a\)Independent Student $t$ tests were used to compare the groups.

\(^b\)Chi-square test was used for analysis between the groups.

\(^c\)N/A: not applicable.

**Spatiotemporal Data in Daily-Life Gait**

The results of the spatiotemporal daily-life gait are presented in Table 2. There were no between-group differences in the number of measurements. Compared with the control group, the dorsiflexion angle measured at any time postoperatively in the fracture group tended to be lower and demonstrated no improvement in the postoperative course. Participants in the fracture group demonstrated significant recovery in gait speed, stride length, and plantarflexion angle between 4 weeks and 6
months postoperatively. Among the CV of each gait parameter, only the CV of stride length between the control group and the 4 weeks–postfracture group demonstrated a significant difference.

Table 2. Daily-life spatiotemporal data. P values <.05 are considered significant.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Control group (n=28)</th>
<th>Fracture group (n=16)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4 weeks after surgery</td>
<td>6 months after surgery</td>
<td></td>
</tr>
<tr>
<td>Number of measurements, mean (SD)</td>
<td>479.3 (432.7)</td>
<td>746.6 (468.7)</td>
<td>.06a</td>
</tr>
<tr>
<td>Median of each parameter, mean (SD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gait speed (m/s)</td>
<td>1.28 (0.12)</td>
<td>1.22 (0.09)</td>
<td>.17a</td>
</tr>
<tr>
<td>Stride length (m)</td>
<td>1.26 (0.12)</td>
<td>1.20 (0.09)</td>
<td>.26a</td>
</tr>
<tr>
<td>Dorsiflexion angle (degree)</td>
<td>26.1 (3.83)</td>
<td>22.8 (4.15)</td>
<td>.03a</td>
</tr>
<tr>
<td>Plantarflexion angle (degree)</td>
<td>75.0 (6.18)</td>
<td>71.5 (4.23)</td>
<td>.08a</td>
</tr>
<tr>
<td>Foot height (cm)</td>
<td>14.0 (1.06)</td>
<td>13.1 (1.35)</td>
<td>.08a</td>
</tr>
<tr>
<td>Circumduction (cm)</td>
<td>2.85 (0.85)</td>
<td>3.16 (0.49)</td>
<td>.12a</td>
</tr>
<tr>
<td>Toe-in or toe-out angle (degree)</td>
<td>13.2 (4.63)</td>
<td>13.6 (3.90)</td>
<td>.93a</td>
</tr>
<tr>
<td>CVd of each parameter (%), mean (SD)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gait speed</td>
<td>15.2 (4.84)</td>
<td>16.3 (2.95)</td>
<td>.36a</td>
</tr>
<tr>
<td>Stride length</td>
<td>10.3 (2.89)</td>
<td>12.6 (3.21)</td>
<td>.05a</td>
</tr>
<tr>
<td>Dorsiflexion angle</td>
<td>20.7 (6.09)</td>
<td>23.2 (5.48)</td>
<td>.19a</td>
</tr>
<tr>
<td>Plantarflexion angle</td>
<td>8.56 (2.74)</td>
<td>10.2 (2.70)</td>
<td>.14a</td>
</tr>
<tr>
<td>Foot height</td>
<td>8.01 (2.15)</td>
<td>11.0 (7.05)</td>
<td>.12a</td>
</tr>
<tr>
<td>Circumduction</td>
<td>51.3 (12.8)</td>
<td>46.5 (12.0)</td>
<td>.45a</td>
</tr>
<tr>
<td>Toe-in or toe-out angle</td>
<td>30.7 (15.3)</td>
<td>37.9 (19.9)</td>
<td>.51a</td>
</tr>
</tbody>
</table>

aKruskal-Wallis test was used to compare the control and fracture groups.
bSteel test was used to compare each group.
cPaired sample t test was used for analysis between the groups.
dCV: coefficient of variation.

HGS and Body Balancing Ability
The HGS in the fracture group demonstrated significant recovery between 4 weeks and 6 months after surgery; however, it was significantly lower in the fracture group compared to the control group. In the TUG test, there were no significant differences between the control and fracture groups or at 4 weeks and 6 months postoperatively in the fracture group (Table 3).
Table 3. Physical tests in the control and fracture groups. P values <.05 are considered significant.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Control group (n=28)</th>
<th>Fracture group (n=16)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4 weeks after surgery</td>
<td>6 months after surgery</td>
<td>Kruskal-Wallis test</td>
</tr>
<tr>
<td>Hand grip strength (kg), mean (SD)</td>
<td>23.3 (3.4)</td>
<td>19.1 (2.6)</td>
<td>20.6 (3.1)</td>
</tr>
<tr>
<td><strong>TUG&lt;sup&gt;d&lt;/sup&gt; test (s), mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal speed</td>
<td>8.07 (1.33)</td>
<td>7.53 (0.85)</td>
<td>8.2 (1.28)</td>
</tr>
<tr>
<td>Faster speed</td>
<td>6.23 (0.89)</td>
<td>6.09 (0.64)</td>
<td>6.4 (0.95)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Kruskal-Wallis test was used to compare the control and fracture groups.

<sup>b</sup>Steel test was used to compare each group.

<sup>c</sup>Paired sample t test was used for analysis between the groups.

<sup>d</sup>TUG: Timed Up and Go.

Discussion

**Principal Results**

We performed daily-life gait assessments with in-shoe IMU sensors and some physical tests, including HGS, to evaluate the differences in spatiotemporal gait and physical ability between patients with DRF and healthy controls. Moreover, we assessed whether these parameters improved during 6 months following DRF. In daily-life gait assessments, patients with DRF demonstrated a lower foot height and higher CV stride length compared to the control group; however, these differences were no longer present 6 months after DRF surgery, reflecting the improvement in some parameters in the fracture group after 6 months of DRF treatment. On the other hand, the dorsiflexion angle in the fracture group tended to be lower consistently during 6 months. In the physical assessments, patients with DRF at 4 weeks and 6 months after the surgery had significantly lower HGS than those without DRF, even though HGS in patients with DRF improved during 6 months.

Some gait parameters, such as gait speed and stride length, improved in the fracture group during 6 months of daily life. Further, the CV of stride length, which is associated with fall risk [26], was higher in the fracture group at 4 weeks after DRF and did not significantly change after 6 months. With the development of wearable sensors, spatiotemporal gait characteristics and their variability in daily life have been increasingly identified. However, few studies have explored the long-term changes in gait characteristics related to falls and fractures. As daily-life gait is influenced by various factors, such as environmental and psychiatric factors [27], changes in patients with DRF could reflect changes in their lifestyle. Conversely, the risk of subsequent fractures is the highest immediately after the initial fracture [28]. Although no subsequent fractures occurred in this study, further research is warranted to determine the relationship between gait changes and subsequent fractures. The in-shoe IMU sensor must play an important role in further evaluating this relationship.

The dorsiflexion angle in the fracture group remained lower, whereas other parameters in the fracture group improved during 6 months. As for vertebral fractures, which are typical fragility fractures as well as DRF, patients with symptomatic vertebral fractures walked with shorter and wider strides at the time of injury. Although those with vertebral fractures show improvement in stride time and stride length over time, even reaching healthy levels again, their gait pattern and stability persist for 6 months, implying a greater risk of incident disability among these patients [29]. Since gait speed is reported to affect other gait parameters, the persistent lower dorsiflexion angle in the fracture group, despite the improvement in gait speed, might be a characteristic of patients with DRF, indicating that patients cannot fully return to healthy states. The dorsiflexion angle in this study, which means the angle between the ground and the sole of the foot, may depend on the movement of all lower extremities. Kyphosis and flexed hip or knee joints, which are common in older people, are related to foot movement or strength [30]. Although further research is needed to determine the cause of this decrease in dorsiflexion angle, the angle could result in stumbling and falls. Older adults with DRF can be assessed as having a high risk of functional decline, particularly those who have access to a health care facility at an early stage. They should receive appropriate intervention to prevent future falls or fractures along with treatment for the initial fracture. Considering that several previous reports have assessed fall risk using machine learning based on gait data from fallers [31,32], our results could be effective in creating a more precise machine learning model for evaluating the risk of falls. Further research is warranted to explore not only the cost of developing sensors and apps but also intervention methods and the extent of fall reduction achievable. Nonetheless, our findings using in-shoe IMU sensors outside the hospital could be valuable for future screening tools to evaluate the risk of falls and fractures.

As for physical assessments, the HGS in the fracture group significantly improved during 6 months after DRF, which is consistent with a previous report [13]. The increased use of the nonaffected hand with DRF in daily life may improve HGS; however, most studies on DRF have focused on the HGS of the nonaffected side, and little is known about that of the nonaffected side. Generally, HGS is associated with health status, including death, falls, and muscle strength [33]. The improvement in HGS could reflect the improvement in health status, and the lower HGS after 6 months of DRF could be associated with a lower...
degree of health status in the fracture group. Contrastingly, there were no differences in the TUG test results, contrary to our previous reports [13,18]. The average TUG test time in both the fracture and control groups in this study was faster compared to previous studies, which may mean that the TUG test, which involves a few steps and seconds in the laboratory, may not reflect the true physical characteristics. Therefore, daily-life gait analysis for a certain period is needed to identify the slight difference between fracture group patients and healthy controls. By using this in-shoe sensor for a certain period, the lower dorsiflexion angle in patients with DRF was revealed, which remained 6 months after the fracture. We would like to further investigate long-term changes in the gait of these patients.

Limitations
This study had some limitations. First, the number of participants was small, which could have affected the power of this study. However, individual changes over 6 months appeared to confirm the statistical outcomes and might not have affected the overall conclusions of our study. Second, we observed progress up to 6 months after the fractures. The HGS on the affected side continued to improve beyond 1 year. Long-term changes in more patients with DRF should be further explored. Third, the participants were all Japanese, who had the habit of taking off their shoes indoors. Considering that 6 patients with DRF fell inside their houses without shoes, the results may not accurately reflect the daily free-living assessments of barefoot individuals.

Conclusions
In summary, we performed a case-control study to investigate the long-term changes in HGS and daily-life gait after DRF. Using an in-shoe IMU sensor, we revealed the recovery of spatiotemporal gait characteristics 6 months after DRF surgery without the participants’ awareness. The dorsiflexion angle in the swing phase and HGS were still lower in the fracture group after 6 months, which could be associated with fracture risk. This in-shoe IMU sensor could be useful for evaluating the future fall and fracture risk outside the hospital and for constructing a new health care system related to preventive medicine using wearable devices outside the hospital.

Acknowledgments
This study was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (grant 21K17469) and joint research funding from NEC Corporation. We would like to thank Masahiro Fujimoto, Kanako Nakajima, Shoma Kudo, and Gaku Kutsuzawa for collecting data; Hiroshi Kajitani for collecting data and software development; Akimoto Nimura, Hidetoshi Kaburagi, and Takafumi Koyama for critical analysis and proofreading; and Editage for English language editing.

Conflicts of Interest
KF has received joint research funding from NEC Corporation. FN and KN are employees of NEC Corporation.

References


Abbreviations

CV: coefficient of variation
DRF: distal radius fracture
HGS: hand grip strength
IMU: inertial measurement unit
TUG: Timed Up and Go
Effectiveness of Telemonitoring in Reducing Hospitalization and Associated Costs for Patients With Heart Failure in Finland: Nonrandomized Pre-Post Telemonitoring Study

Jorma Kokkonen, MD, PhD; Pirjo Mustonen, MD, PhD; Eija Heikkilä, PhD; Riikka-Leena Leskelä, PhD; Paula Pennanen, MSc; Kati Krühn, MSc; Arto Jalkanen, MSc; Jussi-Pekka Laakso, MSc; Jari Kempers, PhD; Sami Väisänen, PhD; Paulus Torkki, PhD

1Tampere Heart Hospital, Tampere, Finland
2The Wellbeing Services County of Southwest Finland, Turku, Finland
3Nordic Healthcare Group, Helsinki, Finland
4Roche Diagnostics (Schweiz) AG, Zug, Switzerland
5The Wellbeing Services County of South Savo, Mikkeli, Finland
6Ceneron Advisor Oy, Helsinki, Finland
7European Health Economics Oy, Jyväskylä, Finland
8Roche Diagnostics Oy, Espoo, Finland
9Department of Public Health, Faculty of Medicine, University of Helsinki, Helsinki, Finland

Corresponding Author:
Jorma Kokkonen, MD, PhD
Tampere Heart Hospital
Elämänaukio 1
Tampere, 33520
Finland
Phone: 358 50 5981078
Email: sydanjopi@gmail.com

Abstract

Background: Many patients with chronic heart failure (HF) experience a reduced health status, leading to readmission after hospitalization despite receiving conventional care. Telemonitoring approaches aim to improve the early detection of HF decompensations and prevent readmissions. However, knowledge about the impact of telemonitoring on preventing readmissions and related costs remains scarce.

Objective: This study assessed the effectiveness of adding a telemonitoring solution to the standard of care (SOC) for the prevention of hospitalization and related costs in patients with HF in Finland.

Methods: We performed a nonrandomized pre-post telemonitoring study to estimate health care costs and resource use during 6 months on SOC followed by 6 months on SOC with a novel telemonitoring solution. The telemonitoring solution consisted of a digital platform for patient-reported symptoms and daily weight and blood pressure measurements, automatically generated alerts triggering phone calls with secondary care nurses, and rapid response to alerts by treating physicians. Telemonitoring solution data were linked to patient register data on primary care, secondary care, and hospitalization. The patient register of the Southern Savonia Social and Health Care Authority (Essote) was used. Eligible patients had at least 1 hospital admission within the last 12 months and self-reported New York Heart Association class II-IV from the central hospital in the Southern Savonia region.

Results: Out of 50 recruited patients with HF, 43 completed the study and were included in the analysis. The hospitalization-related cost decreased (49%; P=.03) from €2189 (95% CI €1384-€2994) to €1114 (95% CI €425-€1803) during telemonitoring. The number of patients with at least 1 hospitalization due to HF was reduced by 70% (P=.002) from 20 (47%) out of 43 patients during SOC to 6 (14%) out of 43 patients in telemonitoring. The estimated mean total health care cost per patient was €3124 (95% CI €2212-€4036) during SOC and €2104 (95% CI €1313-€2895) during telemonitoring, resulting in a 33% reduction (P=.07) in costs with telemonitoring.
Conclusions: The results suggest that the telemonitoring solution can reduce hospital-related costs for patients with HF with a recent hospital admission.

Introduction

The prevalence of heart failure (HF) and related costs is increasing worldwide due to an aging population [1]. The estimated prevalence of HF in the adult population is 1% to 2%, increasing to 10% in older adults aged 70 years or older [2,3]. HF often leads to gradual or acute changes in HF symptoms (decompensation) that require repeated and prolonged hospitalization [4]. Hospital admission is a strong predictor of further hospital admission: 20% to 25% of patients with HF are rehospitalized within 1 month and approximately 50% within 5 months of discharge [5]. Decompensation requiring hospitalization is also linked to increased mortality. A European registry study following patients for 1 year after hospitalization reported mortality rates of 24% for acute HF and 6.4% for chronic HF [6]. Hospitalization accounted for around 80% of HF health care costs [1].

An early return to the hospital following discharge may be a result of incomplete inpatient treatment and poor coordination and planning of follow-up care. Even for patients with regular follow-up care, however, the signs of decompensation may not occur during cardiology visits. Patients often contact clinics when symptoms are at an advanced stage [7]. Self-monitoring of symptoms, such as increased blood pressure, weight gain, or other health status-related symptoms, is particularly important in HF management [4]. Self-monitoring requires patients to be motivated to measure symptoms associated with HF and to have access to clinical advice when symptoms appear [8].

Remote monitoring aims to improve monitoring of patients’ health status and is defined as a part of telehealth [9]. A basic level of remote monitoring involves regular and structured telephone support provided by health care professionals (HCPs) to discuss symptoms, self-monitoring measurements, lifestyle, and drug therapy. Structured telephone support can reduce HF-related hospitalization but does not seem to have an impact on the all-cause hospitalization of patients with HF [10]. Remote monitoring solutions are noninvasive stand-alone systems in which patient data on biometric measurements (such as body weight, blood pressure, and heart rate) and reported symptoms are frequently transmitted to HCPs through a secure digital system. HCPs manually review the data on digital platforms, which may also include integrated automated alerts, and necessary action is taken to optimize treatment.

The effect of noninvasive telemonitoring has been compared to the standard of care (SOC) in several studies, primarily through randomized trials. Some studies found telemonitoring had a beneficial impact on reducing hospitalization [11], while others did not find any effect [12,13]. However, a recent meta-analysis, encompassing 91 randomized trials and observational studies, revealed that noninvasive telemonitoring reduced all-cause mortality by 16%, first hospitalization by 19%, and total HF hospitalizations by 15%. When comparing telemonitoring studies and developing optimal telemonitoring approaches, it is crucial to consider various determinants, including the telemonitoring intervention models, health care systems, and the characteristics of the population with HF in the studies [14].

There are only a few international studies that have explored the cost-effectiveness of telemonitoring compared to the SOC [15-18].

The objective of this nonrandomized pre-post intervention study in patients with HF with a recent (<12 month) hospitalization was to assess the effectiveness of adding a telemonitoring solution to SOC on hospitalizations and related costs in the Finnish health care system. The study compared hospitalization occurrence and related costs with SOC and following the introduction of a telemonitoring solution. Secondary outcomes included hospital admissions and total health care costs.

Methods

Study Design

The nonrandomized pre-post intervention study was performed in Southern Savonia, Finland. During the 12-month study period, patients were treated with SOC for the first 6 months and then with a telemonitoring solution in addition to SOC for the next 6 months. The primary outcome was hospitalization-related costs during 6 months with SOC versus telemonitoring. Secondary outcomes included the number of patients with at least 1 hospital admission due to HF or a cardiovascular cause other than HF emergency care visits and primary care or cardiology (secondary care) calls and visits. Health care costs for secondary outcomes included the total health care costs of primary care, secondary care (for cardiology), emergency visits, and phone calls. The study was designed to demonstrate the effectiveness of remote monitoring within the Finnish health care system. The costs of the telemonitoring service itself were not analyzed.

Health care resource use was collected for each patient during SOC and telemonitoring from the patient register of the Southern Savonia Social and Health Care Authority (Essote). The data was pseudonymized by the register holder. The Health Care Authority is responsible for all social and health care services for the population of approximately 100,000 inhabitants in Southern Savonia, Finland.

Study Patients

Patients were recruited from Mikkeli Central Hospital in Finland’s Southern Savonia region. Patients with an HF
diagnosis confirmed by a cardiologist, at least 1 hospital admission in the 12 months preceding study initiation, and self-reported New York Heart Association (NYHA) class II-IV were eligible for the study (Figure 1).

The inclusion criteria also stated that patients must be able to manage the telemonitoring devices and digital platform used in the study. Palliative care was an exclusion criterion.

Figure 1. Flowchart for patient selection for the study.

Study Procedures
The SOC, during the first 6 months, included regular cardiology appointments and laboratory tests planned by a cardiologist for each patient with HF according to local care guidelines for HF treatment. Nurses followed up with patients through phone calls, depending on the state of HF. After inpatient stays, the cardiologist or internist at the hospital made an individual plan for the follow-up of patients posthospitalization. During the follow-up period, patients measured their weight and blood pressure at home, and nurses followed up with patients through phone calls to discuss their health status and measurement results.

Telemonitoring was added on top of SOC during the next 6 months and consisted of a digital platform, home measurement devices, and nurses monitoring patients through the digital platform. The digital platform used Veta Health’s remote patient monitoring platform (Veta Health Inc), customized for the study. Patients used their smartphones, handheld devices, or personal computers to access the digital platform. Patients measured their weight daily with a digital scale (Omron Corporation) and their blood pressure with a digital blood pressure measuring device (Omron M7000 Intelli IT) and transferred the measurements into the digital platform (Figure 2). The digital platform also included symptom-related questions.
The digital platform automatically compared patients’ body weight and HF symptom answers against preset thresholds and generated semiurgent or urgent alerts predicting HF worsening (Table S1 in Multimedia Appendix 1). Depending on the alert type, the digital platform either advised the patient to contact a nurse or a nurse to contact the patient to validate the health status. Nurses had access to alerts on working days. If needed, nurses referred a patient to a cardiologist to optimize HF care or medication. The treating cardiologist reacted to the nurses’ referrals within 24 hours. For urgent alerts, the digital platform advised patients to go to emergency care. Nurses also provided technical support for patients as required. The digital platform collected blood pressure data and laboratory results from regular health care visits, not for the alert algorithm but to allow nurses to evaluate the patient’s health status.

**Health Care Resource Use and Costs**
Patients’ health care costs and resource use were estimated from the Essote patient register and consisted of public primary care, secondary care (cardiology unit), emergency visits, hospitalizations (cardiology and internal medicine; primary care), and phone calls for primary and secondary care (cardiology). A unique personal identification number for each resident in Finland connected the digital platform data and patient register data. The *International Classification of Diseases, Tenth Revision* (ICD-10) diagnosis code registered for each health care event was used to separate hospitalizations for HF (ICD-10 code I50) from hospitalizations for a cardiovascular cause other than HF (ICD-10 codes I10, I25, I42, I46, I48, I49, and I70). All-cause hospitalizations included hospitalizations with any diagnosis. The costs of health care use were calculated using Essote diagnosis-related group prices.

**Statistical Analysis**
The study analysis included only patients who completed the study. Patients who died during the study or discontinued the study were excluded.

Patient demographics and NYHA class were summarized as n (%) of patients per category or median (IQR). During the SOC and telemonitoring periods, n (%) of patients with at least 1 hospitalization and the mean number of inpatient days per patient (95% CI) were reported. The mean number (95% CI) of visits per patient (primary, secondary, and emergency) and the mean number of calls (primary and secondary care) per patient were also reported for each period. Mean health care costs per patient were reported for each period. The normal distribution of each variable was assessed through visual inspection and the Shapiro-Wilk test. For data found to be nonnormally distributed, differences between SOC and telemonitoring periods were tested using the Wilcoxon signed rank test, and a value of \( P < 0.05 \) was considered statistically significant. The Pearson chi-square test with Yates correction was used for testing the difference between SOC and telemonitoring periods (a binary variable) in the number of patients with at least 1 hospitalization.

**Ethical Considerations**
The ethics committee of the Northern Savonia Hospital District approved the study protocol (1401/2020). The study followed good clinical practice following the Declaration of Helsinki and the laws and regulations applicable in Finland. Patients gave written consent upon recruitment to the study. Participation in the study was voluntary and no financial compensation was awarded for participation.
Results

Study Population and Patient Characteristics
Between December 15, 2020, and March 24, 2021, a total of 50 patients with HF were recruited from the Mikkeli Central Hospital. A total of 7 patients did not complete the study due to their deaths or withdrawals from it. All 43 (86%) patients who completed the 12-month study period were included in the analysis. During the telemonitoring period, 20% (9/43) of the daily weight and blood pressure measurements were missing. The median age of patients was 73 (IQR 66-80) years, 74% (37/50) were male, and 60% (30/50) of patients had NYHA classes III-IV (Table 1).
Table 1. Patient demographics and disease characteristics (n=50).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), median (IQR)</td>
<td>73 (66-80)</td>
</tr>
<tr>
<td>Sex, n (%)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>37 (74)</td>
</tr>
<tr>
<td>Female</td>
<td>13 (26)</td>
</tr>
<tr>
<td>NYHA\textsuperscript{a} score, n (%)</td>
<td></td>
</tr>
<tr>
<td>NYHA class II</td>
<td>20 (40)</td>
</tr>
<tr>
<td>NYHA class III-IV</td>
<td>30 (60)</td>
</tr>
<tr>
<td>Systolic blood pressure (mm Hg), mean (SD)</td>
<td>118 (18)</td>
</tr>
<tr>
<td>Heart rate (beats/min), mean (SD)</td>
<td>72 (11)</td>
</tr>
<tr>
<td>BMI (kg/m\textsuperscript{2}), mean (SD)</td>
<td>27 (6)</td>
</tr>
<tr>
<td>Serum creatinine (µmol/L), mean (SD)</td>
<td>125 (44)</td>
</tr>
<tr>
<td>Pro–B-type natriuretic peptide (ng/L), median (IQR)</td>
<td>3122 (1590-5598)</td>
</tr>
<tr>
<td>Left ventricular ejection fraction (%), mean (SD)</td>
<td>37 (11)</td>
</tr>
<tr>
<td><strong>Etiology of heart failure, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Ischemic cardiomyopathy</td>
<td>15 (30)</td>
</tr>
<tr>
<td>Idiopathic dilated cardiomyopathy</td>
<td>13 (26)</td>
</tr>
<tr>
<td>Hypertensive cardiomyopathy</td>
<td>12 (24)</td>
</tr>
<tr>
<td>Tachycardia cardiomyopathy</td>
<td>&lt;5</td>
</tr>
<tr>
<td>Cytostatic cardiomyopathy</td>
<td>&lt;5</td>
</tr>
<tr>
<td>Valvular cardiomyopathy</td>
<td>&lt;5</td>
</tr>
<tr>
<td>Genetic cardiomyopathy</td>
<td>&lt;5</td>
</tr>
<tr>
<td><strong>Medical history, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>24 (48)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>14 (28)</td>
</tr>
<tr>
<td>Coronary heart disease</td>
<td>17 (34)</td>
</tr>
<tr>
<td>Myocardial infarction</td>
<td>6 (12)</td>
</tr>
<tr>
<td>Atrial fibrillation</td>
<td>31 (62)</td>
</tr>
<tr>
<td>Valvular heart disease</td>
<td>&lt;5</td>
</tr>
<tr>
<td><strong>Medication at recruitment, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Diuretic</td>
<td>46 (92)</td>
</tr>
<tr>
<td>Digitalis</td>
<td>6 (12)</td>
</tr>
<tr>
<td>β-blocker</td>
<td>48 (96)</td>
</tr>
<tr>
<td>Mineralocorticoid antagonist</td>
<td>35 (70)</td>
</tr>
<tr>
<td>ACE\textsuperscript{b}-inhibitor</td>
<td>7 (14)</td>
</tr>
<tr>
<td>Angiotensin receptor-blocker</td>
<td>10 (20)</td>
</tr>
<tr>
<td>Valsartan-sacubitril</td>
<td>30 (60)</td>
</tr>
<tr>
<td>SGLT2\textsuperscript{c}-inhibitor</td>
<td>&lt;5</td>
</tr>
<tr>
<td>Statin</td>
<td>26 (52)</td>
</tr>
<tr>
<td>Anticoagulant</td>
<td>42 (84)</td>
</tr>
<tr>
<td>ASA\textsuperscript{d} or clopidogrel</td>
<td>11 (22)</td>
</tr>
</tbody>
</table>

\textsuperscript{a}NYHA: New York Heart Association.
Significantly fewer patients had an HF hospitalization during the telemonitoring versus SOC period. The number of inpatient days per patient due to HF decreased by 48% during the telemonitoring period (mean 1.2, 95% CI 0.1-2.3 days vs 2.3, 95% CI 1.1-3.6 days with SOC; \(P = .17\)). The number of emergency care visits decreased significantly during the telemonitoring period by 44% (mean 0.7, 95% CI 0.4-1 vs mean 1.3, 95% CI 0.9-1.7 with SOC; \(P = .006\)). Patients with HF made significantly more phone calls to secondary care during the telemonitoring period (mean 8.3, 95% CI 6.6-10 vs mean 2, 95% CI 1.3-2.7 with SOC; 318% increase; \(P = .001\)) and had significantly more primary care visits (mean 4, 95% CI 2.2-5.8 vs mean 2.8, 95% CI 1.7-3.9; 44% increase; \(P = .02\); Table 2).

Table 2. Use of health care per patient in standard of care (SOC) or telemonitoring solution for a 6-month period. Statistics were calculated with the Wilcoxon signed rank test or the Pearson chi-square test with Yates correction for the binary variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>SOC (n=43)</th>
<th>Telemonitoring (n=43)</th>
<th>Absolute change (relative change; %)</th>
<th>(P) value for difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>All-cause hospitalizations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients with (\geq 1) event, n (%)</td>
<td>23 (53)</td>
<td>14 (33)</td>
<td>-9 (-39)</td>
<td>.08</td>
</tr>
<tr>
<td>Inpatient days, mean (95% CI)</td>
<td>2.9 (1.6)</td>
<td>1.7 (1.2)</td>
<td>-1.2 (-41)</td>
<td>.20</td>
</tr>
<tr>
<td>Hospitalizations for cardiovascular cause other than HF(^a)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients with (\geq 1) event, n (%)</td>
<td>6 (14)</td>
<td>&lt;5 (-10)</td>
<td>&lt;5 (-30)</td>
<td>.70</td>
</tr>
<tr>
<td>Inpatient days, mean (95% CI)</td>
<td>0.2 (0.3)</td>
<td>0.02 (0.05)</td>
<td>-0.16 (-88)</td>
<td>.40</td>
</tr>
<tr>
<td>Hospitalizations for HF</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patients with (\geq 1) event, n (%)</td>
<td>20 (47)</td>
<td>6 (14)</td>
<td>-14 (-70)</td>
<td>.002</td>
</tr>
<tr>
<td>Mean inpatient days, days (95% CI)</td>
<td>2.3 (1.3)</td>
<td>1.2 (1.1)</td>
<td>-1.1 (-48)</td>
<td>.17</td>
</tr>
<tr>
<td>Mean number of emergency care visit for cardiovascular cause, n (95% CI)</td>
<td>1.3 (0.4)</td>
<td>0.7 (0.3)</td>
<td>-0.6 (-44)</td>
<td>.006</td>
</tr>
<tr>
<td>Mean number of phone calls to primary care, n (95% CI)</td>
<td>3.9 (1.3)</td>
<td>2.5 (0.9)</td>
<td>-1.4 (-36)</td>
<td>.01</td>
</tr>
<tr>
<td>Mean number of phone calls to secondary care, n (95% CI)</td>
<td>2 (0.7)</td>
<td>8.3 (1.7)</td>
<td>6.3 (+318)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Mean number of all-cause primary care visits, n (95% CI)</td>
<td>2.8 (1.1)</td>
<td>4 (1.8)</td>
<td>1.2 (+44)</td>
<td>.02</td>
</tr>
<tr>
<td>Mean number of secondary care visits (cardiology), n (95% CI)</td>
<td>1.8 (0.5)</td>
<td>1.9 (0.6)</td>
<td>0.1 (+8)</td>
<td>.80</td>
</tr>
</tbody>
</table>

\(^a\)HF: heart failure.

Health Care Costs

Mean hospitalization costs per patient decreased significantly by 49% during the telemonitoring period (mean €1114 vs €2189 with SOC; \(P = .03\); a currency exchange rate of EUR €1=US $1.10589 is applicable), while total health care costs decreased by 33% (mean €2104 vs €3124 with SOC; \(P = .07\); Table 3). The cost of emergency care visits was also significantly lower in the telemonitoring period (mean €209 vs €347 with SOC; 40% decrease; \(P = .009\)), and mean costs per patient for phone calls to secondary care increased significantly (mean €268 vs €114 with SOC; 134% increase; \(P = .001\)) in the telemonitoring period (Table 3).
Table 3. Estimated mean direct health care cost per patient in standard of care (SOC) and in telemonitoring solution, respectively, during a 6-month period (2021). A currency exchange rate of EUR €1=US $1.10589 is applicable. Statistics were calculated with the Wilcoxon signed rank test.

<table>
<thead>
<tr>
<th>Cost category</th>
<th>SOC (n=43), mean cost (€; 95% CI)</th>
<th>Telemonitoring (n=43), mean cost (€; 95% CI)</th>
<th>Absolute change (€; relative change in mean cost; %)</th>
<th>P value for difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hospitalizations^a</td>
<td>2189 (805)</td>
<td>1114 (689)</td>
<td>−1075 (−49)</td>
<td>.03</td>
</tr>
<tr>
<td>Primary care visits</td>
<td>75 (37)</td>
<td>101 (74)</td>
<td>27 (+36)</td>
<td>.30</td>
</tr>
<tr>
<td>Secondary care visits (cardiology)</td>
<td>288 (77)</td>
<td>337 (87)</td>
<td>50 (+17)</td>
<td>.20</td>
</tr>
<tr>
<td>Emergency care visits</td>
<td>347 (122)</td>
<td>209 (99)</td>
<td>−137 (−40)</td>
<td>.009</td>
</tr>
<tr>
<td>Phone calls to primary care</td>
<td>112 (43)</td>
<td>74 (35)</td>
<td>−38 (−34)</td>
<td>.02</td>
</tr>
<tr>
<td>Phone calls to secondary care</td>
<td>114 (42)</td>
<td>268 (68)</td>
<td>153 (+134)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Total cost</td>
<td>3124 (912)</td>
<td>2104 (791)</td>
<td>−1020 (−33)</td>
<td>.07</td>
</tr>
</tbody>
</table>

^aHospitalization from cardiology and internal medicine ward and from primary care.

Discussion

Principal Findings

In this pre-post study of a novel telemonitoring solution for patients with HF in Finland, hospitalization costs were significantly lower during the 6-month telemonitoring period versus the SOC period (from €2189 per patient during SOC to €1114 during telemonitoring). The number of hospitalized patients was significantly lower during telemonitoring (from 20 during SOC to 6 patients), and the mean length of stay decreased from 2.3 days to 1.2 days (not statistically significant). The number of emergency visits and associated costs were also significantly lower during telemonitoring. By contrast, patients with HF had significantly more primary care visits and phone calls to secondary care nurses during telemonitoring versus SOC; however, total health care costs were 33% lower than during SOC (not statistically significant).

The reduction in inpatient days due to HF during telemonitoring was not statistically significant, most probably due to the low number of patients with HF in the study. While secondary care phone calls increased significantly, a substantial part of these were for technical help in using the digital platform at the start of the telemonitoring period. The number of primary care phone calls, on the other hand, decreased during the telemonitoring period, which may be because secondary care nurses were monitoring patients.

Comparison With Previous Work

A comprehensive meta-analysis of recently published studies on telemonitoring has provided evidence that telemonitoring is beneficial in reducing mortality and hospitalizations in patients with HF. However, individual studies show both beneficial and neutral effects of telemonitoring when compared to SOC [11,13,19-21]. The main objective of the studies was to investigate whether remote monitoring can improve the detection of early signs of decompensation and decrease hospitalization and mortality. The variability of the results may be due to differences in the health care system, telemonitoring model, population with HF, and follow-up durations [9,14].

This study found that 47% (20/43) of patients were hospitalized due to HF during the SOC period, versus 14% (6/43) during the telemonitoring period in the Southern Savonia region of Finland. Vuorinen et al [21] conducted a telemonitoring study in Finland from 2010 to 2012. They showed that only 28% of patients were hospitalized in the SOC group and 17% in the remote monitoring group during a 6-month follow-up period. The inclusion criteria for the population with HF included NYHA II-IV but did not require a recent hospitalization, which could explain the lower HF hospitalization risk compared with this study. Our data align with previous studies showing that nearly half of patients with HF are rehospitalized within 6 months after discharge [21,22]. However, further studies using a similar population with HF are needed to confirm these findings.

Similar to this study, Vuorinen et al [21] also found a nonstatistically significant decrease in inpatient days with telemonitoring (mean 0.7 vs 1.4 days with SOC). The significant reduction in hospitalization-related costs and the number of patients hospitalized due to HF in this study support the idea that telemonitoring reduces hospitalizations.

A Spanish telemonitoring trial (n=117) had similar findings to this study. In this trial, 50% of the patients were hospitalized in the SOC group versus 28% in the telemonitoring group over a 6-month follow-up period. The patients were enrolled in the study upon hospitalization [19]. Thus, the results from this study support our findings on the benefits of telemonitoring in reducing hospitalizations. However, a large Better Effectiveness After Transition–Heart Failure (BEAT-HF) trial (n=1437) conducted in California could not see a reduction in readmissions in patients with HF in a telemonitoring group compared to SOC during a 6-month follow-up period [20]. The BEAT-HF trial’s limitations were that patients were recruited from academic medical centers, which may restrict the generalizability of the results, as most patients with HF do not receive care in academic medical centers. Upon receiving alerts, nurses advised patients to contact the physicians, or nurses called the physicians, but physicians were not directly involved with the interventions. Thus, the monitoring may not have affected care in practice. In this study and the Spanish study, nurses and treating physicians collaborated upon receiving alerts, which may have increased the benefits of telemonitoring. For example, in this study, physicians reacted to patient alerts within 24 hours.

There are several studies on the effect of telemonitoring on hospitalization and mortality, but only a few studies on costs.

https://mhealth.jmir.org/2024/1e51841 Jmir Mhealth Uhealth 2024 | vol. 12 | e51841 | p.797 (page number not for citation purposes)
This study estimated the health care cost related to resource use using real-word data and showed that health care costs were 33% (€2104/€3124) lower during telemonitoring versus SOC. As expected, most of the cost reductions originated from reduced hospitalizations. A cost-effectiveness study in a Danish telemonitoring trial (n=274) reported similar results by showing that telemonitoring reduced total health care costs by 35% versus SOC with a 1-year follow-up [16]. In a Spanish telemonitoring trial, Comin-Colet et al [15] found a total cost reduction of 45% with telemonitoring versus SOC, with 178 patients and a 6-month follow-up. A Belgian Telemonitoring in the Management of Heart Failure (TEMA-HF) study (n=160) found a 27% cost reduction (not statistically significant) with telemonitoring versus SOC during a 6-month follow-up [23]. The German Heart Failure II trial (TIM-HF2) showed an 18% reduction in annual costs per patient in the telemonitoring group compared to the SOC group during a 1-year follow-up [18]. These studies support our conclusion that telemonitoring may result in substantial cost savings in HF care. To justify reimbursement for telemonitoring, studies are needed on the cost-effectiveness of large-scale telemonitoring for decision makers. Furthermore, a cost-effective telemonitoring model applicable to different health care systems and settings needs to be developed.

**Strengths and Limitations**

A strength of this study was the cost analysis, which included both HF-related health care costs as well as other costs accrued during the follow-up period. The study had some limitations. The study was conducted in a single region, the Southern Savonia region, which may limit the generalizability of the results. However, the study population is representative of the region, as all patients are directed to the same central hospital where recruitment was done. A randomized controlled trial design was not feasible due to the limited number of suitable patients with HF for remote monitoring. No patients with HF were included from other health care districts, as divergent monitoring practices could potentially bias the analysis results. Patients were not randomized, and patients needed to be able to use the digital platform, which may have resulted in a possible selection bias. As the pre-post design uses a historical control group (ie, patients on SOC in the period before starting telemonitoring), the underlying assumption in the analysis, given the deteriorating nature of HF, is that health care use in the absence of telemonitoring would remain at least at the same level as during SOC. Follow-up with telemonitoring was limited to 6 months, and it is unclear how use of health care services would develop beyond this period. Finally, due to the small patient numbers, the absence of a control group, and the 6-month follow-up period, it was not feasible to conduct mortality analyses.

The following must be considered when generalizing our results and applying our telemonitoring solution to other health care systems: our telemonitoring solution was applied to a patient population with a high risk of readmission due to a recent hospital admission and NYHA class II-IV. Other patient characteristics considered were the mean age (73 years), male proportion (37/50, 74%), proportion of patients in NYHA class II-IV (30/50, 60%), and proportion of patients with at least 1 admission within 6 months (20/50, 47%). These patient characteristics were comparable to those of other reported telemonitoring study populations [11,13,18,19].

**Conclusions**

In conclusion, our results suggest that the novel telemonitoring solution can help reduce hospital admissions and hospitalization costs as well as total health care costs in a population with HF with a recent hospital admission in the past 12 months.

**Acknowledgments**

The authors would like to thank Saara Liukkonen, Kiia Pätsi, and Jenni Tikkala for their valuable contributions.

**Conflicts of Interest**

PM, JK, and J-PL are consultants for Roche Diagnostics. R-LL, EH, and PP receive consulting fees paid to their employer by Roche Diagnostics. KK and SV are employees of Roche Diagnostics. AJ is an employee of Essote.

Multimedia Appendix 1

Definition of the alert triggers.

[DOCX File , 16 KB - mhealth_v12i1e51841_app1.docx ]

**References**


Abbreviations

BEAT-HF: Better Effectiveness After Transition–Heart Failure
HCP: health care professional
HF: heart failure
ICD-10: International Classification of Diseases, Tenth Revision
NYHA: New York Heart Association
SOC: standard of care
TEMA-HF: Telemonitoring in the Management of Heart Failure
TIM-HF2: German Heart Failure II trial
Correction: The Importance of Activating Factors in Physical Activity Interventions for Older Adults Using Information and Communication Technologies: Systematic Review

Ellen Bentlage, MSc; John Jnr Nyamadi, BSc; Rosemary Dubbeldam, PhD
Department of Movement Science, Institute of Sport and Exercise Sciences, University of Münster, Münster, Germany

Corresponding Author:
Ellen Bentlage, MSc

Related Article:
https://mhealth.jmir.org/2023/1/e42968

In “The Importance of Activating Factors in Physical Activity Interventions for Older Adults Using Information and Communication Technologies: Systematic Review” (JMIR Mhealth Uhealth 2023; 11(1): e42968), the authors noted some errors.

In the “Conclusions” section of the Abstract, the following phrase:

So far, only a limited number of available BCTs (21/99, 21%) have been integrated.

Has been changed to:

So far, only a limited number of available BCTs (21/102, 21%) have been integrated.

In the “Introduction” subsection “Activation Using Information and Communication Technology,” the following sentence:

Dugas and colleagues [12] added 2 more categories (ie, gamification and personalization).

Has been changed to:

Dugas and colleagues [12] added 2 more categories (ie, personalization and gamification), including 9 BCTs in total.

In the “Methods” subsection “Selection of Studies and Data Extraction,” the line:

Additionally, the two categories suggested by Dugas and colleagues [12]—gamification and personalization—were incorporated as categories 17 and 18.

Has been changed to:

Additionally, the two categories suggested by Dugas and colleagues [12]—personalization and gamification—were incorporated as categories 17 and 18.

In Table 1, under the “Skills (ability)” column, the following text:

• Category 7: Repetition and substitution

Has been changed to:

• Category 8: Repetition and substitution

Elsewhere in Table 1, under the “Knowledge (awareness)” column, the following content:

• Category 5: Natural consequences
• Category 9: Comparison of outcomes
• Category 11: Regulation
• Category 14: Scheduled consequences
• Category 16: Covert learning

Has been changed to:

• Category 5: Natural consequences
• Category 9: Comparison of outcomes
• Category 11: Regulation

In Table 1, under the “Motivation (triggers)” column, the following text:

• Category 1: Goals and planning
• Category 2: Feedback and monitoring
• Category 3: Social support
• Category 8: Associations
• Category 10: Reward and thread
• Category 12: Antecedents
• Category 17: Gamefication
• Category 18: Personalization

Has been changed to:

• Category 1: Goals and planning
• Category 2: Feedback and monitoring
• Category 3: Social support
• Category 7: Associations
• Category 10: Reward and thread
• Category 12: Antecedents
• Category 14: Scheduled consequences
• Category 16: Covert learning
• Category 17: Personalization
• Category 18: Gamification
In the “Results” subsection “Delivering Activation Factors,” the following sentence:

Of the 22 available BCTs [11,12] that target skills, the aforementioned 5 (23%) were used among all included articles.

Has been changed to:

Of the 23 available BCTs [11,12] that target skills, the aforementioned 6 (23%) were used among all included articles.

Further in the same section, the following paragraphs:

Others were prompts and cues (4 intervention groups), social comparison (1 intervention group), and problem-solving discussions for finding ways to overcome barriers (1 intervention group). Of the 64 available BCTs [11,12] that target motivation, 12 (19%; inclusive of the 7 aforementioned BCTs) were used among all included articles.

Altogether, of the 99 potential BCTs [11,12], 21 (21%) were integrated in all articles.

Have been changed to:

Others were prompts and cues (4 intervention groups), adjustment of intervention content to the performance (4 intervention groups), action planning (2 intervention groups) and problem-solving discussions for finding ways to overcome barriers (1 intervention group). Of the 66 available BCTs [11,12] that target motivation, 11 were used among all included articles.

Altogether, of the 102 potential BCTs [11,12], 21 (21%) were integrated in all articles.

Within Textbox 1, the “Skills” subheading which previously appeared as:

Skills
• Promoted in 19 of 20 interventions
• Five BCTs were used; main behavior change techniques (BCTs) were the instruction of optimal behavior performance and the demonstration of behavior

Has been changed to:

Skills
• Promoted in 19 of 20 interventions
• Six BCTs were used; main behavior change techniques (BCTs) were the instruction of optimal behavior performance and the demonstration of behavior

Further within Textbox 1, the “Motivation” subheading previously read as:

Motivation
• Promoted in 17 of 20 interventions
• Twelve BCTs were used; main BCTs were self-monitoring, feedback on behavior, social support, and goal setting

And will now appear as:

Motivation
• Promoted in 17 of 20 interventions
• Eleven BCTs were used; main BCTs were self-monitoring, feedback on behavior, social support, and goal setting

Within the “Discussion” subsection “Delivering Activation Factors,” the sentence:

To target participants’ motivation, 12 BCTs, such as self-monitoring, feedback on behavior, social support, and goal setting, were used.

Was changed to:

To target participants’ motivation, 11 BCTs, such as self-monitoring, feedback on behavior, social support, adjusting intervention content to the performance and goal setting, were used.

The “Discussion” subsection “Conclusions” includes the claim that:

Although a broad variety of BCTs were used in the articles, they were limited to about 21% (21/99) of available BCTs.

This has been adjusted to appear as:

Although a broad variety of BCTs were used in the articles, they were limited to about 21% (21/102) of available BCTs.

In Multimedia Appendix 1, the rows “Li, 2020,” “Mansson, 2020,” “Rowley, 2019,” and “Van Dyck, 2016” have been amended from:

18 Personalization

And now appear as:

17.4 Adjusting intervention content to performance

Finally, Multimedia Appendix 2 has been changed to match the main manuscript in the ways listed below.

The “Skills” column has been changed from:

Skills (30 BCTs)
5 BCTs
To read as:
Skills (23 BCTs)
6 BCTs

The “Knowledge” column has been changed from:

Knowledge (27 BCTs)
4 BCTs
To read as:
Knowledge (13 BCTs)
4 BCTs

The “Motivation” column has been changed from:

Motivation (50 BCTs)
12 BCTs
To appear as:
Motivation (66 BCTs)

11 BCTs

The correction will appear in the online version of the paper on the JMIR Publications website on February 1, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Multimedia Appendix 1
Study characteristics.
[DOCX File, 67 KB - mhealth_v12i1e55486_app1.docx]

Multimedia Appendix 2
Behaviour change techniques overview corrected.
[DOCX File, 21 KB - mhealth_v12i1e55486_app2.docx]

Submitted 14.12.23; this is a non–peer-reviewed article; accepted 14.12.23; published 01.02.24.

Please cite as:
Bentlage E, Nyamadi JJ, Dubbeldam R
Correction: The Importance of Activating Factors in Physical Activity Interventions for Older Adults Using Information and Communication Technologies: Systematic Review
JMIR Mhealth Uhealth 2024;12:e55486
URL: https://mhealth.jmir.org/2024/1/e55486
doi:10.2196/55486
PMID:38329098

© Ellen Bentlage, John Jnr Nyamadi, Rosemary Dubbeldam. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 1.2.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Corrigenda and Addenda

Correction: Mobile App Intervention of a Randomized Controlled Trial for Patients With Obesity and Those Who Are Overweight in General Practice: User Engagement Analysis Quantitative Study

Vera Helen Buss¹*, PhD; Margo Barr¹*, PhD; Sharon M Parker¹*, MPH; Alangir Kabir¹*, PhD; Annie Y S Lau²*, PhD; Siaw-Teng Liaw³*, PhD; Nigel Stocks³*, MD; Mark F Harris¹*, MD

¹Centre for Primary Health Care and Equity, University of New South Wales, Sydney, Australia
²Australian Institute of Health Innovation, Macquarie University, Sydney, Australia
³School of Population Health, University of New South Wales, Sydney, Australia
⁴Adelaide Medical School, University of Adelaide, Adelaide, Australia
*all authors contributed equally

Corresponding Author:
Margo Barr, PhD
Centre for Primary Health Care and Equity
University of New South Wales
AGSM Building
High Street, Kensington Campus
Sydney, 2052
Australia
Phone: 61 290656041
Email: Margo.barr@unsw.edu.au

Related Article:
Correction of: https://mhealth.jmir.org/2024/1/e45942
doi:10.2196/58507

In “Mobile App Intervention of a Randomized Controlled Trial for Patients With Obesity and Those Who Are Overweight in General Practice: User Engagement Analysis Quantitative Study” (JMIR Mhealth Uhealth 2024;12:e45942), the authors noted two errors. The following corrections have been made:

Under “Methods”, the third paragraph of the “Outcome Measures” subsection has been changed from:

Specifically, the Health Literacy Questionnaire domain 8 questions were the following [11]: please indicate how difficult or easy the following tasks are for you now: (1) find information about health problems; (2) find health information from several different places; (3) get information about health so you are up to date with the best information; (4) get health information in words you understand; and (5) get health information by yourself. There is a 5-point response option scale for each question (cannot do or always difficult, usually difficult, sometimes difficult, usually easy or always easy). The scores are reported as averages for the domain (with a range between 1 and 5) with high scores representing higher health literacy.

And now reads as follows:

The Health Literacy Questionnaire domain 8 questions were used [11]: (1) find information about health problems; (2) find health information from several ... ; (3) get information about health so you...; (4) get health information in words you...; and (5) get health information by yourself. There is a 5-point response option scale for each question (cannot do or always difficult, usually difficult, sometimes difficult, usually easy or always easy). The scores are reported as averages for the domain (with a range between 1 and 5) with high scores representing higher health literacy.

Additionally, in Table 1 the definition of the “Health literary” row has been changed from:

Health literacy, specifically the self-reported ability to find good quality health information, according to domain 8 of the Health Literacy Questionnaire [9], at baseline and 6-month follow-up.

To reference a different source, this now appears as:

Health literacy, specifically the self-reported ability to find good quality health information, according to domain 8 of the Health Literacy Questionnaire [11], at baseline and 6-month follow-up.
The correction will appear in the online version of the paper on the JMIR Publications website on April 2, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 18.03.24; this is a non–peer-reviewed article; accepted 22.03.24; published 02.04.24.

Please cite as:
Buss VH, Barr M, Parker SM, Kabir A, Lau AYS, Liaw ST, Stocks N, Harris MF
Correction: Mobile App Intervention of a Randomized Controlled Trial for Patients With Obesity and Those Who Are Overweight in General Practice: User Engagement Analysis Quantitative Study
JMIR Mhealth Uhealth 2024;12:e58507
URL: https://mhealth.jmir.org/2024/1/e58507
doi: 10.2196/58507
PMID:

©Vera Helen Buss, Margo Barr, Sharon M Parker, Alamgir Kabir, Annie Y S Lau, Siaw-Teng Liaw, Nigel Stocks, Mark F Harris. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 02.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Trial Participants’ Perceptions of the Impact of Ecological Momentary Assessment on Smoking Behaviors: Qualitative Analysis

Elizabeth R Stevens¹, MPH, PhD; Rina Li¹; Grace Xiang¹, BA; Rachel Wisniewski¹, BA; Sidney Rojas¹, BA; Katherine O’Connor¹, MS; Olivia Wilker¹, MPH; Mahathi Vojjala¹2, MPH; Omar El-Shahawy¹2, MPH, MD, PhD; Scott E Sherman¹, MPH, MD

¹ Corresponding Author: Elizabeth R Stevens, MPH, PhD

Abstract

Background: Ecological momentary assessment (EMA) is an increasingly used tool for data collection in behavioral research, including smoking cessation studies. As previous addiction research suggests, EMA has the potential to elicit cue reactivity by triggering craving and increasing behavioral awareness. However, there has been limited evaluation of its potential influence on behavior.

Objective: By examining the perspectives of research participants enrolled in a tobacco treatment intervention trial, this qualitative analysis aims to understand the potential impact that EMA use may have had on smoking behaviors that may not have otherwise been captured through other study measures.

Methods: We performed a qualitative analysis of in-depth interviews with participants enrolled in a pilot randomized controlled trial of a tobacco treatment intervention that used SMS text messaging to collect EMA data on smoking behaviors. In the pilot randomized controlled trial, combustible cigarette and e-cigarette use and smoking-related cravings were measured as part of an EMA protocol, in which SMS text messaging served as a smoking diary. SMS text messaging was intended for data collection only and not designed to serve as part of the intervention. After a baseline assessment, participants were asked to record daily nicotine use for 12 weeks by responding to text message prompts that they received 4 times per day. Participants were prompted to share their experiences with the EMA text messaging component of the trial but were not directly asked about the influence of EMA on their behaviors. Transcripts were coded according to the principles of the framework for applied research. The codes were then examined, summarized, and grouped into themes based on the principles of grounded theory.

Results: Interviews were analyzed for 26 participants. The themes developed from the analysis suggested the potential for EMA, in the form of an SMS text messaging smoking diary, to influence participants’ smoking behaviors. The perceived impacts of EMA text messaging on smoking behaviors were polarized; some participants emphasized the positive impacts of text messages on their efforts to reduce smoking, while others stressed the ways that text messaging negatively impacted their smoking reduction efforts. These contrasting experiences were captured by themes reflecting the positive impacts on smoking behaviors, including increased awareness of smoking behaviors and a sense of accountability, and the negative impacts on emotions and smoking behaviors, including provoking a sense of guilt and triggering smoking behaviors.

Conclusions: The collection of EMA smoking behavior data via SMS text messaging may influence the behaviors and perceptions of participants in tobacco treatment interventions. More research is needed to determine the magnitude of impact and mechanisms, to account for the potential effects of EMA. A broader discussion of the unintended effects introduced by EMA use is warranted among the research community.

(JMIR Mhealth Uhealth 2024;12:e52122) doi:10.2196/52122

KEYWORDS

smoking behavior; ecological momentary assessment; bias; behavioral impact; smoking; smoker; qualitative analysis; pilot study; tool; data collection; tobacco; text message; accountability; mHealth; mobile health; message; trigger; cigarette
Introduction

Ecological momentary assessment (EMA) is a data collection method that is increasingly being used in health and behavioral sciences [1,2]. EMA has been shown to be a useful tool for measuring behaviors associated with substance use [3]. Consequently, there has been a strong interest in the use of EMA in smoking cessation studies [4]. However, as indicated in previous addiction research, EMA has the potential to elicit cue reactivity by triggering craving and increasing behavioral awareness [5-8]. As such, there is potential for the use of EMA to create an assessment effect and inadvertently influence behaviors in some settings.

Assessment effects refer to the phenomenon in which the outcome of interest (e.g., a behavior) is modified simply by assessing it, arising due to the assessment method or its interaction with the intervention [9,10]. The frequent prompts for self-reporting that are often required by EMA might inadvertently influence participants’ behaviors by, among other mechanisms, increasing their self-awareness, altering an emotional response, or serving as reminders of the behavior [2,8,11,12]. Depending on the behavior of interest, these EMA consequences may differentially impact outcomes. These assessment effects can have a significant impact on the interpretation of trial results but are rarely considered in trial design [13].

Despite its potential importance to the interpretation of research results, there are a limited number of studies investigating the potential for EMA to produce an assessment effect. Within the research that does exist, there have been mixed results reported, with some studies reporting no impact on behavior [11,14,15] and others indicating that EMA likely has an impact [2,5-8,11,12,14]. For this reason, it is important to understand if and how the use of EMA data collection in smoking research could impact smoking behaviors.

To better understand the potential impact of using EMA to measure smoking behaviors, we performed a qualitative analysis of in-depth interviews with participants in a randomized controlled trial (RCT) pilot study of a smoking intervention that used SMS text messaging–based EMA as a data collection strategy [16]. By examining the perspectives of the trial participants, this qualitative analysis aims to understand the potential impact that the use of EMA may have had on smoking behaviors that may not have otherwise been captured through other means of data collection.

Methods

Study Design

For the purposes of assessing the acceptability of a smoking intervention and determining points for potential program improvement, we performed a qualitative analysis of in-depth interviews conducted with participants of an RCT pilot study that compared the effectiveness of behavioral counseling and the use of e-cigarettes on smoking outcomes to that of behavioral counseling and nicotine replacement therapy (NRT).

Ethical Considerations

The interviews and analysis procedures were reported in accordance with the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist (Checklist 1), as applicable [17]. The study protocol was approved by the New York University Langone Health Institutional Review Board (approval number i20-00839), and written documentation of informed consent was received prior to starting data collection. Participants were provided with a US $20 incentive for their participation. Participant data were maintained on a secure server. After transcription, all participant data were deidentified prior to data analysis.

Setting and Participants

Interview participants were recruited upon completion of the intervention phase of the RCT at the 12-week follow-up study visit. All RCT participants were invited to participate in an in-depth interview to discuss their experiences with the intervention and other aspects of the trial. Interview recruitment ended once thematic saturation was reached. Interviews lasted approximately 30 minutes and were performed between April 2021 and November 2022.

The pilot RCT was performed to determine the feasibility and acceptability of an e-cigarette–based smoking intervention and to compare the effectiveness of counseling and e-cigarette use on smoking outcomes to that of counseling and NRT [16]. Text message–based EMA data collection was used to record smoking patterns. Patients from the electronic health record system of New York University Langone Health—a private hospital system serving New York, New Jersey, and Connecticut—were recruited as RCT participants. The RCT participant sample was initially restricted to patients with a diagnosis of chronic obstructive pulmonary disease, but the scope was later expanded to include patients with a diagnosis of coronary artery disease, peripheral artery disease, or asthma. In addition, to be eligible, RCT participants were required to smoke ≥4 days per week, with at least 5 cigarettes smoked on the days that participants did smoke; be motivated to quit smoking; and possess a phone with SMS text messaging capabilities. A total of 121 participants were recruited into the pilot RCT.

RCT EMA Protocol

In the pilot RCT, combustible cigarette use and e-cigarette use, as well as smoking-related cravings, were measured as part of an EMA protocol, in which SMS text messaging served as a smoking diary. The SMS text messaging was intended for data collection only and not designed to serve as part of the intervention. After a baseline assessment, participants were asked to record their daily nicotine use for 12 weeks by responding to text message prompts that they received 4 times per day. The coverage design prompted participants to provide brief check-in reports via text message over the course of each day, wherein they were asked to report combustible cigarette and e-cigarette use based on their study arm (Figure 1). Responses to 1-item measures of cigarette craving and satisfaction were also collected from each report.
Data Collection

Semistructured, in-depth telephone interviews, which lasted approximately 30 minutes, were conducted by research staff (OW and MV). Interviews were audio-recorded, transcribed, and imported into Dedoose software (SocioCultural Research Consultants LLC) for qualitative data analyses.

The interview guide covered topics that were designed for the following goal: gaining a deeper understanding of the participants’ experiences, intervention satisfaction, attitudes toward e-cigarettes, and intentions to quit. The interviews were intended to assist with the further adaptation of the e-cigarette smoking intervention and behavioral counseling manual used in the RCT. The interview guide was developed based on the pilot RCT procedures to provide additional insights about the barriers and facilitators of e-cigarette use and how to refine the current approach to enhance program retention and outcomes. Interviews covered topics such as program aspects that the participants liked or disliked, features of the intervention that should be modified, participants’ experiences with using e-cigarettes or NRT, intentions of using e-cigarettes after the intervention, and whether participants’ health symptoms interfered with their ability to engage in the intervention. The participants were prompted to share their opinions on the EMA texting component of the trial but were not directly asked about the influence of the EMA on their behaviors. Participants were prompted to discuss the texting in the interview, as follows: “During the program, you answered questions over text on daily
basis—What was that experience like for you?” This was followed by probes, including “How may texting have affected your overall experience with the program?” A copy of the interview script is available in Multimedia Appendix 1.

Data Analysis

Interview transcripts were coded by using procedures that were designed to ensure thoroughness and reliability. We used Dedoose software to manage the data and coded the data according to the principles of the framework for applied research [18], which consists of the following 5-stage process: familiarization, identifying themes, indexing, charting, and interpretation. Codes were primarily developed a priori based on intervention components and the quality improvement goals of this study. Additional codes were developed by reviewing a random sample of interviews and via discussion with the coding team. The general development of themes arose from the data, using the principles of grounded theory [19]. To enhance reliability, 2 researchers took part in the coding and analysis process for each interview. Prior to full coding, a random sample of interviews was double coded, and intercoder reliability was assessed based on percent agreement (range 79.5%-87.9%). Afterward, all disagreements were discussed as a group to improve concordance among coders. All interviews were then independently coded by 5 coders (GX, KO, RL, RW, and SR) who worked in various pair combinations; each pair met with the other coders and a coauthor (ERS) to resolve discrepancies. When coding was completed, the quotations for each code were examined, summarized, and grouped together into themes.

Results

Participants

Interviews were performed with a total of 27 participants. Due to an audio malfunction, 1 interview was not included in the analysis. As such, 14 participants were in the e-cigarette study arm and 12 were in the NRT study arm. The average age of the participants included in the analysis was 57.1 (range 28-74) years; 54% (14/26) of participants identified as female, 46% (12/26) were White, 31% (8/26) were Black, 15% (4/26) were Hispanic, and 4% (1/26) were Asian. The majority (17/26, 65%) of participants had completed at least some college; 35% had a high school education or less. Around two-thirds (17/26, 65%) of participants had a diagnosis of chronic obstructive pulmonary disease, 15% (4/26) of participants were diagnosed with coronary artery disease or peripheral artery disease, and 19% (5/26) of participants were diagnosed with asthma.

Themes

Overview of Themes

Without being directly prompted, 18 of the 26 participants described EMA impacting their behaviors or emotions, and several themes indicating a potential, inadvertent influence of EMA on smoking behaviors emerged. The perceived impacts of EMA texting on smoking behaviors were polarized; some participants emphasized the positive impacts of the text messages on their efforts to reduce smoking, while others stressed the ways in which the texting negatively impacted their smoking reduction efforts. These contrasting experiences were captured by 2 major themes and their subthemes, reflecting the positive impacts on smoking behaviors, including increased awareness of smoking habits and a sense of accountability, and the negative impacts on emotions and smoking behaviors, including provoking a sense of guilt and triggering smoking behaviors (Table 1).

Table 1. Major themes and subthemes.

<table>
<thead>
<tr>
<th>Themes and subthemes</th>
<th>Frequency, n</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMA\textsuperscript{a} texting has a positive impact on smoking behaviors</td>
<td>16</td>
</tr>
<tr>
<td>EMA texting serves as a source of accountability</td>
<td>13</td>
</tr>
<tr>
<td>Anticipation of the next text message serves as a deterrent to impulsive smoking</td>
<td>6</td>
</tr>
<tr>
<td>Texting prompts increase awareness of smoking habits</td>
<td>16</td>
</tr>
<tr>
<td>Reminders of the goal to reduce cigarette smoking</td>
<td>11</td>
</tr>
<tr>
<td>Check-ins serve as markers of progress made toward quitting</td>
<td>5</td>
</tr>
<tr>
<td>Negative impact of EMA texting on emotions and smoking behaviors</td>
<td>7</td>
</tr>
<tr>
<td>Repeated text messages asking about smoking behaviors produce negative emotions</td>
<td>4</td>
</tr>
<tr>
<td>Text messages inquiring about cigarette use may have a triggering effect</td>
<td>5</td>
</tr>
</tbody>
</table>

\textsuperscript{a}EMA: ecological momentary assessment.
**Perceived Positive Impacts of EMA Texting Procedures**

**Overview of Positive Impacts**
A major theme that arose was the perceived positive impacts of the EMA methods on efforts to reduce smoking, as many (n=16) participants perceived the SMS text messaging smoking diary as an important component of the tobacco treatment intervention and their experience during their efforts to reduce their cigarette smoking. Participants described the EMA text messages as helpful because “you could use [them] for yourself as a tool” (Participant E09), and while stating that “the text messages are a pain in the butt,” a participant thought that “they were very useful….A useful pain in the butt” (Participant E04). Within this overarching theme—the positive influence of EMA—four subthemes that highlighted the potential roles of EMA text messages emerged: EMA text messages (1) increase awareness of smoking habits, (2) serve as reminders of smoking goals, (3) mark progress made, and (4) provide a sense of accountability.

**Texting Prompts Increase Participants’ Awareness of Their Smoking Habits**
Many (n=16) participants noted that the EMA text messages helped them with their smoking behaviors because the act of recording the number of cigarettes smoked increased their awareness of their smoking behaviors. This theme—increased smoking habit awareness—encompassed the following three layers: (1) awareness of the number of cigarettes smoked and smoking habits, (2) reminders of the goal to reduce cigarette smoking, and (3) markers of progress made.

The predictable and repeated EMA text message requests for participants to report the number of cigarettes smoked prompted reflection on their smoking habits. By engaging in the conscious effort of quantifying the number of cigarettes smoked, the participants heightened their self-awareness and gained a clearer understanding of the frequency of their cigarette use. One participant remarked:

> [The text message would say] “…have you smoked? When was the last time you smoked your last cigarette?” So, it helped me to be aware of how many cigarettes I was smoking per day. [Participant N24]

With regard to the act of recording cigarette use, a participant noted:

> …you’re so aware with the…text messages…of just how often you smoke….Because with the text, I was physically writing it and seeing it. [Participant E27]

**Reminders of the Goal to Reduce Smoking**
For some participants (n=11), the EMA text messages also served as reminders of their goal to reduce or quit smoking and heightened their sense of purpose and determination. One participant said, “It [the text messaging] does remind you what you’re supposed to be focused on” (Participant N20), and in this way, the messages served as frequent reminders of participants’ intentions to change their smoking behaviors. The messages also invigorated their commitment to achieving these goals; one participant said:

> It [the text messages] kept me going. It kept me wanting to quit, you know, and to keep doing it, to keep at the program. [Participant N36]

The text messages also kept participants feeling involved:

> …[questions] like, “how many cigarettes have you smoked?”…kind of keep you involved in it instead of letting it go on the way side. Like it would kind of keep reminding you that…this is what you’re working on, you know? [Participant N50]

**Markers of Progress Made Toward Smoking Reduction**
The EMA text messages served as markers of progress made toward reducing smoking (n=5). By regularly quantifying and reporting their smoking behaviors, participants saw tangible evidence and took note of the accomplishments resulting from their efforts. By tracking progress over time, participants observed patterns of improvement, which reinforced their motivation to continue cutting back on smoking. When describing why they found the text messages useful, one participant said:

> …it’s like a progress type thing. So, I enjoyed the texts…it kind of gave me a reminder [of my progress] because as I went on, I had less and less craving. [Participant N58]

**EMA Texting Serves as a Source of Accountability for Progress Toward Reducing Cigarette Smoking**
Many participants (n=13) believed that the EMA text messages had a positive effect on their sense of accountability in their efforts to reduce smoking. Knowing that they would receive inquiries about their smoking habits increased participants’ mindfulness in reducing their smoking. One participant stated:

> I loved it [the text messaging] because it kept me...accountable...because it kept asking the questions over and over again. In the beginning, it was like a little stressing because I was like, “Oh, my God, these messages, I don’t want to deal with it.” But, it kept me accountable. And it was good. [Participant N36]

Similarly, another participant remarked:

> I had to answer for all the cigarettes I smoked today and hold myself accountable. I thought that was ingenious. [Participant N06]

In some instances (n=6), the text messages served as deterrents to impulsive smoking. One participant shared:

> [There] were a couple of times when they [cigarettes] were right on hand… I’m upstairs, I’m getting ready to light and then boom [sic], oh, “did you smoke today? How many times did you smoke since we last communicated?” [Participant N31]

Similarly, the anticipation of upcoming EMA messages served as a motivator for refraining from smoking or delaying the next cigarette. The participant went on to describe their morning routine:

> …while you’re trying to wake up and organize yourself, you pick up the cigarette…[but] you know...
you’re going to get a text at 9 o’clock saying, “did you smoke?” right? So that notion pops in your head, so you don’t smoke right away. [Participant N31]

Having the knowledge that there was a regular time for reporting their smoking activities encouraged participants to make more deliberate choices regarding their smoking behaviors. One participant discussed how they used EMA check-ins to consciously reduce their cigarette consumption during the time leading up to the EMA prompt:

I knew at a certain time, I was going to get this text. So, when I went to the text, I wanted to have everything in line. I wanted it to be right. So therefore, I would only smoke four cigarettes because that was the allotted cigarettes that I was supposed to smoke at the time. During that time, I would only smoke three cigarettes because I had cut down to three during that period of time and I looked forward to doing it. [Participant N04]

EMA Text Messages Can Negatively Impact Participants’ Emotions and Smoking Behaviors

Overview of Negative Impacts

A second major theme that emerged from the data was the potential negative effects of EMA texting could have on participants. Although many participants reported experiencing positive effects of EMA on their smoking behaviors, some participants (n=7) reflected on the potential negative impacts of the EMA text messages. This theme—the negative effects of EMA—was further distilled into the following two subthemes: the potential roles of EMA in (1) producing negative emotions and (2) triggering cigarette cravings.

Repeated Text Messages Asking About Smoking Behaviors Produce Negative Emotions

Some participants (n=4) mentioned feeling “bad” when they had to admit to smoking or experiencing relapses via the text messages. However, some respondents admitted that the guilt experienced as a result of the text messages, while being a negative emotion, reinforced their determination and prompted them to make renewed efforts, with one participant stating that the text messages “helped me because when I wrote how…I smoked a cigarette…I kind of felt bad” (Participant N58). Moreover, other participants emphasized the negative emotions and guilt experienced when a spotlight was focused on their perceived failures. One participant described the experience by saying, “being able to tell about my progress, or lack thereof, I would have felt bad if I had a slip up” (Participant N21).

Text Messages Inquiring About Cigarette Use May Have a Triggering Effect

Several participants (n=5) expressed their desire to remove the texting component of the program due to its triggering effect on their smoking. Some participants reported that the text messages acted as triggers for cigarette use, as the text messages reminded them of smoking, thereby eliciting an urge to smoke. One participant said that the EMA “was a reminder, actually;” and “…with the reminder came in the struggles” (Participant N20). Notably, another participant said:

Discussion

Principal Findings

This study indicates that smoking intervention participants perceive the collection of EMA smoking behavior data via SMS text messaging as a potential influence on their smoking behaviors. The themes developed from the analysis revealed that EMA, in the form of an SMS text messaging smoking diary, may be perceived as a source of accountability for smoking reduction but may also be a trigger for cigarette use among some people. The results of this study emphasize the need to examine the potential influence of EMA data collection techniques on participants’ behaviors within smoking interventions, as well as in other behavioral research.

The perceived impacts of EMA on smoking behaviors are consistent with previously made observations that the act of receiving EMA prompts can increase behavioral awareness and act as a trigger for craving [6-8], in addition to altering participants’ moods [2,12]. This suggests that EMA for data collection purposes has the potential to unintentionally create an assessment and intervention effects in itself. Although a lack of EMA impact on behavior has been reported in some studies [11,14,15], these studies may be limited by the choice of measures used. As seen in suicide research, EMA prompts have been observed to have an effect on some measures, such as mood, but a minimal effect on other measures, such as suicidal ideation [2]. Therefore, when designing a study, it is important for researchers to reflect on the various factors that may influence the behaviors of interest and consider the potential effect that EMA may have on these factors, in addition to the primary outcomes of interest. Due to its potential effect on participants, investigators should consider and discuss the potential for a behavioral influence to be introduced into a study through the use of EMA data collection.

There is a need for further investigation into the ways that and the degree to which EMA affects participants. Within the EMA literature, there is a general lack of discussion around the effects of EMA on participants’ behaviors. When seeking to improve EMA methods, focus is often placed on participant retention and the validity of the data collection method [20-23], with little to no discussion on the potential behavioral impact of EMA. Indeed, when discussing strengths and limitations of EMA, a large portion of the EMA study literature discusses and reports measures of adherence to and reliability of EMA data collection [20-23], with few studies exploring the potential limitation of EMA in which the data collection itself may affect behaviors of interest [2,5-8,11,12,14,15]. A potential effect from EMA may influence the interpretation of the results; therefore, investigators ought to be encouraged to report considerations related to EMA when designing and publishing a study. Future
research may benefit from randomizing a subset of participants to receive one EMA modality (eg, texting) while observing behaviors among all participants with another measurement modality (eg, Bluetooth e-cigarette monitor).

When the potential for EMA to influence study outcomes is identified, less obtrusive EMA methods could be considered, when available. The participants in this study expressed the omnipresent awareness and anticipation of the SMS text messaging–based EMA. This awareness altered participants’ behaviors and resulted in negative emotions that likely would not have been emphasized had the EMA not been used or had been subtler. There are numerous types of EMA strategies used in smoking research [4]. Future intervention research studies could consider less frequent SMS text messaging or EMA data collection methods outside of SMS text messaging that may have a more minor impact on smoking behaviors, such as the use of biosensors [24], Bluetooth-enabled devices [25], or puff counters [26].

This study had a few limitations. First, the interview guide was not designed to investigate the impact of EMA on participants’ smoking behaviors. Therefore, further details on the effects of EMA were not deeply explored, limiting the scope of this analysis. The unprompted nature of the participants’ observations of behavioral impact, however, strengthens the conclusion that the SMS text messaging–based EMA had a meaningful impact on the trial participants. Second, the sample of interview participants was not randomized, and the interview was not required; rather, it was offered to all participants sequentially as an optional component. This potentially introduced selection bias, as those with stronger opinions on the program may have been more likely to participate. Third, as EMA data were collected as part of a smoking reduction trial, it is difficult to completely disentangle the effects of the intervention on changes in behavior from the effects of EMA. Finally, the impact of EMA on behavior change was based on the self-reported perceptions of interview participants, and behavior changes were not directly observed. Therefore, this study cannot be used as conclusive evidence that the EMA had a significant impact on smoking behaviors, and further research is needed.

**Conclusion**

The collection of EMA smoking behavior data via SMS text messaging may influence the behaviors and perceptions of participants in smoking interventions. More research is needed to determine the magnitude of impact and mechanisms, to account for the potential effects of EMA on behavior change. Furthermore, a broader discussion of the behavioral influence introduced by the use of EMA may be warranted among the EMA research community.


Abbreviations

- **COREQ**: Consolidated Criteria for Reporting Qualitative Research
- **EMA**: ecological momentary assessment
- **NRT**: nicotine replacement therapy
- **RCT**: randomized controlled trial
Investigating Receptivity and Affect Using Machine Learning: Ecological Momentary Assessment and Wearable Sensing Study

Zachary D King¹, MS; Han Yu¹, MS; Thomas Vaessen²,³, PhD; Inez Myin-Germeys², PhD; Akane Sano¹, PhD

¹Department of Electrical and Computer Engineering, Rice University, Houston, TX, United States  
²Center For Contextual Psychiatry, Katholieke Universiteit Leuven, Leuven, Belgium  
³Department of Psychology, Health & Technology, University of Twente, Enschede, Netherlands

Abstract

Background: As mobile health (mHealth) studies become increasingly productive owing to the advancements in wearable and mobile sensor technology, our ability to monitor and model human behavior will be constrained by participant receptivity. Many health constructs are dependent on subjective responses, and without such responses, researchers are left with little to no ground truth to accompany our ever-growing biobehavioral data. This issue can significantly impact the quality of a study, particularly for populations known to exhibit lower compliance rates. To address this challenge, researchers have proposed innovative approaches that use machine learning (ML) and sensor data to modify the timing and delivery of surveys. However, an overarching concern is the potential introduction of biases or unintended influences on participants’ responses when implementing new survey delivery methods.

Objective: This study aims to demonstrate the potential impact of an ML-based ecological momentary assessment (EMA) delivery system (using receptivity as the predictor variable) on the participants’ reported emotional state. We examine the factors that affect participants’ receptivity to EMAs in a 10-day wearable and EMA–based emotional state–sensing mHealth study. We study the physiological relationships indicative of receptivity and affect while also analyzing the interaction between the 2 constructs.

Methods: We collected data from 45 healthy participants wearing 2 devices measuring electrodermal activity, accelerometer, electrocardiography, and skin temperature while answering 10 EMAs daily, containing questions about perceived mood. Owing to the nature of our constructs, we can only obtain ground truth measures for both affect and receptivity during responses. Therefore, we used unsupervised and supervised ML methods to infer affect when a participant did not respond. Our unsupervised method used $k$-means clustering to determine the relationship between physiology and receptivity and then inferred the emotional state during nonresponses. For the supervised learning method, we primarily used random forest and neural networks to predict the affect of unlabeled data points as well as receptivity.

Results: Our findings showed that using a receptivity model to trigger EMAs decreased the reported negative affect by >3 points or 0.29 SDs in our self-reported affect measure, scored between 13 and 91. The findings also showed a bimodal distribution of our predicted affect during nonresponses. This indicates that this system initiates EMAs more commonly during states of higher positive emotions.

Conclusions: Our results showed a clear relationship between affect and receptivity. This relationship can affect the efficacy of an mHealth study, particularly those that use an ML algorithm to trigger EMAs. Therefore, we propose that future work should focus on a smart trigger that promotes EMA receptivity without influencing affect during sampled time points.

(JMIR Mhealth Uhealth 2024;12:e46347) doi:10.2196/46347

Published: June 20, 2024

Corresponding Author:  
Zachary D King, MS  
Department of Electrical and Computer Engineering  
Rice University  
6100 Main St  
Houston, TX, 77005  
United States  
Phone: 1 713 348 0000  
Email: zdk2@rice.edu

https://mhealth.jmir.org/2024/11/e46347
KEYWORDS

mobile health; mHealth; affect inference; study design; ecological momentary assessment; EMA; just-in-time adaptive interventions; JITIs; receptivity; mobile phone

Introduction

User Engagement in Mobile Health Systems

Mobile health (mHealth) technologies continue to grow within the health care sector and are imperative for precision medicine initiatives. mHealth can provide beneficial interactions between health care providers and patients outside clinical settings. An engaged and responsive user base in any mHealth system is vital for maximizing the knowledge that researchers and providers acquire. Mental health research mainly depends on active users because investigators rely on participant survey responses to establish ground truth. Researchers can only adequately interpret the relationships between physiology and psychological state with a population that is compliant with sensors and surveys. Evaluating a health construct is only possible with highly receptive participants in mHealth studies.

Here, we discuss 2 forms of interaction between participants and mHealth systems: ecological momentary assessments (EMAs) and just-in-time interventions (JITIs). EMAs gather in situ data from users in real time. EMAs are commonly used in mHealth studies, as they allow researchers to prompt participants regularly throughout the day [1]. In the case of mHealth studies focusing on psychological states, EMAs enable users to report their momentary symptoms or context in a natural environment, often using smartphones, because of their accessibility. JITI is a method that allows investigators to send interventions as needed. The just-in-time adaptive intervention (JITAI) uses incoming information (physiological, contextual, or psychological markers) as context to determine when an intervention is required [2]. Researchers have been working on enhancing the efficiency of these interactions. As mentioned previously, this effort is crucial because ineffective interactions in an mHealth study can have significant effects on outcomes. Failing to collect EMA responses may impede researchers’ ability to identify real-world measures of health behaviors, and without participants receiving or engaging in JITIs, researchers may find it challenging to measure the efficacy of the intervention.

Improving EMA Receptivity

To enhance compliance with EMAs and JITIs, it is imperative to gain a comprehensive understanding of the factors that influence participant adherence. Ho and Intille [3] described 11 factors that influence a person’s interruptability (willingness to follow through if notified or interrupted). These factors encompass contextual aspects, such as social engagement, ongoing activities, future schedule, and emotional state, as well as message-related attributes, including frequency, complexity, modality, and utility.

Currently, many researchers have reduced interruptability by altering message-related attributes, often involving strategies such as reducing the complexity or frequency of an EMA or increasing the incentives for a response [4,5]. Reducing the size of the instrument relieves some of the burden associated with answering an EMA [6]. This is done by excluding redundant questions or by choosing a less complex instrument. The Perceived Stress Scale [7] was initially a 14-item question set. However, after some statistical analysis, researchers found that a 10-item instrument was sufficient for measuring stress. Another factor affecting receptivity is the frequency at which users are sampled. In 2 separate reviews, researchers demonstrated conflicting findings regarding the effects of frequency on EMA compliance [8,9]. These conflicting results can be attributed to the author’s focus on differing populations and the many other factors that play a role in EMA compliance. The third method for improving receptivity rates is to increase the incentives based on EMA compliance. However, this method can be costly and seen as exploitative, especially when dealing with susceptible populations.

An emerging method for improving receptivity rates is the use of machine learning (ML). This can be achieved by using wearable data to predict the likelihood of a response, which can help deliver EMAs that mitigate interruptibility. Mishra et al [10] used ML models built from previously collected data to improve the receptivity of a JITAI by contacting users at points where they are more likely to be receptive. The study showed a difference of >38% in receptivity rates between an ML-based static model (using previously collected data) and a control model (using a set schedule) to distribute EMAs. Mishra et al [11] built a model for predicting the optimal time to send an EMA. Their results demonstrated that a model built from contextual cues such as activity, audio, conversation, and location could significantly outperform a baseline model (prediction based on the proportion of responded EMAs). Researchers have also shown that contextual cues, including location [12,13], personality traits [14,15], physical activity [14,16], and time of day [17], influence participants’ willingness to respond to regular surveys. Together, these methods can predict and respond to the unobserved contextual aspects of an interruption, thus offering a more holistic approach to addressing participant engagement. However, a system that reacts to these contextual aspects may have unintentional effects on the response of the user. For instance, emotional state is an underlying factor that affects receptivity. A model designed to initiate EMAs when a participant is most likely to respond favors prompting users when experiencing positive emotions. Consequently, this approach could influence the reported emotional state during each prompt, potentially making it challenging to collect subjective responses during negative emotions. Understanding the influence of ML-based EMA triggers on these underlying receptivity factors allows us to incorporate additional variables into an algorithm. Integrating predicted affect into the decision-making of an ML-based EMA trigger will ensure that participants receive prompts across a broad spectrum of emotions.
Relationship Between Affect and EMA Receptivity

Clark and Watson [18] described how positive and negative affect (NA) can influence participation in activities of daily living. Their results show differences in the expected mean across many social activities, with reported positive affect (PA) having more significance in differentiating the two groups. Similarly, research has also demonstrated a negative relationship between students’ emotional state and academic achievement [19,20]. Although none of these studies demonstrate the relationship between affect and EMA receptivity during mHealth studies, they all demonstrate the effect of emotional state on a participant’s general ability to engage in normal activities of daily living.

Several authors have examined the effect of emotional state on EMA adherence by using the preceding response as a gauge of affect during instances of nonresponse. Murray et al [21] conducted a study (N=261) demonstrating that NA and stress reduce the chance of a response during the next prompt. Other researchers have expanded on this by examining various contextual cues within an EMA that precede instances of nonresponse. The authors found that variables such as medication use, activity, battery life, and being away from home negatively impacted the compliance of the following EMA [22,23]. This work contributes to understanding how affect can influence participants’ response behavior but falls short of providing real-time explanations for the absence of responses. Alternatively, real-time explanations for receptivity can be derived through passive sensing and ML. Leveraging these explanations allows for delivering EMAs at moments of heightened receptivity, guided by current contextual and physiological factors.

Objectives and Hypothesis

This study aims to analyze the relationship between participant EMA receptivity and affect in a 10-day wearable and EMA–based affect-sensing study (N=45). We hypothesize that a relationship exists between EMA receptivity and affect in mental health–related mHealth studies. We can establish the relationship between emotions when participants respond. However, to investigate this connection during nonresponses, we need to infer affect when a participant fails to provide a response. Therefore, we implemented ML models for identifying receptive time points and predicting emotional states. This allowed us to determine whether there was a statistically significant difference in emotions between responses and nonresponses. If this relationship exists and the likelihood of a response is dependent on emotional state, it would bias the outcome of an ML-based EMA delivery mechanism.

Methods

Ethical Considerations

Ethics approval was granted by the Sociaal-Maatschappelijke Ethische Commissie of Katholieke Universiteit Leuven (G-2018 09 1339) [24]. Informed consents were obtained from the participants. All data was de-identified prior to analysis.

Data Collection

This study included 45 healthy adult participants from Leuven, Belgium [24]. The average age of the participants was 24.5 (SD 3) years and ranged from 19 to 35 years. In total, 84% (38/45) of the participants were female. The participants were recruited via flyers distributed to areas around Leuven.

The study lasted for 10 days. The participants wore a sensor suite (Figure 1), including a chest patch with 2 electrodes for gathering electrocardiography (ECG) at 256 Hz and a wristband for electrodermal activity (EDA) at 256 Hz, skin temperature at 1 Hz, and accelerometer at 32 Hz. Participants were allowed to remove the device while they slept and were asked to remove the devices while bathing or participating in rigorous activities. The sensors had a battery life that surpassed the duration of the study, and the data were recorded on the device on an SD card.
Participants were given a research phone, and 10 EMAs were sent to the participants daily at random time points between 15 and 90 minutes apart. EMAs were initiated via text messages, and the participants had a specific amount of time to respond to the survey attached to the text message before it closed. The EMAs contained a question set to assess mood [25] in 3 languages: English, Belgian, and French. In total, there were 13 questions, including 9 negative (worried, stressed, anxious, annoyed, down, restless, tense, under pressure, and ashamed) and 4 positive (relaxed, cheerful, confident, and in control) affect-related questions. The questions were prefaced with the phrase “At the moment, I feel...” followed by a rating scale for each emotion, ranging from 1 (not at all) to 7 (very much). The participants were given €0.5 (US $0.54) for each EMA they responded to.

**EMA Analysis**

Our EMA question set was scored by adding the numerical interpretation of the 9 negative responses to the inverse (1 is 7 and 7 is 1) of the positive questions. The range of possible scores was between 13 and 91, with higher scores indicating more negative emotions. Owing to the low variance in reported positive and NA, we used a composite score of both positive and NA.

We also analyzed the participants’ response time (time between the notification and onset of EMA) and the response rate to EMA. We then investigated the potential for loss of engagement over time, which may lead to reduced participant receptivity. The lack of engagement may impede our capacity to discern the underlying causes of nonresponsiveness, particularly when assessing the relationship between affect and receptivity.

**EMA Receptivity and Affect Detection Models**

**Overview**

In the following sections, we discuss the sequential methodology, which encompasses the collection of raw signal data, the subsequent data processing and feature extraction, and the design of ML models for inferring 2 constructs—receptivity and affect. This framework is shown in Figure 2.

We began by processing our 4 sets of time series data: skin temperature, ECG, EDA, and accelerometer. Once we processed the data, we segmented them and attached labels to each segment based on the conditions explained in the EMA Receptivity Labels section. Next, we built and tested multiple ML algorithms to infer EMA receptivity and affect and verified the results using several statistical techniques.

**Preprocessing**

**Time Series Processing**

We began by extracting all the data from the 4 time series data sets. Table 1 shows the features computed for the 4 sets of the data. We used IQR to process skin temperature to remove outliers. We used biosppy [26] for the ECG to process the data and extract the R peaks. Biosppy uses a bandpass filter with frequencies of 3 Hz and 45 Hz, a sampling rate of 256, and the Hamilton segmentation algorithm to extract R peaks. We then validated the R peaks using an algorithm by Hovsepian et al [27]; this algorithm uses the criterion beat difference based on the maximum expected difference for a beat and the minimal artifact difference. We then used heart rate variability analysis to extract heart rate and heart rate variability features such as number of pairs of successive normal-to-normal intervals that differ by more than 20 ms and root mean square of successive differences between normal heartbeats [28]. We also obtained some frequency- and geometric-based features. For EDA, we used the method proposed by Taylor et al [29] to process and extract the statistical and wavelet features. Finally, for accelerometer, we smoothed the signal by using a fourth-order 10-Hz low-pass Butterworth filter and obtaining an average, and then, we used a package from the study by Simon [30] to extract step features. The features we extracted and the information on how those features were calculated are shown in Table 1.
Table 1. Features from our 3 raw sources and definitions of the features that are less commonly used.

<table>
<thead>
<tr>
<th>Signal</th>
<th>Features</th>
<th>Description</th>
<th>Prior work</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ST</strong>&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Mean, median, mode, minimum, range, root mean square, zero cross, kurtosis, skew, and IQR (25th percentile and 75th percentile)</td>
<td>Zero cross here is based on the number of times ST crosses over the mean ST. Kurtosis measures the extremity of the data in the segment, and skew is the measure of asymmetry.</td>
<td>[31]</td>
</tr>
<tr>
<td><strong>ECG</strong>&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Mean, median, mode, minimum, range, root mean square, zero cross, kurtosis, skew, IQR (25th percentile and 75th percentile), RMSSD&lt;sup&gt;c&lt;/sup&gt;, CVSD&lt;sup&gt;d&lt;/sup&gt;, CVNNI&lt;sup&gt;e&lt;/sup&gt;, SDNN&lt;sup&gt;f&lt;/sup&gt;, NNI50&lt;sup&gt;g&lt;/sup&gt;, NNI20&lt;sup&gt;h&lt;/sup&gt;, PNNI50&lt;sup&gt;i&lt;/sup&gt;, PNNI20&lt;sup&gt;j&lt;/sup&gt;, LF&lt;sup&gt;k&lt;/sup&gt;, VLF&lt;sup&gt;l&lt;/sup&gt;, HF&lt;sup&gt;m&lt;/sup&gt;, high/low-frequency ratio</td>
<td>Normal to Normal or RR&lt;sup&gt;n&lt;/sup&gt; interval indicates time between heartbeats. NNI20 or NNI50 refers to the number of successive intervals that differ by more than 20 or 50 ms. “P” indicates the proportion of NNI20 or NNI50 in the segment. RMSSD is the root mean square of successive differences between heartbeats. CVNNI and CVSD are the coefficients of variation SDNN/mean and RMSSD/mean, respectively. Our frequency domain features are based on how much of the signal lies between 0.003 and 0.04 Hz (VLF), 0.04 and 0.15 Hz (LF), and 0.15 and 0.40 Hz (HF).</td>
<td>[32-35]</td>
</tr>
<tr>
<td><strong>Electrodermal activity</strong></td>
<td>Wavelet: maximum, mean, SD, median, and above zero (1-second and half-second wavelet); raw: amplitude, maximum, minimum, and mean; filtered: amplitude, maximum, minimum, and average</td>
<td>A 1-second and a half-second window were used for wavelet features. Features were calculated for both the first and second derivatives of each window size.</td>
<td>[31,36-39]</td>
</tr>
</tbody>
</table>

<sup>a</sup>ST: skin temperature.
<sup>b</sup>ECG: electrocardiography.
<sup>c</sup>RMSSD: root mean square of successive differences between normal heartbeats.
<sup>d</sup.CVSD: coefficient of variation of differences between adjacent normal-to-normal intervals.
<sup>e</sup.CVNNI: coefficient of variation of the normal-to-normal intervals.
<sup>f</sup>SDNN: SD of the normal-to-normal intervals.
<sup>g</sup>NNI50: number of pairs of adjacent normal-to-normal intervals differing by more than 50 ms.
<sup>h</sup>NNI20: number of pairs of adjacent normal-to-normal intervals differing by more than 20 ms.
<sup>i</sup>PNNI50: percentage of pairs of adjacent normal-to-normal intervals differing by more than 50 ms.
<sup>j</sup>PNNI20: percentage of pairs of adjacent normal-to-normal intervals differing by more than 20 ms.
<sup>k</sup>LF: low frequency.
<sup>l</sup>VLF: very low frequency.
<sup>m</sup>HF: high frequency.
<sup>n</sup>RR: R-peak to R-peak.

**Segmentation**

We segmented the data into 1-minute windows with a 30-second overlap. We then calculated the statistical features for each of the sensors, excluding steps. For each of these windows, we calculated historic features. To do so, we elongated each of the windows by 5, 30, and 60 minutes and then extracted the features with the extended window size (ie, for each 1-min window, we have not only the features from the 1 min but also the features going back to these 4 time frames).

**EMA Receptivity Labels**

Labels for receptivity were based on whether the user responded to the EMA and were assigned to segments based on whether it was within a specified time of the scheduled notification. By expanding the window of labeled data, we can increase the size of the labeled data set (pseudolabeling). However, as this window increases, so does the distance between some of our time points and the corresponding label. We tested windows that are 5, 30, 60, and 120 minutes long. For instance, for the 5-minute window, if an EMA was sent at midnight, the segments that fell between 11:55 AM and midnight would be labeled “responded” if they did respond and “no response” if they did not. We applied the same method for the affect labels (Figure 3). We ultimately chose 30-minute windows owing to the balance between the size of the training set and the labeled points being relatively close in terms of time to the actual response (or nonresponse).
Receptivity and Affect Model Design and Hyperparameter Tuning

We designed ML models to infer EMA receptivity and affect. A wide variety of ML algorithms are used in affect and receptivity prediction including random forest (RF) [31,32,39], support vector machine [33,34,39], logistic regression, k-nearest neighbors [30], neural network (NN; long short-term memory, recurrent NN, convolutional NN, etc) [31,39], and naïve Bayes [39,41]. On the basis of our sensor data, initial tests, and drawing inspiration from previous studies, especially those by Mishra et al [10,11]. We selected (1) RF for predicting emotional state and receptivity, (2) an NN for predicting emotional state, and (3) a baseline model. This baseline model serves as a benchmark for evaluating whether our models outperform random chance, whereas the NN algorithm was introduced as a possible improvement on existing model implementations. Unlike the research mentioned previously, we used physiological data rather than contextual data. These signals are sampled at higher frequencies compared with contextual data and allow the extraction of more fine-grained features, making NNs more feasible. We designed personalized models to infer the receptivity and effect of EMA.

To optimize our personalized model, we selected hyperparameters using the grid search method for each participant, explicitly using the GridSearchCV method defined in scikit-learn. This method uses an exhaustive search method (ie, testing each user-defined parameter permutation). The hyperparameters tested included the number of estimators, maximum depth of the estimator, minimum number of samples per leaf, minimum number of samples for split, and maximum number of features that can be used for the split. Using training and validation sets, we selected the parameters and then applied the optimal model to our test set. The optimal set of hyperparameters differed for each participant, although the most common optimal hyperparameters chosen included 60 estimators, maximum depth=3, minimum sample leaf=2, minimum sample split=2, and maximum features=square root of the number of features.

Our NN model was structured to use 3 densely connected layers using a rectified linear unit activation function at each layer. The output dimension of each layer was 256, 128, and 64, and the output layer was a densely connected layer with 2 output
dimensions. The reasoning for an output layer of 2 is to define a CI for our regression model.

The baseline model was built by predicting random output based on the distribution of the class labels in the training set (ie, if 10/100, 10% of the labels were nonresponses and 90/100, 90% were responses, the model would predict nonresponses 10/100, 10% of the time). We can determine the expected outputs for this model; our true positive rate should be equal to \( Pr(\text{response in the training set}) \times Pr(\text{response in the test set}) \). The more evenly the class labels are distributed, the worse the performance of the model. For the affect regression models, we used a normal sampling method with the mean and SD based on the training set class labels.

As there are more labeled responses compared with nonresponses, we considered this imbalance in the receptivity prediction model, weighting the classes based on the distribution in our training set. All models were built using the Python packages `scikit-learn` [42] or Tensorflow [43].

**Model Uncertainty**

To determine the relationship between affect and receptivity, we must use predictions to infer the emotional state of our participants during nonresponses. As affect is a complex and difficult-to-predict construct, we need a method for filtering our predictions based on some level of confidence. Therefore, we introduced a method for calculating uncertainty for regression using an NN.

Determining a confidence value for a regression model is difficult compared with a binary or categorical model. We can use a custom loss function in our NN to estimate epistemic and aleatoric uncertainty for our regression model, where epistemic uncertainty is based on our ability to predict our class labels with the available data (affected by lack of knowledge or data), and the aleatoric uncertainty is affected by randomness, which is unknown or unmeasured in the model [44].

Our affect prediction model outputs are 2D rather than a single predicted output. The first output is the predicted affect, \( \mu(x) \), and the second output, \( \ln(\sigma(x)) \), is the predicted variance (the log allows us to take the exponent to ensure a positive value for \( \sigma \)). Both \( \mu \) and \( \sigma \) are functions of our training set \( x \).

The loss function \( L \) is shown in the equation (custom loss function for measuring model uncertainty) and is derived from the mean square error (MSE) calculation and the maximum likelihood of a normal Gaussian distribution [45]. The numerator of this equation is identical to the MSE loss function, where \( \mu(x) \) is the predicted output of our model. Unlike the MSE loss function, we continuously update not only our predicted output \( \mu \) but also the predicted variance \( \sigma \). The \( \sigma \) output of our model is based on error; the sigma value increases to account for higher error and decreases to account for lower error. This \( \sigma \) value can be used as an uncertainty or error metric. Although it is still a predicted value, it should align with how confident the model is in the \( \sigma(x) \) output. The \( \sigma \) value plays a crucial role in assessing the confidence of our affect predictions, given that we use predicted affect to infer emotional states during nonresponses.

Consequently, to illustrate the relationship between the predicted sigma value and model uncertainty, we performed a mixed effect model analysis using affect scores and the predicted sigma values and tested whether greater uncertainty will occur in emotional states that are less frequently represented and when the testing error is larger. As uncertainty is a measure of the model’s confidence in its predictions, we can reasonably assume that predictions associated with larger testing errors would correspond to higher levels of uncertainty.

**Model Evaluation**

For cross-validation, we used a personalized random train-test split cross-validation method. We randomly split the data into training and testing sets using the response label (whether they responded to the EMA or not) to stratify the split. Responses and nonresponses can encompass multiple segments; by grouping them together, we avoid splitting up segments from a single response or nonresponse. As our response labels are imbalanced, we want to ensure that our training, validation, and test sets have a relatively even number of responses and nonresponses. For the purpose of fairness, we excluded 3 participants who had a single nonresponse from our receptivity results.

We first normalized the training and test sets independently of one another based on the participant. In total, we obtained approximately 230 features from the sensor signals. We reduced our feature set using principal component analysis. Our implemented principal component analysis was set such that the number of produced components explained 99% of the variance (48 features). This method was used for each model, excluding the RF model, in which the original normalized data were used as the input.

**Analysis of the Relationship Between Affect and Receptivity**

**Overview**

We conducted two different analyses to understand the relationship between affect and receptivity better:

1. To infer emotional state during nonresponses, we clustered the physiological data and then examined the makeup of the clusters. By doing so, we can assume the emotional state of different clusters and unlabeled data points.
2. For EMAs the participants did not respond to, we used the affect prediction model described in the previous section to infer the emotional state at the time of a nonresponse. With these newly predicted affect scores, we can analyze the differences in the emotional state during a response and nonresponse.

**Cluster Evaluation**

We used the most significant features (based on correlation) when predicting receptivity for our clustering analysis. To determine the optimal clustering method, we tested several clustering methods, including hierarchical and \( k \)-means clustering, with a maximum number of iterations of 300. We then calculated the silhouette score across all clusters using receptivity as our ground truth and selected our best-performing set of hyperparameters. On the basis of the cluster distribution,
we analyzed the difference in the perceived emotional state of the participants. We calculated the average NA, PA, and receptivity rates in the clusters for each participant and then characterized the clusters based on receptivity rates (high receptive and low receptive clusters). Next, using repeated measure ANOVA, we demonstrated the statistical difference between affect and the clusters. Given that the clusters were created from physiological data, we know that the data points within each cluster are physiologically similar; therefore, we inferred that they would also exhibit similar psychological states. This allowed us to assign affect scores to nonresponsive data points within each cluster based on the labeled data points within that cluster. Unlike affect prediction, we used the raw NA and PA values in our evaluation as the clustering was performed independently of affect scores; therefore, the lack of variance in responses did not affect the output of the clustering. These results gave us a sense of participants’ perceived emotional state during nonresponses. We also investigated differences in receptivity in 2 clusters using the chi-square test.

**Analysis of the Receptivity and Affect Relationship**

Ideally, we would show the interaction between affect and receptivity using the data collected. However, because nonresponses do not have a corresponding affect score, we designed and implemented our models for receptivity and emotional state.

After generating predictions for our test data set, we assessed the agreement (using Cohen κ) and correlation (using the point biserial method) between receptivity and predicted affect, leveraging true labels at time points when affect measures were reported. A high level of agreement or correlation would suggest a strong relationship between these 2 constructs, thereby highlighting the potential influence each construct would have on an ML algorithm to predict the other construct. We then examined the disparities between the predicted affect during nonresponses and the reported affect during responses. By doing so, we can establish the extent to which emotional state influences receptivity. Substantial disparities in affect between responses and nonresponses suggest that participants’ emotional states impact their receptivity. Consequently, a model designed to predict receptivity would indirectly include emotional state as a determinant of a participant’s receptiveness. However, it is essential to acknowledge that some of these variations could be attributed to model error. As a result, we also compared the predicted and reported affects during responses to investigate the significance of the model error. We then calculated and visualized the cumulative distribution of these 3 sets of values to illustrate the influence of affect on receptivity and the associated model error.

Finally, we investigated the potential effects that an ML-based receptivity algorithm would have on reported affect, influencing the outcome of the study. On the basis of our receptivity model, we can estimate the difference in the reported perceived emotional state between our true findings and predicted affect during time points that would initiate an EMA.

**Results**

In the following sections, we discuss the results of our study, particularly the methods of evaluation that were discussed in the previous section.

**EMA Analysis: Affect and Receptivity**

The distribution of EMA responses is shown in Figure 4. Although participants rarely indicated high negative emotions, this trend is evident in Figure 5, illustrating a box plot of composite scores for each participant. Participants’ average and median reported affect were <26, meaning that, on average, the participant responded to each question with a relatively low score of 2 (on a scale between 1 and 7, where 1 indicates high positive emotion and 7 shows high negative emotion). We also investigated participants’ emotional states as the study advanced and observed minimal to no variations based on their duration of enrollment or time of day.

On average, participants responded with a 4.5 for PA questions and a 1.8 for NA questions. This disparity in affect intensity was consistent with previous research [23]. There was a slight difference in the reported affect between male and female participants. On average, female participants responded with a 1.9 (SD 1.08) for the NA questions and 4.5 (SD 1.3) for the PA questions, whereas male participants responded with a 1.8 (SD 0.9) for the NA questions and 4.7 (SD 1.0) for the PA questions.

Of the 3885 notifications sent to the 45 participants, there were 3066 (78.92%) responses. As the study persisted, there was little to no drop-off in receptivity rates over time. This finding helped confirm that loss of engagement was not a contributing factor to receptivity. Most studies have stated that the quality receptivity rate is at 80%. The range of response time (time between notification and initiation of the EMA; Figure 3) was between 0.5 seconds and 306 seconds. Participants responded to the notification on average in 20.9 seconds and had a median response time of 8.7 seconds. There were no responses after 306 seconds of a notification. The reason for this fast response time is that participants were allowed 90 seconds to begin the survey, after which the survey would no longer be accessible (we had a few responses after the 90-second restriction owing to software or design issues). This restriction makes it challenging to relate response times to participant affect, as has been done by other researchers.

We found that none of the mood responses were strongly correlated with the time to respond. Across each question, we did not obtain a correlation coefficient >0.03 (all correlations indicated significant confidence; P<.05). This low correlation coefficient indicates that the participant’s mood had little to do with how long it took the participant to initiate the EMA. Although considering the limit we put on the response time, this relationship might be difficult to assume.
Figure 4. Question set: includes the 13 questions used to measure affect with their mean, SD, and correlation to the final affect score. For each question, participants were asked to rate the degree they were experiencing each emotion. These 13 questions can be split into positive affect (orange) and negative affect (blue).

Analysis of Features

The features that we found to be the most significantly related to receptivity were ECG low frequency (1 min, momentary: \( F_{2,54}=6.7; P<.001 \)) and very low–frequency features (1 min, momentary: \( F_{2,54}=4.7; P=.02 \) and 60 min: \( F_{2,54}=4.1; P<.001 \)); EDA mean (\( F_{2,54}=10.2; P<.001 \)) and median (\( F_{2,54}=15.4; P<.001 \)); number of pairs of adjacent NN intervals differing by more than 20 ms in the 5 and 60 minute windows; percentage of pairs of adjacent NN intervals differing by more than 50 ms in the 30 minute window (\( F_{2,54}=11.2; P<.001 \)); and maximum (\( F_{2,54}=6.3; P<.001 \)), minimum (\( F_{2,54}=3.6; P=.009 \)), and absolute maximum (\( F_{2,54}=6.6; P=.002 \)) of the first and second derivatives for EDA. These results show that ECG and EDA-related features were best at differentiating between responses and nonresponses compared with features derived from accelerometer and skin temperature.

When running the LMMs to determine the relationship between features and emotional state, we found a nonsignificant relationship between affect scores and steps or sleep features. However, heart rate was significant when predicting emotional state, particularly negative emotion. This LMM showed a significant positive relationship between heart rate and affect (\( \beta=.007; P<.04 \)). This underscores the significance of heart rate as a predictor of emotional state, although it does not necessarily imply that steps and sleep features lack importance in this context.

Receptivity and Affect Models

After processing, cleaning, and filtering out segments with confounding values, we obtained 1368 responses with usable physiological data. As our class labels were expanded to include segments 30 minutes before the point of response (pseudo labeling), we ended up with 13,477 data points for determining affect and 17,254 data points for predicting response.

Model Performance

Table 2 shows the results of our receptivity (binary) and affect (regression) models. On the basis of these results, there was little difference between the RF and NN models, although we used the NN models to demonstrate the relationship between affect and receptivity in the following section.

Table 2. Model results for predicting receptivity (binary) and affect (regression).

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy, mean (SD)</th>
<th>Precision, mean (SD)</th>
<th>Recall, mean (SD)</th>
<th>( F_1 )-score, mean (SD)</th>
<th>Root mean square error (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.73 (0.001)</td>
<td>0.83 (0.002)</td>
<td>0.84 (0.002)</td>
<td>0.83 (0.002)</td>
<td>11.1 (4.3)</td>
</tr>
<tr>
<td>Neural network</td>
<td>0.84 (0.19)</td>
<td>0.82 (0.006)</td>
<td>0.85 (0.10)</td>
<td>0.86 (0.20)</td>
<td>7.3 (2.7)</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.83 (0.11)</td>
<td>0.82 (0.15)</td>
<td>0.94 (0.10)</td>
<td>0.87 (0.12)</td>
<td>7.5 (3.1)</td>
</tr>
</tbody>
</table>

Analyzing Uncertainty in the Affect Model

Figure 6A shows the relationship between the calculated sigma value (uncertainty) and the reported affect scores. Uncertainty should follow a pattern where class labels that are more represented in the training set should have lower uncertainty. Conversely, values that are less represented in the data set should have larger uncertainty. As can be seen, Figure 6A \( \sigma \) values are smaller when the reported emotional state is more positive. As shown in Figure 5, most respondents indicated relatively low composite scores, with a few participants reporting an affect score >40. We also observed a statistically significant relationship between sigma and affect scores, as shown in Figure 6A, using a mixed effect model. In this model, we accounted
for the random effect associated with participants, as indicated by the mixed linear model results (intercept: 7.090; \( P < .001 \) and affect score: 0.002; \( P = .046 \)).

Figure 6B shows the relationship between \( \sigma \) and the testing error; in particular, \( \sigma \) values were larger when the model was further from the ground truth. This relationship shows that our \( \sigma \) value is an accurate representation of model uncertainty. On the basis of Figure 6, we can say that the \( \sigma \) value we calculated is related in some way to uncertainty. Figure 6B shows that most responses indicating an affect score of <39 had a \( \sigma \) of <6. Therefore, we chose 6 as the cutoff for uncertainty. This cutoff filters out many of the predictions that are more likely to have higher errors because we cannot look at errors during nonresponses, as we have no affect label.

Figure 5. Box plot of perceived emotional state, minimum is 13 (negative) and maximum is 91 (positive). The average perceived emotional state is 26.42, denoted by the blue horizontal line.

Receptivity and Affect Analyses

Cluster Analysis

On the basis of the “elbow rule” of silhouette scores, we chose \( k \)-means as our clustering method with 2 clusters. We found that the distribution of receptivity was somewhat different between clusters. Cluster 0 contained a higher density of responses, with just <15% nonresponses, whereas cluster 1 had a higher density of nonresponses of just >21%. We first analyzed the overall affect scores in the 2 clusters, where we found the average reported affect score in cluster 1 to be >3 points higher than the average reported affect in cluster 0 (repeated measure ANOVA, \( F_2 = 23.16; P < .001 \)). The receptivity rates and average reported affect scores for the 2 clusters are shown in Table 3. We also found that the distribution of receptivity was different between the 2 clusters using the chi-square test of independence (\( \chi^2 = 898.8; P < .001 \)). These results indicate distinctions between response and affect across the cluster labels. Considering that the cluster with a higher density of nonresponses (cluster 1) also had a higher average affect score (higher scores indicate more intense negative emotions or lower positive emotions), we can assume that there was a relationship between EMA receptivity and reported affect.

Figure 7 shows a scatter plot of the difference in perceived PA between the 2 clusters and the difference in perceived NA between the 2 clusters for each participant. The results show that participants’ perceived emotion was more negative regarding lower PA and higher NA in cluster 1 compared with their perceived emotional state in cluster 0. As stated earlier, cluster 1 contains a higher percentage of nonresponses compared with cluster 0, indicating that cluster 1 is a better representation of a nonresponse. Therefore, it appears that there is a relationship between negative perceived emotional state and receptivity. Using the cluster labels as groups, we calculated the \( F \) test statistic using an ANOVA test for each feature. The features that separated the 2 clusters were mostly calculated from the ECG signal, including the minimum heart rate, low or very low frequency, mean heart rate, coefficient of variation of the NN intervals, coefficient of variation of differences between adjacent NN intervals, high frequency, and maximum heart rate (in order of \( F_1 \)-score). Features obtained from the EDA, accelerometer, and body temperature did not return significant \( P \) values when calculating the \( F \) test statistic.
Table 3. Receptivity rates and average reported affect scores in each cluster.

<table>
<thead>
<tr>
<th>Cluster number</th>
<th>Receptivity rate</th>
<th>Reported affect score, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 0</td>
<td>0.85</td>
<td>24.3 (4.7)</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>0.78</td>
<td>27.3 (4.9)</td>
</tr>
</tbody>
</table>

Figure 7. Each point represents a participant, where the x-axis denotes the difference between average negative affect (NA) of cluster 0 and 1, whereas the y-axis represents the difference between average positive affect (PA) of clusters 0 and 1.

Relationship and Analysis Between Receptivity and Affect

Figure 8 shows the cumulative distribution of reported affect scores for responses and predicted affect scores for responses and nonresponses. On the basis of this figure, there is a clear difference between the predicted affect during nonresponses and our true affect scores. Although this could be a model error, we also predicted affect scores during these responses and found that our model consistently predicted lower affect values (higher PA).

There was a fair amount of agreement between our affect and our binary response model, with a Cohen $\kappa$ score of 0.33 and a correlation of 0.44. When our model predicted a response, 77.42% (22,761/29,399) of the segments were during times when the affect model predicted PA. Only 69.72% (7760/11,131) of the predicted nonresponses reported PA. This indicates that the predicted response is negatively related to affect (ie, responses are associated with PA, whereas nonresponses are associated with NA). The reason determining the relationship between our constructs is important is because this bias can, and as we show, affect the overall outcome of a study. For instance, the average predicted affect score for times that we predicted as low likelihood for a response was a full 1.5 (SD 1.35) or 2.01 points higher than the average predicted affect for points predicted to be of high likelihood for a response. When observing only the segments where we misclassified a response (ie, we had a true affect, but the response was misclassified as a nonresponse), we found that the average affect score dipped slightly from 26.1 (predicted nonresponse) to 25 (predicted response). This difference in affect between responses and nonresponses is evidence that our receptivity model is indirectly based on affect. The SD of the affect score also decreased from 11.1 (true labels) to 9.8 (true affect and predicted response) during responses.

The average predicted affect score for a nonresponse was 30.9 (SD 11.2), and the average affect score for a response was 29.3 (SD 10.7; true) and 27.7 (SD 8.9; predicted). The predicted affect scores during nonresponses were higher than the reported and predicted affect scores during responses. Given that our average testing error was $-1.6$, we could also assume that the predicted affect during these nonresponses could be more negative than the true predictions. The distribution of these scores is shown in Figure 9. In Figure 9, all 3 groups’ affect scores peaked at around 20 to 25; this is probably owing to the large number of reported affect scores in this range. However, nonresponses had a second peak at an affect score of 40. This bimodal distribution could indicate that our affect distribution
during nonresponse was affected by ≥2 factors. Some nonresponses may not be affected by their affect but perhaps by their daily life activities (seeing a movie, spending time with family, showering, etc). In contrast, the second peak indicates that NA is related to nonresponses.

**Figure 8.** Cumulative distribution of predicted and actual affect scores for responses and nonresponses.

<table>
<thead>
<tr>
<th>Cumulative distribution of affect score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affect score</td>
</tr>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>80</td>
</tr>
</tbody>
</table>

**Figure 9.** Distribution of predicted and true affect scores for responses and nonresponses. Density is specific to the response and nonresponse.

<table>
<thead>
<tr>
<th>Distribution of predicted and true values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affect score</td>
</tr>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>80</td>
</tr>
</tbody>
</table>

**Discussion**

In this section, we discuss the outcome of our study, particularly the relationship between emotional state and receptivity, what that means, how it affects our results, and how we might implement a receptivity model that removes this bias. We also mention the limitations of this study.

**Principal Findings**

This study aimed to understand how ML models used to improve participant receptivity can affect the outcome of a study. Although we focus on emotional state in this study, we feel as if there are many health constructs and outcomes that can be affected by these receptivity models. Improving receptivity is not a new concept, but in the realm of mHealth, it is an emerging problem. The factors influencing study adherence have been analyzed and discussed in depth in previous research. One such scope is in medication adherence. Researchers have found many factors that influence medication adherence, from social, therapeutic, patient-related, and disease-related factors [46]. However, few have examined the momentary factors that affect adherence to medication or a health construct, and few have had the ability to do so without wearable sensors and momentary assessments.

Our findings using supervised learning and clustering indicate a clear relationship between emotional state and user receptivity. The clustering method demonstrated clear differences in affect between a highly receptive cluster and a less receptive cluster. The results of supervised learning demonstrate that users experience more negative emotions during nonreceptive time points. Although our results showed promise for a model dedicated to predicting response, we also showed the biases inherent in such a model. Ideally, we would want a receptivity model that is completely independent of emotion. Otherwise, we are influencing the participant’s responses.

Our results demonstrate that an mHealth study implementing a receptivity trigger based purely on the likelihood of responding (a model that triggers EMAs and JITIs using predicted receptivity) will bias the participant’s response. In this case, the model would initiate an EMA or JITI during times of more positive emotions, thereby decreasing the overall affect score.
for the EMA and possibly sending the JITI during times when the intended construct was not being met. As our ability to predict binary affect is limited with this data set, we believe that using the affect regression and ground truth labels for responses will return the most realistic representation of affect during nonresponses.

Comparison With Prior Work

Our findings are consistent with those of the previous studies. Many prompt-level studies [21-23] found a relationship between nonresponses and higher levels of NA in previous prompts. Although these results can provide insight into what makes a participant less compliant with EMAs, they do not offer a reasonable method for using this information in real-time decision-making. Using ML, wearable sensors, and contextual cues allows researchers to predict noncompliance components and distribute EMAs accordingly.

Consistently, our models either surpassed or achieved equivalent performance compared with previous research efforts. We achieved $F_1$-scores ranging from 0.83 to 0.87 when predicting receptivity. In contrast, Künzler et al [14] reported $F_1$-scores of approximately 0.4 while relying solely on contextual features. It is important to note that these results are not directly comparable, as contextual data lack the granularity of the data collected in our study.

Regarding affect prediction, our results present a unique challenge for comparison because we used regression in our predictions, unlike most researchers who typically use binary or categorical labels for emotion recognition. We chose not to convert our ground truth data into binary or categorical labels because of the inherent ambiguity in setting the thresholds and the limited variance in user responses. The effectiveness of affect prediction can vary significantly depending on the specific construct of interest and the sensors and signals available. Schmidt et al [40] conducted a review and reported an emotion recognition accuracy ranging from 40% to 95% using wearable sensors and signals. In terms of regression analysis, Tuarob et al [47] achieved nearly identical root MSE scores ($PA=7.37; NA=7.40$) when forecasting positive and NA scores from Positive and Negative Affect Scale using RF regression and previously collected questionnaire data.

Limitations

In this section, we address the limitations of our study, which can be categorized as limitations in our population, study design, data collection, and affect prediction models.

The major concern of our study population is that our results may be specific to this cohort. The study population was very receptive, even with 10 EMAs sent daily. This could be difficult for other researchers to implement, as the frequency and complexity of the EMA are fairly burdensome. Although we believe that the relationship between affect and receptivity would extend to other studies, it is important to note that our population was relatively small ($N=45$), predominately young (age 24.5 y), and had a higher representation of female participants (38/45, 84%). Consequently, our results may be specific to our cohort and EMA question set, but previous studies analyzing medication adherence and prompt-level relationships between EMAs and nonresponses indicate that the effect of emotional state on receptivity is common across multiple populations [21-23]. Further research is needed to explore the extent of this relationship between different emotional states and receptivity across multiple populations.

One limitation of the study design is that we cannot examine how loss of engagement over time affects the relationship between emotional state and receptivity. There was little to no drop-off in receptivity rates as our study progressed. This may have been because of the relatively short time frame in which the participants were enrolled. As a result, it is difficult to explore the effect emotional state would have on EMAs in the latter part of a study when participants can be more fatigued and less engaged. In future work, we intend to study a population for an extended period to analyze how emotional state affects participant response rates later in the study. Ideally, this will allow us to see the rate at which responses decay, the causes, and how we might combat it. Furthermore, we believe that a measure of this decay in engagement could be included in our ML-based decision-making for delivering EMAs that mitigate study fatigue, similar to how we would use model uncertainty to diversify emotional response.

Another potential study design limitation is that the app and research phone were shared with participants. Carrying 2 phones, especially one dedicated solely to responding to EMAs, can be burdensome for participants. In addition, the app designed for EMA distribution requires further usability evaluation. In future work, we aim to develop an app that can seamlessly integrate into users’ devices and assess its ease of use.

The data gathered in this study were limited to physiological features and user-defined responses. Although the physiological features make up a large portion of what researchers consider important for predicting psychological constructs, the data set lacks sampling contextual data. Certain contextual information is imperative for recognizing emotions and improving EMA response rates that cannot be obtained using physiology, such as social context. The social context can help infer the participant’s emotional state and willingness to respond to an EMA or JITAI.

Similarly, by incorporating more psychological and environmental cues (personality traits, working hours, etc), we can better understand what to expect from our participants regarding receptivity and affect before the start of the study. Using these prestudy measurements, we could assess the type of participants enrolled. Specifically, what will be their needs regarding receiving and responding to EMAs. This will help us develop and personalize our ML models for affect and receptivity.

The last significant limitation of our study is the use of predicted affect labels in determining the relationship between emotional state and receptivity. We can never collect reported affect during nonresponses for this or any data set. We attempt to reduce this limitation by using uncertainty to filter out less-confident predictions. Nevertheless, the predicted affect is only as good as our models. The only way to overcome this limitation is to improve the affect models. Although some may argue that the quality of our models needs to be more robust to claim a
The relationship between affect and receptivity, the effects of emotional state on engagement in social and daily life activities are well documented and consistent with our conclusion.

Conclusions
This paper presents the possibilities for bias in ML models to trigger surveys and interventions for participants in mHealth studies. Our results show a clear relationship between emotional state and user EMA receptivity. By designing an mHealth study using a "trigger" to improve participant response, it is imperative to consider some biases that may arise, in this case, affect. Participants were more likely to respond to an EMA during positive emotional states. If we distribute those EMAs to times when they are more likely to respond, we would further be biasing our participants’ recorded emotional state. Although this may not be a significant problem for less responsive populations, for the general population, this could change researchers’ perception of the participant’s perceived emotional state. In this study, we did not examine other constructs that might be a factor of receptivity because affect is the focal point of this study. For this objective, we are collecting both subjective and physiological data. Although this may be broad, it can be applied to any construct, particularly the intended construct of an mHealth study.

The pitfall of any mHealth study, particularly those involving psychological concepts, is the dependency on subjective user responses. The sampling rate of subjective responses will always be less than that of the physiological sensors and even some contextual cues. As our feature set became increasingly comprehensive, our labeled data remained relatively sparse. Considering that our proposed trigger considers factors beyond receptivity, it would likely have lower receptivity rates compared with triggers solely based on receptivity. However, the importance of even a minimal increase in a user’s adherence or engagement in a study can drastically improve researchers’ understanding of the health construct.

The models discussed in this paper have mostly proposed single-objective optimization functions that try to optimize based on whether the model considers that a user will respond to an EMA. In future work, we will propose a multiobjective optimization function for triggering EMAs and JITAs based on the likelihood of responding and an active-learning measurement of the health construct. This multiobjective function would base the timing of the EMAs on 2 separate objectives: receptivity and model uncertainty. By initiating EMAs or JITAs based on these 2 objectives, we can obtain an expected response that is more diverse in terms of affect. We hope that the work presented in this paper can be used to further enhance communication and the ability to gain knowledge from participants.

Conflicts of Interest
AS has received travel reimbursement or honorarium payments from Leuven Mindgate, American Epilepsy Society, the Institute of Electrical and Electronics Engineers, and Apple. AS has also received research support from Microsoft, Sony Corporation, NEC Corporation, Pola Chemicals, and Meta and consulting fees from Gideon Health and Suntory Global Innovation Center. All other authors declare no other conflicts of interest.

References


Abbreviations

ECG: electrocardiography
**EDA:** electrodermal activity

**EMA:** ecological momentary assessment

**JITAI:** just-in-time adaptive intervention

**JITI:** just-in-time intervention

**LMM:** linear mixed model

**mHealth:** mobile health

**ML:** machine learning

**MSE:** mean square error

**NA:** negative affect

**NN:** neural network

**PA:** positive affect

**RF:** random forest