Contents

Review

Human-Centered Design of Mobile Health Apps for Older Adults: Systematic Review and Narrative Synthesis (e29512)
Zethapong Nimmanterdwong, Suchaya Boonviriya, Pisit Tangkijvanich ......................................................... 3

Original Papers

Knowledge and Expectations of Hearing Aid Apps Among Smartphone Users and Hearing Professionals: Cross-sectional Survey (e27809)
Jae Han, Yong-Ho Park, Jae-Jun Song, Il Moon, Woojoo Lee, Yoonjoong Kim, Young Cho, Jae-Hyun Seo, Moo Park ................................. 24

Assessing Elderly User Preference for Telehealth Solutions in China: Exploratory Quantitative Study (e27272)
Nuoya Chen, Pengqi Liu ......................................................... 36

Dose–Response Effects of Patient Engagement on Health Outcomes in an mHealth Intervention: Secondary Analysis of a Randomized Controlled Trial (e25586)
Yiran Li, Yan Guo, Y Hong, Yu Zeng, Aliza Monroe-Wise, Chengbo Zeng, Mengting Zhu, Hanxi Zhang, Jiaying Qiao, Zhimeng Xu, Weiping Cai, Linghua Li, Cong Liu ......................................................... 56

Effect of an Integrative Mobile Health Intervention in Patients With Hypertension and Diabetes: Crossover Study (e27192)
Sang Oh, Kyung-Kon Kim, Sung Kim, Su Park, Sangshin Park ................................................................. 68

A Mobile App to Increase Fruit and Vegetable Acceptance Among Finnish and Polish Preschoolers: Randomized Trial (e30352)
Henna Vepsäläinen, Essi Skaffari, Katarzyna Wojtakowska, Julia Barli ska, Satu Kinnunen, Riikka Makkonen, Maria Heikkilä, Mikko Lehtovirta, Carola Ray, Eira Suhonen, Jaakko Nevalainen, Nina Sajaniemi, Maijaliisa Erkkola ................................................................. 82

Multipurpose Mobile Apps for Mental Health in Chinese App Stores: Content Analysis and Quality Evaluation (e34054)
Xiaojian Wu, Lin Xu, PengFei Li, TingTing Tang, Cheng Huang ................................................................. 93

Russian-Language Mobile Apps for Reducing Alcohol Use: Systematic Search and Evaluation (e31058)
Anna Bunova, Veronika Wiermer, Boris Gornyi, Carina Ferreira-Borges, Maria Neufeld ................................................................. 106

Prioritization of Quality Principles for Health Apps Using the Kano Model: Survey Study (e26563)
Christin Malinka, Ute von Jan, Urs-Vito Albrecht ................................................................. 117
A Smartphone App to Improve Oral Anticoagulation Adherence in Patients With Atrial Fibrillation: Prospective Observational Study (e30807)
Keitaro Senoo, Tomonori Miki, Takashi Ohkura, Hibiki Iwakoshi, Tetsuro Nishimura, Hirokazu Shiraishi, Satoshi Teramukai, Satoaki Matoba.
1 9

App-Based Relaxation Exercises for Patients With Chronic Neck Pain: Pragmatic Randomized Trial (e31482)
Daniel Pach, Susanne Blödt, Jiani Wang, Theresa Keller, Beatrice Bergmann, Alizé Rogge, Jürgen Barth, Katja Icke, Stephanie Roll, Claudia Witt.

Enabling Research and Clinical Use of Patient-Generated Health Data (the mindLAMP Platform): Digital Phenotyping Study (e30557)
Aditya Vaidyam, John Halamka, John Torous.
Human-Centered Design of Mobile Health Apps for Older Adults: Systematic Review and Narrative Synthesis

Zethapong Nimmanterdwong¹, MD; Suchaya Boonviriya², BNS, MSc; Pisit Tangkijvanich¹, MD, PhD

¹Department of Biochemistry, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand
²Center of Excellence in Hepatitis and Liver Cancer, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

Abstract

Background: The world is aging. The number of older patients is on the rise, and along with it comes the burden of noncommunicable diseases, both clinical and economic. Attempts with mobile health (mHealth) have been made to remedy the situation with promising outcomes. Researchers have adopted human-centered design (HCD) in mHealth creation to ensure those promises become a reality.

Objective: This systematic review aims to explore existing literature on relevant primary research and case studies to (1) illustrate how HCD can be used to create mHealth solutions for older adults and (2) summarize the overall process with recommendations specific to the older population.

Methods: We conducted a systematic review to address the study objectives. IEEE Xplore, Medline via Ovid, PubMed, and Scopus were searched for HCD research of mHealth solutions for older adults. Two independent reviewers then included the papers if they (1) were written in English, (2) included participants equal to or older than 60 years old, (3) were primary research, and (4) reported about mHealth apps and their HCD developments from start to finish. The 2 reviewers continued to assess the included studies’ qualities using the Mixed Methods Appraisal Tool (MMAT). A narrative synthesis was then carried out and completed.

Results: Eight studies passed the eligibility criteria: 5 were mixed methods studies and 3 were case studies. Some studies were about the same mHealth projects with a total of 5 mHealth apps. The included studies differed in HCD goals, target groups, and details of their HCD methodologies. The HCD process was explored through narrative synthesis in 4 steps according to the International Standardization Organization (ISO) standard 9241-210: (1) understand and specify the context of use, (2) specify the user requirements, (3) produce design solutions to meet these requirements, and (4) evaluate the designs against requirements. The overall process and recommendations unique to older adults are summarized logically with structural order and time order based on the Minto pyramid principle and ISO 9241-210.

Conclusions: Findings show that HCD can be used to create mHealth solutions for older adults with positive outcomes. This review has also summarized practical HCD steps and additional suggestions based on existing literature in the subfield. However, evidence-based results are still limited because most included studies lacked details about their sampling methods and did not set objective and quantifiable goals, leading to failure to draw significant conclusions. More studies of HCD application on mHealth for older adults with measurable design goals and rigorous research strategy are warranted.

(JMIR Mhealth Uhealth 2022;10(1):e29512) doi:10.2196/29512
However, while policy barriers tend to impede new innovations at the individual level to a higher level of the policy governing its use. Understanding of the end users and (2) failing business models cannot scale to their own target population and fail to achieve sustainability. Although mHealth has remarkable potential, most projects focused on pediatric asthma management reported increased treatment adherence in 13 studies, reduced exacerbations in 5, and improved quality of life in 4 [17].

mHealth and telehealth are the means to achieve timely and accurate health management; they help enable a seamless sharing of medical information between all those involved, creating the so-called connected health environment that the current trend strives for [12]. Successful integration of such innovations is believed to ensure universal health coverage, reduce health care costs, and improve clinical outcomes [9]. There were 5.2 billion mobile phone users at the end of 2019 with the estimation that the number will reach 5.8 billion by 2025, roughly 70% of the entire human population [13]. Diffusion of health care through a mobile medium in such a large populace will surely guarantee impact on a global scale.

Real-world mHealth implementations across the globe are committed to educating patients, offering easier access to medical care, improving medical data storage and transfer, empowering health care providers, and boosting the efficiency of its institutions [14]. The synthesis of clinical evidence in the field is also on the rise. A meta-analysis of 11 lifestyle modification apps reported a significant reduction in the mean HbA1c of the users in both short- and long-term observations [15]. Self-management interventions in 24 studies were shown to be able to decrease both systolic and diastolic blood pressures in patients with hypertension [16]. One systematic review that focused on pediatric asthma management reported increased treatment adherence in 13 studies, reduced exacerbations in 5, and improved quality of life in 4 [17]. Although mHealth has remarkable potential, most projects cannot scale to their own target population and fail to achieve the intended results. This can be attributed to (1) poor understanding of the end users and (2) failing business models [18]. Barriers to user adoption of mHealth can range from an individual level to a higher level of the policy governing its use. However, while policy barriers tend to impede new innovations or hinder the successful ones from a larger adoption [19], user-related barriers pose a more tangible challenge as no one might use the technology in the first place. A survey in the United States showed that about half of those who have downloaded health apps stop using them eventually [20]. The cause of this begins when inadequate user involvement makes it impossible to draft concise software requirements [21], which results in poor user acceptance and failure to scale [22].

These issues get even more complicated with older adults. The United Nations defines older persons as those aged over or equal to 60 or 65 years; now, over 703 million people are aged over 65 years, and that number is projected to double by 2050 [23]. Moreover, about 2 out of 3 older adults suffer from multiple chronic diseases [24], a condition to which mHealth proves to be a highly possible solution [3-7]. A myriad of frameworks and techniques have been employed to ensure the success of mHealth development and implementation with varying outcomes. Suggestions from research up to date stress the importance of having an in-charge multidisciplinary team working together with real end users rather than giving them the finished product out of the blue [25]. The International Organization for Standardization (ISO) 9241-210 further elaborates this concept in the term “human-centered design” (HCD) as the “approach to systems design and development that aims to make interactive systems more usable by focusing on the use of the system and applying human factors/ergonomics and usability knowledge and techniques”, in which the word “human-centered” is used to highlight that the process includes all stakeholders and not just the users [26]. Thus, in this review, the term “user centered” will be referred to as “human centered” to reflect its definition better.

Review Objective and Question

In searching for the best methodology to create the most usable mHealth, many have put the said value at the core of their work: having the humans at the center of focus. This review aims to explain how HCD can be applied to create mHealth suitable for older adults and to summarize the overall process with recommendations from relevant primary research studies of mHealth design and development.

The research question of this review is the following: How can HCD be used to create mHealth solutions for older adults? This issue was formed during the first author’s attempt to develop an mHealth app for older adults to solve their current pain points in a geriatric wellness clinic. Despite the constant mentioning of HCD, previous scoping searches of literature bear a heterogeneous group of research studies differing in interpretation, execution, and the extent of evaluation. The need for further clarification on the procedural details is identified.
Methods

Design

A systematic approach following Siddaway et al’s guide [27] was employed to ensure a robust acquisition of the existing literature related to the topic with a method as reproducible, transparent, and unbiased as possible. The review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [28] (Multimedia Appendix 1). Detailed methods are described in the review as no prior protocol was published.

Eligibility Criteria

Textbox 1 presents the eligibility criteria. As this review aims to draw from studies of a relatively new and emerging subfield of study, the criteria are inclusive. However, a certain degree of clarity in participants, qualitative or quantitative methods, analysis of the results, and discussions of the implementation results are required. Moreover, to best answer the review question, the included studies have to have these 3 key steps starting from (1) designing solutions based on existing problems, (2) developing the designed solutions, and (3) evaluating the developed solutions, all stated to be conducted in accordance with the HCD philosophy.

Textbox 1. Eligibility criteria.

<table>
<thead>
<tr>
<th>Inclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Community, primary, secondary, or tertiary care.</td>
</tr>
<tr>
<td>• Any qualitative, quantitative, or mixed methods study of original primary research.</td>
</tr>
<tr>
<td>• Participants must include, but not limited to, older adults (aged ≥60 years).</td>
</tr>
<tr>
<td>• Design goals must focus on mobile health (mHealth) solutions in the form of mobile apps intended for older adults.</td>
</tr>
<tr>
<td>• Study procedures must be in line with the human-centered design (HCD) philosophy.</td>
</tr>
<tr>
<td>• Studies must include details of mHealth apps and their development process, participants, design goals, and some implementation data.</td>
</tr>
<tr>
<td>• Studies depicting different processes of the same product/project are included. For example, an mHealth project might have 2 separate papers such as 1 for design and 1 for evaluation; both are included in this review.</td>
</tr>
<tr>
<td>• Trial and pilot studies are included.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Non-English language papers.</td>
</tr>
<tr>
<td>• Any type of literature review, narrative review, or systematic review.</td>
</tr>
<tr>
<td>• Studies with no relevant data or information that is of interest to the review question.</td>
</tr>
</tbody>
</table>

Search Strategy and Study Selection

Systematic searches were conducted from the following 4 databases: IEEE Xplore, Medline via Ovid, PubMed, and Scopus. To best ensure comprehensive search results, search strings were compiled from keywords of the review question. Listed below are those strings with their corresponding similar terms:

• “mHealth” OR “mobile health”, for the app to be reviewed;
• “human centered” OR “human centered” OR “user centered” OR “user centered”, the approach in question;
• “design” OR “development”, the process required;
• “usability”, an outcome of HCD according to ISO 9241-210;
• “elderly” OR “older adults” OR “geriatric”, the target population.

Each group of strings was put together with the “AND” Boolean operator in the search engines as all of the above key terms were required by the set eligibility criteria. No date range was set. Manual searches on Google Scholar and the references of the eligible papers were also conducted to identify possible additional relevant papers for screening. All searches were performed by a single reviewer (ZN) on the same day (November 12, 2020). The reason why the ACM Digital Library was not included is discussed in the “Limitations” section.

Microsoft Excel was used to record and manage the search results; duplications were removed. Two independent reviewers (SB and ZN) screened the deduplicated results by titles and abstracts. The full-text screening was done by the same reviewers using the eligibility criteria from Textbox 1. The results were in agreement. The reviewers then proceeded to appraise the study qualities using the Mixed Methods Appraisal Tool (MMAT) for mixed methods studies [29]. Disagreements were resolved through discussions. As this review aimed to be inclusive, study quality was not used to exclude any paper from the review but rather to inform about the present research quality of the existing literature of interest. We chose MMAT as our appraisal tool because (1) it can appraise the heterogeneous methodologies of design studies and (2) its methodological focus helps reflect on the existing research critically. Table 1 presents the qualities of the included studies appraised by MMAT.
Table 1. Quality appraisal of included studies.

<table>
<thead>
<tr>
<th>Studies</th>
<th>Criteria from the Mixed Methods Appraisal Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.1&lt;sup&gt;a&lt;/sup&gt;  1.2&lt;sup&gt;b&lt;/sup&gt;  1.3&lt;sup&gt;c&lt;/sup&gt;  1.4&lt;sup&gt;d&lt;/sup&gt;  1.5&lt;sup&gt;e&lt;/sup&gt;  4.1&lt;sup&gt;f&lt;/sup&gt;  4.2&lt;sup&gt;g&lt;/sup&gt;  4.3&lt;sup&gt;h&lt;/sup&gt;  4.4&lt;sup&gt;i&lt;/sup&gt;  4.5&lt;sup&gt;j&lt;/sup&gt;  5.1&lt;sup&gt;k&lt;/sup&gt;  5.2&lt;sup&gt;l&lt;/sup&gt;  5.3&lt;sup&gt;m&lt;/sup&gt;  5.4&lt;sup&gt;n&lt;/sup&gt;  5.5&lt;sup&gt;o&lt;/sup&gt;</td>
</tr>
<tr>
<td>Cornet et al [30]</td>
<td>1  1  1  1  1  N/A  N/A  N/A  N/A  N/A  N/A  N/A  N/A  N/A</td>
</tr>
<tr>
<td>Cornet et al [31]</td>
<td>1  1  1  1  1  0  0  1  0  1  1  1  0  1  0</td>
</tr>
<tr>
<td>Fortuna et al [8]</td>
<td>1  0  0  1  1  0  0  0  0  1  0  0  0  0  0</td>
</tr>
<tr>
<td>Harte et al [32]</td>
<td>1  1  1  1  1  0  1  1  0  1  1  1  1  1  1</td>
</tr>
<tr>
<td>Harte et al [33]</td>
<td>1  1  1  1  1  N/A  N/A  N/A  N/A  N/A  N/A  N/A  N/A  N/A  N/A</td>
</tr>
<tr>
<td>Petersen et al [34]</td>
<td>1  1  1  1  1  0  0  1  0  0  1  1  1  1  0</td>
</tr>
<tr>
<td>Srinivas et al [35]</td>
<td>1  1  1  1  1  0  0  0  0  1  0  1  1  1  0</td>
</tr>
<tr>
<td>Stara et al [36]</td>
<td>1  1  1  1  1  N/A  N/A  N/A  N/A  N/A  N/A  N/A  N/A  N/A  N/A</td>
</tr>
</tbody>
</table>

<sup>a</sup>Is the qualitative approach appropriate to answer the research question?
<sup>b</sup>Are the qualitative data collection methods adequate to address the research question?
<sup>c</sup>Are the findings adequately derived from the data?
<sup>d</sup>Is the interpretation of results sufficiently substantiated by data?
<sup>e</sup>Is there coherence between qualitative data sources, collection, analysis, and interpretation?
<sup>f</sup>Is the sampling strategy relevant to address the research question?
<sup>g</sup>Is the sample representative of the target population?
<sup>h</sup>Are the measurements appropriate?
<sup>i</sup>Are the confounders accounted for in the design and analysis?
<sup>j</sup>Is the statistical analysis appropriate to answer the research question?
<sup>k</sup>Is there an adequate rationale for using a mixed methods design to address the research question?
<sup>l</sup>Are the different components of the study effectively integrated to answer the research question?
<sup>m</sup>Are the outputs of the integration of qualitative and quantitative components adequately interpreted?
<sup>n</sup>Are divergences and inconsistencies between quantitative and qualitative results adequately addressed?
<sup>o</sup>Do the different components of the study adhere to the quality criteria of each tradition of the methods involved?

P/N/A: not applicable.

**Data Extraction**

One independent reviewer (ZN) performed data extraction from the 8 eligible papers. The information from 5 mixed methods studies included (on the data extraction form) the year of the study, the country of the study, the name of the project (if stated), study design, design goals, participants, study methods, quantitative or qualitative data used, results, and key discussions. The information from the other 3 case studies included the year of the study, the country of the study, goals, and results. All extracted texts were manually typed in Microsoft Excel.

**Synthesis of Results**

Because of the heterogeneous nature of the included studies, narrative synthesis was chosen. Following Popay et al’s guide [37], the narrative data synthesis was performed iteratively between the 4 key elements as explained in Textbox 2.
Textbox 2. Key elements for the narrative data synthesis.

- Developing a theory of how the intervention works, why, and for whom

Previous studies were carried out under the same hypothesis that human-centered design (HCD) helps make a more usable system for its users. This review adopted that same assumption and aimed to elaborate on how HCD works, especially for older adults, in steps.

- Developing a preliminary synthesis of findings of included studies

Textual descriptions together with tabulation were chosen to summarize and display the extracted data. A recurring concept was identified across the studies: the HCD process. To ensure transparency, suggested HCD activities from ISO 9241-210 were chosen to categorize these patterns into 4 steps as follows: (1) understand and specify the context of use, (2) specify the user requirements, (3) produce design solutions to meet these requirements, and (4) evaluate the designs against requirements [26].

- Exploring relationships in the data

Qualitative case descriptions were used to explore details and findings among included studies that correlate with each theme/step. A conceptual diagram was then created to answer the review question. The diagram was structured according to the Minto pyramid principle, using the following rules: (1) ideas at any level in the pyramid must always be summaries of the ideas grouped below them, (2) ideas in each grouping must always be the same kind of idea, and (3) ideas in each grouping must always be logically ordered [38].

- Assessing the robustness of the synthesis

All included studies were appraised by Mixed Methods Appraisal Tool (MMAT), and the synthesis process was reflected on critically.

Results

Study Selection

Figure 1 shows the selection process of the included studies. The initial search yielded 44 studies, of which 40 were from the 4 databases and the other 4 were from Google Scholar. A total of 25 studies remained after the removal of duplications. Two independent reviewers (SB and ZN) screened titles and abstracts according to the criteria. The remaining 13 full-text studies were then assessed by the same 2 separate reviewers for eligibility. Five studies were excluded, as shown in Figure 1. Eventually, 8 studies were retained for this systematic review.
Study Characteristics

Five studies were mixed methods, employing both quantitative and qualitative techniques to design and evaluate mHealth apps. Three were qualitative case studies focusing on describing the methodology and problems found during the process with little or no actual quantitative or qualitative data from the research shown. One of the 3 case studies was about the same process from the same project of another included mixed methods study; it was included for its qualitative reflections on the design and development processes. All studies stated clear aims of the research and were deemed relevant to this review question.

To illustrate the overall quality of the included studies, their quality appraisal scores were reported in the MMAT-suggested format [39] in Table 1. Of the 8 included studies, all were rated to have adequate quality in their qualitative part. However, only 1 study had a passable rating of 60% in its quantitative part, while the rest were rated poor. The quantitative criteria that all studies failed were about the sampling strategy and the account for possible confounders. The description of the processes was lacking, leading to questionable results and interpretation. This issue was further explored in the narrative synthesis of results.

Table 2 summarizes HCD processes of the 5 included mixed methods studies. Four of the studies were conducted in the United States, with only 1 study conducted in Ireland. Two studies under the same project focused on patients with heart failure [31,35]. One study aimed at patients with psychiatric disorders [8]. The other 2 studies dealt with fall risk assessment and detection [32] and sarcopenia prevention [34], respectively.

In addition, of the 3 included qualitative case studies, 2 reflected on the same project as the mixed methods study dealing with falls in Ireland [33,36]. By contrast, the remaining 1 study reflected on a different project targeted at patients with cardiac implantable electronic devices (CIEDs) in the United States [30].
<table>
<thead>
<tr>
<th>Study</th>
<th>Project</th>
<th>Setting</th>
<th>Design goal</th>
<th>Participants</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cornet et al [31]</td>
<td>Engage</td>
<td>Academic health center, the United States.</td>
<td>To evaluate and test the usability of a self-managing heart failure system for older adults developed in a study by Srinivas et al [35].</td>
<td>(1) 13 older adults and (2) 2 caregivers</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Study I: (1) A structured interview was used to assess participants’ daily self-management routines and technology familiarity. (2) The think-aloud method was employed as each participant completes 8 given tasks on the system. (3) Feedback from the patients after they finish was used.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Study II: (1) The system was re-designed after Study I. (2) A structured interview was used. (3) The think-aloud method was employed as each participant completed a given scenario in which he/she was to act as if he/she were the assigned fictitious character. (4) Feedback from the patients after they finish was used. (5) SUS&lt;sup&gt;a&lt;/sup&gt; was used after usability evaluations. (6) NASA-TLX&lt;sup&gt;b&lt;/sup&gt; was used after usability evaluations.</td>
</tr>
<tr>
<td>Fortuna et al [8]</td>
<td>...&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Specialized center, the United States.</td>
<td>To incorporate an existing psychosocial intervention into a selected mobile platform.</td>
<td>Phase I and Phase II: (1) authors; Phase III: (1) older adults and (2) experts; and Phase IV: (1) 10 middle-aged and older adults</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Phase I: (1) A literature review was done to identify requirements.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Phase II: (1) A literature review was done to find a suitable existing mobile platform.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Phase III: (1) The interdisciplinary panel of end users and experts work together to incorporate an existing psychosocial intervention into the chosen mobile platform.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Phase IV: (1) The think-aloud method was employed as each older adult goes through task-based usability testing. (2) Feedback from patients was collected. (3) Surveys based on SUS; Post-Study System Usability Questionnaires; and USE&lt;sup&gt;d&lt;/sup&gt; questionnaires were used after each usability testing. (4) The ability to perform tasks without help was recorded in percentage.</td>
</tr>
<tr>
<td>Study</td>
<td>Project</td>
<td>Setting</td>
<td>Design goal</td>
<td>Participants</td>
<td>Methods</td>
</tr>
<tr>
<td>-----------------------</td>
<td>----------------------------------------------</td>
<td>--------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Harte et al [32]</td>
<td>Wireless Insole for Independent and Safe Elderly Living</td>
<td>Academic health center, Ireland.</td>
<td>To develop, assess, and enhance usability and user experience of a mobile app of a connected health system designed for fall risk assessment and fall detection.</td>
<td>Phase I: (1) 10 experts and (2) 12 older adults; Phase II: (1) 10 experts from Phase I; and Phase III: (1) 10 older adults from Phase I.</td>
<td>Phase I: (1) Likert scales were used to rate mock-ups called use cases. (2) The think-aloud method was employed during use case analysis. (3) Self-reported measures of the experts were collected. (4) Visual perception and cognitive processing metrics of older adults were collected.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Petersen et al [34]</td>
<td>Academic health center, the United States.</td>
<td></td>
<td>To create a mobile app for older adults to monitor their use of a Bluetooth-connected resistance band for sarcopenia prevention.</td>
<td>Round 1: (1) 6 older adults; Round 2: (1) 3 clinicians and (2) 4 older adults; Round 3: (1) 3 clinicians and (2) 6 older adults</td>
<td>Phase II: (1) Likert scales were used to rate the paper prototypes based on use cases by experts. (2) ASQ&lt;sup&gt;e&lt;/sup&gt; and chosen usability metrics were used to rate the developed mobile working prototypes by experts after scenario-based usability testing. (3) The think-aloud method was employed during experts' mobile working prototype runs.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:**
- **ASQ<sup>e</sup>:** Assessment of Subjective Quality
- **SUS:** System Usability Scale
- **NASA-TLX:** NASA Task Load Index
- **USE score:** User Experience Score

**References:**
- Harte et al [32]
- Petersen et al [34]
### Study Participants Design goal Methods

<table>
<thead>
<tr>
<th>Study</th>
<th>Project</th>
<th>Setting</th>
<th>Design goal</th>
<th>Participants</th>
<th>Methods</th>
</tr>
</thead>
</table>
| Srinivas et al [35] | Engage  | Specialized center, the United States. | To design, develop, and evaluate a consumer-facing health information technology system that supports heart failure self-care. | Phase I: (1) 63 older adults, (2) 35 caregivers, and (3) additionally data on 66 patients obtained from other literature; Phase II: (1) experts; Phase III: (1) 1 expert and (2) 5 older adults | • Phase I: Major themes of the app were synthesized from data gathered through direct observations at patient outpatient visits, standardized surveys on patient self-care, patients' electronic medical record reviews, and semi-structured interviews focused on patient self-care.  
• Phase II: Core activities of the app were determined through educating, brainstorming, and design sessions of the research team.  
• Phase III continues in Cornet et al (2017) [31]: (1) heuristic evaluation done by the team's expert identified and classified usability flaws. (2) Structured interviews focusing on patients' self-care routines were done before usability testing. (3) The think-aloud method was employed during laboratory-based usability testing of the developed prototype as each older adult goes through the tasks given on a mobile. (4) Questionnaires adapted from the SUS were used after each usability testing. |

---

**Overview**

All 5 mHealth projects, from the included 5 mixed methods studies and 3 case studies, have the 4 key steps from ISO 9241-210 in their HCD processes, albeit described and mentioned to varying degrees. This section explores and illustrates these recurring steps across all included studies using the qualitative case description technique. All 8 studies are summarized and described in 4 HCD steps. Each step has 5 paragraphs representing a total of 5 mHealth projects: the first for patients with heart failure [31,35], the second for patients with psychiatric disorders [8], the third about falls in the elderly [32,33,36], the fourth for sarcopenia prevention [34], and the fifth about CIEDs [30].

**Step 1: Understand and Specify the Context of Use**

Understanding the context of use such as the end users, their current tasks, key activities, and working environment is essential to the design process; it helps guide how solutions should be tailored and set practical goals for the project [26].

Srinivas et al [35] used various HCD frameworks to develop an mHealth app that helped older patients with heart failure to improve their self-care engagement, health behaviors, and knowledge of the disease. In 2 years, the researchers collected data from 65 older patients with heart failure and 35 caregivers through direct observations at outpatient clinics, electronic medical record reviews, and semi-structured interviews; the patients' health care routines, health literacy, environments, and supports were the priority. They conducted these field-based investigations in an academic medical center in Southeast United States. In addition, the authors included 66 other patient data from the United States and Singapore in an urban emergency setting. Details on the sampling method and rationales for the number of patients were not provided. It was also noted that not all data could be utilized fully in the design process. The qualitative quality of this study was adequate.

Fortuna et al [8] aimed to integrate self-management intervention into a mobile app for middle-aged and older patients with psychiatric disorders to promote self-care for better health outcomes. However, rather than obtaining data directly from potential users, the researchers gathered rationales and pain points of the project from a literature review. Details about the method were not specified in the paper. No quality appraisal of the included literature was presented. They then used the review results in the subsequent design. For example, integrating an existing intervention to an existing mobile platform was chosen over developing a new one because it was more practical. Characteristics specific to the elderly such as declining cognitive functions affecting their self-management and motivation, multimorbidity, and limited digital literacy were considered. The researchers also decided the intervention to be implemented based on the literature review: Integrated Illness Management and Recovery (I-IMR), an evidence-based medical practice for...

---

**Notes**

- **SUS**: System Usability Scale.
- **NASA-TLX**: NASA-Task Load Index.
- **Not stated.**
- **USE**: Usefulness, Satisfaction, and Ease of Use.
- **ASQ**: After Scenario Questionnaire.
psychiatric patients, was cited to be promising and thus chosen. The authors reported successful implementation and noted that identifying the unique needs of the intended users to guide the design process helped build a more usable product. The qualitative methodologies were appraised to be of adequate quality.

Stara et al [36] integrated HCD into the development process of their connected health system: the Wireless Insole for Independent and Safe Elderly Living (WISESL), consisting of a pair of chargeable insoles with Bluetooth transmission, its charger, a smartphone app, a gait analysis desktop application, and an administrative web application. The authors drafted a preliminary concept of the system and then discussed it in 3 focus groups with 6 older adults and 6 stakeholders in each group; the focus groups were conducted at 3 separate sites: a primary care center in Ireland, a tertiary care center in Israel, and a specialized center in Italy. The sampling method was not specified. The qualitative quality is adequate. The authors concluded that barriers to technology-enabled care acceptance in older adults were related to security, intrusiveness into their home environment, lack of control, confidentiality, and usability issues worsened by aging. Thus, involving users early in the process proves vital in crafting a health care technology that matches actual older adult user needs, with elderly friendly user interfaces and safety being a priority.

Petersen et al [34] used HCD to develop an mHealth app featuring exercise videos to work with a Bluetooth-connected resistance band that together would help health care providers monitor older adults’ exercise progress for sarcopenia prevention. A convenience sampling method was used. Six older adults were recruited from a primary care clinic at an academic health center in the United States. The researchers then conducted semistructured interviews to assess the patients’ general views regarding mHealth, their current activities, and their opinions of the Bluetooth-connected band and sample exercise videos. They further explored the participants’ opinions in using technology to help with their exercise therapy. The quality of the study was appraised to be adequate. The participants had positive responses to the idea. All had experience using smartphones. Notes from these interviews were then used as key information to guide further design processes.

Cornet et al [30] implemented HCD in developing an mHealth app that shared the information stored in CIEDs of patients with heart failure with the patient themselves. In 3 months, 24 older patients with heart failure, 12 of whom had CIEDs, were recruited from a major health system in the Midwestern United States for semistructured interviews to gain context about their health decision-making processes. No sampling method was stated. The interview utilized 2 notable approaches: (1) the critical incident technique, which involves asking the interviewee to recall a particular past event to gain insights through their actions and experience at the time; and (2) the think-aloud method, which lets the interviewee talk about what he/she was currently doing or would do in a given event. The researchers then analyzed and synthesized the gathered data into 2 outputs: (1) personas, a design technique that groups users based on their behaviors; and (2) use-case scenarios (or as-is scenarios), another design technique that depicts how users make decisions in hypothetical situations. These outputs were then cross-checked with 2 patient advisors, older adult patients who volunteered to help with design, and a group of 7 clinician experts from the same major health system. The patient advisor meeting was held early to gain additional inputs and feedback to help the team make more relatable personas and use-case scenarios. The clinician meeting was held later and focused more on the validity and feasibility of the subsequent processes. Besides, direct observations at the CIED clinic and meeting with 2 cardiologists were also done with the same objectives. The methods were appraised to be adequate in quality. As the paper is a case study, challenges and recommendations by the authors were reported. First, logistics issues including but not limited to compensation, conflicts of interests, older adult limitations, patient data, recruitment criteria, and stakeholder meetings need to be addressed or consulted with professionals to ensure efficiency and efficacy. Second, stakeholders should be involved early in the design process, and their roles should be identified clearly in how active they would be; for example, it might just be getting informed about the process, giving their opinions to the team, or having specific tasks given to them. The authors also added that more roles are not always better, and stakeholder involvement should be carefully balanced. Third, an adequate recruitment method should be employed to secure a representative group of potential users. Also, a selection of stakeholders who work well with the development team is key. Fourth, direct and timely communication between development team members and relevant stakeholders is recommended, although it might be difficult to achieve at times.

**Step 2: Specify the User Requirements**

The second step of HCD focuses more on synthesizing further outputs from the first step. The goal is to derive what the users need to do and their objectives based on the gathered context and then set a clear statement of user requirements for the solution designs [26]. User requirements lay down the performance or criteria should be measured to evaluate the product. These requirements are often created along with other requirements of the product such as the requirements of the system stating that the system needs to be able to do a certain task because it will help users accomplish their goals.

Srinivas et al [35] reported a successful translation of major themes from the gathered data and created a set of requirements for the subsequent design. Thematic analysis was done to identify user needs; it was concluded that the patients lack adequate health information and communication regarding their conditions and disease progresses, they are disengaged from their self-care due to the added burden, and they are not equipped with practical knowledge or tools for optimal self-management. The authors then held educating sessions with the design team, composed mainly of experts from technical and HCD backgrounds, on the phenomena of interest through various media and means from the collected data. Next, brainstorming sessions were held. The requirements were derived from the previously identified major themes from the research: the system needed to be viewable by the patients and potentially their health care providers, simple to use, and potentially...
complementing well with their self-care routines, and customizability. Details of the performance goals to be evaluated were not stated.

Fortuna et al [8] derived requirements from their literature review of user interfaces for older adults, I-IMR contents, and the interdisciplinary panel of end users and experts recruited from the site where the project was intended to be launched. I-IMR is a clinical psychological intervention that requires both health care providers and patients to work together in 10 training modules/sessions covering 4 topics on psychoeducation, behavioral tailoring, relapse prevention training, and coping skills training over 8-10 months. Adapting this face-to-face intervention and its contents into a suitable mobile experience for middle-aged and older patients was key to the project. Details about the performance goals were not stated.

Stara et al [36] stated in their study that user requirements for the WIISEL-connected health system were defined by the design team together with 18 older adults who were potential users and 18 stakeholders who were geriatricians, neurologists, nurses, and physical therapists in 2 sessions of focus groups. Details of the participants and the final user requirements and performance goals were not provided.

Petersen et al [34] concluded pain points from their previous research regarding the exercise videos as follows: specific movements were hard to identify from low-contrast backgrounds, and instruction sounds were not heard clearly. Participants also stated that big and clear repetition numbers would help them better keep track, feedback and instructions would help them finish the exercises at home, and tablets were preferred as they have large screens. The authors did not show an explicit statement of user requirements or performance goals in the paper.

Cornet et al [30] did not provide details regarding the process of writing user requirements in their case study.

**Step 3: Produce Design Solutions to Meet These Requirements**

This HCD step focuses on designing how the users interact with the system based on the requirements from the previous step [26]. HCD strove for the best user experience. The process needs to be iterative and flexible to address user needs and requirements that are often hard to identify completely in 1 cycle. The outputs from this step are also used to explain and communicate the design concepts with stakeholders, simulate possible scenarios of its uses, and ultimately specify how the system is to be developed.

Srinivas et al [35] created design solutions from the requirements specified. The team members raised diverging ideas from those requirements. They then worked together to converge those ideas into 4 main potential design solutions: (1) a short-term intervention of 30 days to encourage user adoption of the system, (2) an avatar representing the results of different self-care routines to teach users about cause and effect in a more engaging way, (3) a function that allowed users to set and keep track of their goals to promote health behaviors, and (4) a tool that helped enhance clinical visit experience to improve communication and collaboration with their health care providers. Finally, the team decided to develop the mHealth solution based on the 30-day intervention idea. The system was to have 3 main modules to serve all user requirements previously set: (1) LOG, for users to log their health information; (2) HINT, a collection of short materials about heart failure disease and self-care; and (3) GOAL, gamified daily goals for better self-care behaviors. User–system interaction and user interfaces were then developed in subsequent design sessions composed of 3-6 members of the design team. Paper prototypes, Microsoft PowerPoint wireframes, and software prototypes were developed successfully. The authors noted that although clinical experts were consulted from time to time during this step, stakeholders, including end users, did not really participate in the design process; logistics issues and not knowing how to involve older adults were accountable for this approach. The authors also added that their waterfall approach to development, meaning the design process was linear and required time before evaluation, caused delays in solving design problems that could be prevented if a more agile approach was adopted.

Fortuna et al [8] created scenarios of uses, user–system interaction, and user interfaces based on the identified user needs and requirements. The system was designed to have the following: an ability to be customized, a tracking system to show users their progress, a monitoring system to send data back to health care providers, and a messaging system from health care providers for more human interaction and a smoother workflow. The text contents from I-IMR manual, originally intended for clinicians, were also modified to fit smartphone pages and rewritten using the Flesch-Kincaid Grade Level formula in Microsoft Word to a simpler sixth-grade level. The 4 core topics of I-IMR were also re-designed to fit the mobile experience. Psychoeducation used short videos that showed clinicians teaching self-management techniques to patients to help users master the skills. Behavioral tailoring utilized educating modules, a medication schedule function, and a reminder system to make patients take their medication on time. Relapse prevention training, usually done by exploring a patient’s experience and identifying triggers to create a prevention plan for a possible relapse, offered an already made plan that was accessible and editable on mobile at any time. Coping skills training also used videos as media to equip users with the tools to help them in the real world, for example, relaxation videos that guide users to self-soothe and calm themselves down. In addition, issues regarding data security and mHealth user disengagement were addressed: the project adhered to Health Insurance Portability and Accountability Act (HIPAA) compliance and involved health care providers to encourage adoption.

Stara et al [36] stated in their study that the subsequent design process of the smartphone app of the WIISEL system is elaborated in 2 studies included in this review: 1 mixed methods study focused on the findings by Harte et al [32] and the other case study focused on the HCD methodology also used by Harte et al [33,36]. Rapid development was employed to create, test, and produce 4 versions of prototypes, 2 on paper and 2 on mobile. The first paper prototypes (also called “use cases”) consisting of scenarios of use, descriptive end user profiles, storyboards, and interface mock-ups were created from the
opinions of all project stakeholders. These use cases were based on key activities that the users needed to carry out. Ten multidisciplinary experts and 12 older adults were then recruited by a purposive sampling method to analyze these use cases. This analysis quickly identified usability problems that were in turn fixed by the development team. The first mobile prototypes together with user manuals and the updated, second paper prototypes were then produced. The manuals were created to help address usability problems that could not be fixed by the development team such as the built-in buttons, the operating system keyboard design, the impracticality of an automatic data sync, and the connection limitations. The same experts then evaluated the first mobile prototypes simultaneously with the guide of the second paper prototypes and the user manuals. The results from this second expert analysis were then used to design the second mobile prototype for another usability test with end users. The authors added that multiple inputs from the relevant stakeholders, although divergent in nature, are essential to HCD and could be obtained only with enough rounds of iterations. Thus, the rapid cycles of using paper prototypes and expert evaluation for fast feedback before end user usability testing are recommended.

Petersen et al [34] updated the exercise videos from user pain points and created mobile prototypes as black and white wireframes showing simple outlines of the designed user interfaces. The researchers then presented exercise videos and the wireframes with different design approaches to 3 clinicians and 4 older patients from an academic health center in the United States for additional inputs. The pain points gathered from the older patients about the videos and the wireframes were as follows: the video instructions should be slower and have more details, the videos should have subtitles with large fonts, the video sound frequency should be adjustable, and the progress bars in the wireframes should be vertical. By contrast, the clinicians were content and also suggested the use of the Borg Scale of Perceived Exertion to measure each exercise difficulty in comparison to the others. Finally, the team created the interactive mobile prototypes featuring playable updated exercise videos with clear instructions and the colored user interfaces for usability testing in the next step. The authors noted that health management is a process that needs both health care providers and patients; therefore, the more stakeholder groups involved, the more complete the design of the mHealth that aims to assist the process.

Cornet et al [30] produced and improved 4 versions of prototypes through iterative prototyping and testing. The mHealth app for patients with heart failure with CIEDs was designed to have 4 main features: a heart health score derived from CIED data for the patient audience, self-assessments covering topics of recommended self-care routines, guides for better heart failure self-management, and logs showing data from CIEDs. The first prototype design was reported to take 5 months to complete. The 2 patient advisors from the research phase helped review this early prototype design. The later 3 prototypes then took 2, 1, and 3 months, respectively, with the final prototype going through refinement for heuristic evaluation for another 1 week. Feedback from each prototyping and testing was used to improve the later versions. More details of the process were not stated in this case study; however, challenges and recommendations were reported. First, the authors found that design solutions should be based on evidence gathered from the potential users of the project to avoid bias or assumptions of the design team. This challenge benefits from stakeholder involvement and rapid cycles of testing and feedback. Second, design solutions and features should be prioritized and focused on. Grouping these solutions into modules and structurally planning how to develop and test them to determine what works and what does not could help simplify the process. The third point elaborates on the first point but focuses on the feasibility of the proposed design solutions: to balance the design team’s creativity with practicality. All limitations or regulations regarding the mHealth and its implementation should be worked out properly with the stakeholders to avoid project failure.

Step 4: Evaluate the Designs Against Requirements

The human-centered evaluation activity is vital to HCD and is iterative by its nature [26]. As illustrated in the third step, producing design solutions usually follows by evaluating them to assess their abilities to fulfill the requirements, obtain user feedback, gain more user needs, and quantify the results as baselines or for comparisons.

Srinivas et al [35] conducted a series of evaluations on the wireframes and the prototypes in parallel with the design process. First, heuristic evaluation by the team’s human–computer interaction expert guided by Nielsen’s usability heuristics [40,41] was done before the software prototype development. This helped transition the static wireframes into the interactive software prototypes and identify usability flaws early in the process for correction. The authors reported 45 flaws, of which 6 were major flaws. The corrected software prototypes were then evaluated by older adults and caregivers as elaborated further in another study by Cornet et al [31]. The researchers conducted the evaluation in 2 phases in a laboratory setting: (1) a task-based usability test with 5 users and (2) a scenario-based usability test with 10 users. A total of 13 patients with heart failure and 2 informal caregivers aged over or equal to 60 years were recruited from an urban and another suburban outpatient cardiology clinic of an academic health system in the Midwestern United States. All consented to the study and were compensated with US $40 gift cards. Details of the sampling method were not specified. All participants were given mobile devices with the software prototypes installed ready for testing. Both tests involved structured interviews about users’ self-care routines and familiarity with technology at the start, the think-aloud method by talking out loud about what they were thinking during testing, and the use of standardized evaluation tools at the end. The tools include (1) System Usability Scale (SUS) consisting of 10 questions about the overall usability of the product and reporting in a score of 0-100 with 68 defined as average usability [42] and (2) NASA-TLX (NASA-Task Load Index) consisting of 6 scales to assess the cognitive load expended during product use [43]. SUS was used in both phases, but NASA-TLX was used only in the second phase. The tests were video recorded. The software prototypes were updated between the 2 usability tests. It was reported that SUS rating improved in the second phase from below average to above average. However, the
authors stated that the result could be affected by the design changes made, the different usability testing methods, and the sampling techniques. Moreover, the wording of SUS was shown to be difficult to understand to a certain group of older adults and might not reflect the real usability of the system [44]. The methodologies were appraised to be of inadequate quality. The authors added that quantitative results from the standardized tests did not capture the whole picture of the system usability issues and should be interpreted together with the qualitative results. Some older adults also showed resistance toward these usability techniques, that is, the think-aloud method was strange and the fictitious event of the scenario-based testing was counterintuitive as they had to remember the mock details that were irrelevant to them and got distracted. Logistics issues such as the locations of the testing sites, the set ups of recording tools, and the transportation of the older participants also need to be addressed.

Fortuna et al [8] conducted 2 cycles of task-based usability testing with 2 different groups of 5 participants each. The authors deemed a minimum of 5 participants could identify most usability issues [45]. All participants were middle-aged and older patients with both medical and psychiatric illnesses recruited from 2 mental outpatient programs in New Hampshire. A purposive sampling method of reviewing medical charts and reaching out to potential patients for informed consent was used. Gift cards worth US $20 were provided upon participation. The participants were given mobile devices with the app installed and a list of tasks to complete. They were orientated on how to use the devices and what the think-aloud method was. The researchers also asked the participants about the user interfaces and assigned them adapted surveys based on SUS, Post-Study System Usability Questionnaire (PSSUQ), and the Usefulness, Satisfaction, and Ease of use (USE) questionnaire. PSSUQ is an 18-item questionnaire with 7 rating scales and 1 not-applicable rating, assessing user satisfaction with the system [46]. The USE questionnaire also contains multiple items with 7 rating scales that explore 3 dimensions: usefulness, satisfaction, and ease of use [47]. All sessions took approximately an hour and were audio recorded if allowed or noted in detail if not. Updates on the app were made between the 2 cycles from user feedback: the text and video contents were shortened and the reading level was reduced from sixth grade to fourth grade. The authors reported that all participants could finish given tasks independently and both the qualitative comments and quantitative surveys had positive results, suggesting the users were satisfied with the app and would continue to use it if encouraged to do so. The mixed methods methodology was appraised to be inadequate in quality. The authors remarked that (1) future behaviors, that is, whether the patients would use the app in a real-world environment or not, were hard to predict from 1 hour of usability testing in a controlled environment, (2) the patients recruited specifically for the purpose of usability testing might lack heterogeneity and did not fully represent the intended vulnerable group of interest, and (3) technology constraints of utilizing an existing platform were reported. The authors also added the results might prove relevant and beneficial to the research of a similar fashion, and more studies on the mHealth intervention effectiveness were needed.

Stara et al [36] incorporated evaluation early in their HCD process as shown in 2 studies by Harte et al [32,33]; the authors then conducted user testing of the finished system with 54 older users [36]. During the course of producing design solutions, evaluation was done on (1) the first paper prototypes or the use cases, (2) the second paper prototypes, (3) the first mobile prototypes, and (4) the second mobile prototypes. The participants were 10 multidisciplinary experts and 12 older adults recruited using a purposive sampling method. Self-reported measures regarding the experts’ knowledge together with the older adults’ visual perception and cognitive processing metrics were reported. For the first paper prototypes, 10 experts and 12 older adults analyzed the prototypes by going through each use case. The think-aloud method was used to gather qualitative inputs, and Likert scales that asked the users to rate the user interfaces and task flows were used after each use case to obtain quantitative results. Likert scales are 5-point scale questionnaires that can be used to quantify user satisfaction; the question can be, for example, “I have no problems using the system.” Usability problems of the first paper prototypes were then identified from think-aloud transcripts and grouped according to a derived set of heuristics [48]; the problems were then given severity rating based on the results of the related use case Likert scores. The prototypes were then updated accordingly. For the second paper prototypes, the experts analyzed the updated use cases in the same manner again to compare them with the first ones: most usability problems were reported to improve. For the first mobile prototypes, scenario-based usability testing was done by the experts as if they had been first-time users. The experts were able to use user manuals during testing. The sessions were also video recorded. Think-aloud scripts together with After Scenario Questionnaire (ASQ), SUS, and 3 usability metrics (i.e., time taken to complete task, errors made, and completion rate) were used to update the user manuals and the user interfaces. The ASQ is a 3-item questionnaire regarding ease of completion, time taken to complete, and support information of the system with a 7-point scale, where a lower score indicates greater satisfaction [49]. For the second mobile prototypes, both task- and scenario-based usability testing were done by 10 older adults. They had access to the user manuals during testing. The sessions were video recorded. Data were obtained using the same methods as the first mobile prototypes with the addition of posttest interviews about general impressions of the system and NASA-TLX. The authors reported that the system achieved acceptability among end users. This is the only included mHealth project that has adequate quality appraised by M-MAT. The authors concluded in their study that (1) older adults needed clear feedback from the app due to technology unfamiliarity, but imposing feedback such as alerts or cautions should be used only when necessary to avoid anxiety for the same reason; (2) older adults were found to be uncomfortable with touchscreen keyboards, thus minimizing or simplifying them would be ideal; (3) standardized tests such as SUS, ASQ, and NASA-TLX proved to give concordant and valuable information regarding the system usability, but they should be interpreted together with the more objective metrics, such as time taken to complete task, errors made, and completion rate for more tangible results; (4) expert evaluation before end user usability testing was efficient and
thus recommended; and (5) multiple inputs from different stakeholders, despite being divergent in nature, were essential to HCD. The process took 12 months to complete, of which the first prototypes took around 6 months; it was noted that interviewing and testing all the participants in the first phase were the causes of the long duration. Finally, 54 older users then tested the WISEL system, both the mobile app and the soles, in Ireland, Israel, and Italy [36]. The usability testing had 2 stages: (1) the 3-day pilot stage had 15 participants test the system in a laboratory setting for a day and then at home for 2 days without specific instructions, and (2) the validation stage had 39 participants use the system at home for 14 days also without specific instructions. The participants completed the 12-item Quebec User Evaluation of Satisfaction with Assistive Technology questionnaire (QUEST) and SUS after each stage. Both had positive results. The authors did not detail the process as the paper was a case study. They concluded that technology acceptance was most affected by the system effectiveness but could also be positively influenced by proper user training and support.

Petersen et al [34] evaluated the wireframes and the prototype app using both qualitative and quantitative methods. Three clinicians from an academic health center in the United States participated; 6 older patients were recruited from a primary clinic of the same health center. A convenience sampling method was used. Think-aloud and verbal prompting methods were employed during testing to gather qualitative feedback from the participants. SUS and USE questionnaires were used after each participant finished testing the wireframes and the prototype app. The USE questionnaires comprise 30 questions asking about usefulness, ease of use, ease of learning, and satisfaction of the system with a 7-point scale to rate them. All sessions were audio recorded. Usability scores of the wireframes and the prototype app were calculated together and showed no statistically significant differences between the clinician and the patient participant groups with mean SUS scores of 65.8 and 66.8, respectively. The mixed methods study was appraised to be of inadequate quality. In addition, sentiment analysis of the participants’ recorded statements was done; its results were in accordance with the SUS scores. A further application of natural language processing–based Dirichlet allocation topic modeling of the recorded statements showed that clinicians and older patients had different topics of interest regarding the mHealth system. The authors concluded that (1) inclusion of different stakeholder groups was vital to HCD because each has a different perspective on the mHealth system as illustrated in the study; (2) sentiment analysis could prove useful to HCD by effectively and efficiently analyzing qualitative inputs alongside traditional usability techniques; and (3) future research on incorporating sentiment analysis and natural language processing in HCD was encouraged.

Cornet et al [30] conducted 3 usability evaluations and 1 heuristic evaluation concurrently with the iterative production of 4 prototypes. The 3 rounds of usability evaluations had 4, 8, and 12 older patients with heart failure as participants, respectively. Details of the sampling method were not specified. The software prototypes were installed on prepared smartphones for the tests. The first 2 rounds were 90-minute task-based usability testing. Participants had to complete demographic surveys and the Newest Vital Sign (NVS) health literacy screening before the test. NVS is a screening tool that takes 3 minutes to complete; it has 6 questions asking about a nutrition label that is given to the patient to assess their health literacy [50]. The think-aloud method was used during the test. Participants had to complete SUS, NASA-TLX, and user acceptance survey once finished. They were also interviewed about the system after the test. The third round was a 90-minute scenario-based usability testing simulating the use of the system in the first 10 days. The participants in this round were also required to have CIEDs. Data were gathered in the same manner as in the first 2 rounds. Finally, heuristic evaluation was done by 3 outside HCD experts for refinement of the system; the process took 2 weeks. Details of the results were not stated. However, the authors reported challenges and recommendations found as the results of their case study. First, laboratory usability testing is good for detecting general software issues (eg, user interfaces and navigation), but it might not be able to address real-world usability issues. Thus, system evaluation at the actual site of the intended setting should be considered as time and the budget allow. Second, standardized tests should be adapted to fit real users, for example, the word “cumbersome” in SUS was changed to “awkward” in the study as the older adults could understand it better. Third, the testing process tends to get complicated and lengthy with numerous tools and techniques employed, therefore the HCD team should opt to cut reducible workload, manage time between testing and analyzing, and look for the possibility of utilizing automated data collection or analysis.

Discussion

Principal Findings

This systematic review has shown how HCD can be used to create mHealth for older adults, with additional recommendations reported. Eight studies are included in this review: 5 are mixed methods studies and 3 are case studies. All studies were published recently starting from 2017 onward, suggesting that the subfield is relatively new. All were conducted in developed countries and mostly in academic or specialized health care settings. Because of the diverse methodologies and details of the included studies, we used the Minto pyramid principle and the 4 HCD steps from ISO 9241-210 to guide the creation of 3 conceptual models: Figure 2 shows a structure of HCD team members and stakeholders in the HCD process, and Figure 3 shows how HCD can be applied to create mHealth solutions for older users. The following discussion explains the models further and also explores limitations with recommendations for future research.
First, mHealth ideas, either novel or of existing concepts, should be based on what the users need, not what the creators want. As illustrated in the included studies, the authors, usually acting as the management team that oversees the project, identify and base their proposed mHealth solutions on real stakeholders both directly [30-36] and indirectly [8]. The first step of HCD investigates whether the solution fits well with the target users; this step also aims to produce outputs that ensure all HCD team members and stakeholders share the same vision. Stara et al [36] conducted focus groups to discuss their preliminary concept with relevant stakeholders. Researchers, or the functional team members, then recruit relevant stakeholders, both health care providers and patients, to gain more insight into their context of use: the users, their environment, and their current activities.

The included studies’ details and rationales for the number of participants and the sampling method were diverse and vague. For example, Fortuna et al [8] had no participant and relied solely on a literature review to identify older adults’ needs, whereas Srinivas et al [35] had a total of 100 participants and remarked that the gathered data proved to be more than they used for design. However, it should be noted that the 2 studies differ in their design goals, where the first wanted to implement a known intervention on mobile phones (app), while the second sought to identify new problems from an existing routine entirely.

Besides quantity, most included studies recommended that the sampling method include diverse groups of participants to ensure HCD solutions reflect real-world problems. Reaching out to older adults who are more physically inept or socially disadvantaged can be challenging. Fortuna et al [8] tackled this by building on the results of existing research in middle-aged and older adults with serious mental illness. Cornet et al [30] suggested recruiting key stakeholders who know how to approach such a group of patients to help. Harte et al [32] used purposive sampling and evaluated their participants’ visual perception and cognitive processing to ensure the process was inclusive. Thus, the recruitment of older adults for HCD projects should be flexible and inclusive to best serve HCD goals.
Information on the context of use was mostly gathered through qualitative methods in the included studies. Observation is valued more than opinions in HCD: it shows how users currently pursue their goals from an unbiased perspective. Interviewing techniques that can be employed are (1) the critical incident technique, and (2) the think-aloud method of a fictitious scenario [30]. This factual information of the context of use is crucial to HCD as the functional team needs it to create HCD outputs such as personas and use-case scenarios to communicate with the technical team to develop a suitable mHealth solution. All 3 actions, which are (1) the management team setting design goals and identifying stakeholders, (2) the functional team gaining insight into the users, and (3) the functional team creating design outputs, can and should be done iteratively to truly understand the context of use.

Second, mHealth solutions need to address the current pain points of the users and ensure they achieve their intended goals; a clear understanding of user needs and a concise list of user requirements help the HCD team accomplish that. The context of use plays a vital role in identifying the user problems from their current activities and what the users need to solve them. Then, user requirements based on these user needs are created to guide the HCD team on how the solutions should be designed. These requirements can be obtained through a literature review [8] or a direct contextual inquiry of recruited participants [30,35]. For example, Srinivas et al [35] successfully derived user requirements from the user needs identified through thematic analysis of the established context of use.

The included studies also pointed out that there was a set of requirements unique to older users; however, most had not listed these requirements at the beginning and dealt with them only after the users raised the problems in usability testing. These design considerations for older adults are well-established: Harte et al [51] reported an extensive list of HCD considerations for connected health devices for older adults, and Li et al [52] identified barriers to mHealth adoption by older adults in their narrative review. With these guidelines, mHealth solutions can be designed to suit older adults’ physical and cognitive limitations prior to testing for efficiency. Psychosocial factors such as motivation, technology perception, and social influence need to be addressed as well to ensure adoption.

Specifying user requirements also means setting measurable goals for the mHealth system. This usually requires gathering quantitative data for the context of use, such as the duration to complete the conventional I-IMR, which is approximately 8-10 months [8]; the HCD team might set the goals for their system to take only 4-5 months accordingly. If no goal is set, the way to assess the system in the subsequent steps will be limited.

Third, HCD seeks to create an ideal system through iterative prototyping together with the stakeholders, to make certain all user needs and requirements are accounted for. In the beginning, the functional team should design how the users will interact with the system and how the interfaces will be like. Harte et al [33] demonstrated this in their study by creating use cases as outputs to be analyzed by stakeholders for feedback. These outputs are called low-fidelity prototypes because they are easy to create, simple to change, and able to quickly convey the design concepts to all relevant parties. Besides, user interfaces can be based on existing design guidelines, such as the literature review about the unique user needs and requirements of older adults done by Fortuna et al [8]. Once the design is refined and approved, the technical team could then create an interactive software of the system or high-fidelity prototypes.

It should be noted that high-fidelity prototypes are not open for major changes or costly to do so; the best approach would be to finalize user–system interaction and user interfaces before their creation. As this HCD step requires iteration, an agile approach is recommended [35]. Agile is a software methodology based on rapid and iterative prototyping to gain continuous feedback from users, allowing developers to quickly create, evaluate, and improve their solutions to best fit the users [53]. Communication with stakeholders is key in this step, especially with the older population whose participation tends to be low due to their technology ineptness, physical and cognitive limitations, and logistical issues, for example, timing and travel [30]. Cornet et al [30] recruited patient advisors, older adult volunteers, to bridge these gaps: these patient advisors were able to give rapid feedback from users’ perspectives. Due to HCD being agile in nature, this step is often done together with the evaluation step.

Fourth, evaluation comes after the production of design solutions. The functional team should work closely with the technical team to evaluate the produced solutions. In HCD, evaluating the designs is done by usability experts and end users. A number of included studies recommended system evaluation by usability experts before end users. Expert evaluation can identify and classify usability problems early in the process where changes are less punishing; it is also much simpler to arrange compared with its user-based counterpart [32,33,35]. The process can be conducted according to standards such as usability heuristics [30,32,33,35] or by having the experts role-play as real end users [32,33]. However, it should be noted that the greater the difference in usability knowledge between the experts and the users, the more divergent the results from the 2 groups might be [26].

Evaluation by end users is critical to the process because getting feedback from the target users and improving accordingly would surely make the system usable for them. This result can be obtained by recruiting the right participants. Fortuna et al [8] reported that their sampling method with the intention to select only willing participants to test the system could have excluded the group of real end users who might be less eager to participate. Harte et al [32] dealt with this problem by verifying that the recruited participants fit well with HCD goals: the participants’ visual perception and cognitive processing were measured to confirm that the sampling method was inclusive enough. As for the number of participants, no standard has been agreed upon, but 5-10 participants are typically enough to discover major usability problems [45].

User-based evaluation methods range from giving the users specific tasks in a controlled environment to letting them use the system in the real world; the complexity also increased respectively so. Cornet et al [30] remarked in their study that usability testing in a controlled laboratory setting is more
older people could lead to design failures if left unchecked [56]. To avoid such pitfalls, we have to learn from the untold stories [30]. This review has gathered and summarized practical HCD challenges and strategies from primary research to aid HCD implementation with older adults.

Hastened by the COVID-19 pandemic, the field of mHealth will only expand. Moving forward, digital health solutions are aiming further than empowering patients and enhancing delivery. They are going for “digital therapeutics.” These evidence-based interventions aim to prevent and manage medical conditions through digital platforms and mobile devices; one of its focuses is to deliver lifestyle therapy to combat chronic diseases such as type 2 diabetes [57]. Older adults are major target users as most have chronic conditions and can benefit greatly from these digital lifestyle therapies. However, the field is in need of solutions for effective development, testing, and deployment [58]. Future research on implementing HCD in digital therapeutics might be able to solve these issues and improve the health of the older population as a whole.

**Limitations**

Limitations of this systematic review are acknowledged. First, the ACM Digital Library was not included in this review despite being in the relevant field. We did search the database on the same day as the others: no studies from the ACM Digital Library passed our criteria. We then failed to mention this once we proceeded with the review. However, we ran another search with the same strategy on the database in May 2021 to recheck; 14 studies found did not pass our abstract screening according to our established eligibility criteria.

Second, the research question aims to address the whole HCD process, but an existing body of literature proves to be limited as the topic is an emergent subfield, especially with older adults as the target group. Although the criteria are forgiving, the search strategy and the inclusion criteria still demand that all steps of HCD are implemented in each app development. This excludes a large number of studies that feature only a part of HCD. For example, one study might focus on qualitative interviews without applying them, while another might test a newly developed system that is based solely on the authors’ vision, not actual user needs. Nevertheless, the included studies complement one another and thus can accommodate the research question as illustrated in this review.

Third, the highly diverse HCD goals and methodologies in the included mHealth apps restrict the means of analysis and synthesis of results. All studies relied heavily on various qualitative means for HCD such as literature reviews, interviews, and field notes from direct observations. Even the seemingly same approaches, such as interviews, still differ in detail such as the time, the duration, the focus, and the questions. Most studies also focus more on the process not the result, or in the case of case studies, the process itself is the result. This might be due to the fact that baselines of the existing activities are not established in the second step of HCD, specifying the user requirements, so comparisons for effectivity and efficiency of the newly developed mHealth interventions cannot be made with objective metrics. Because of that fact, the included studies have to be reviewed with qualitative techniques using narrative
synthesis and guided by ISO 9241-210 together with logical ordering of the Minto pyramid principle [26,37,38]. Quantitative results of the included studies, which are based on the less tangible satisfaction results of standardized tools and often lack a definite conclusion, are also underutilized.

This leads to the fourth limitation regarding the included studies: all but 1 of the 5 mixed methods studies are rated to be of inadequate quality by the MMAT. Their quantitative components lack clarity. They do not explain their sampling methods or have done so insufficiently, resulting in the inability to deem their samples representative of the target population and failure to address possible confounding factors in making the conclusion that HCD helps make a usable product. This issue of the sampling methodology is also raised by the authors of the included studies; future HCD research should note this point in their strategic planning accordingly.

Finally, the authors stress that the aim of this systematic review was not to assess the implementation of HCD in creating mHealth for older adults or the effectiveness of mHealth interventions. The objective was to explore existing literature and establish recommendations and pitfalls for subsequent HCD projects. The older adults might be a narrow target population, but being the more sensitive and vulnerable group, the insight gained could be applicable to a wider range of users and help make future mHealth solutions more inclusive as well.

Conclusions

This review concludes that HCD can be used to create mHealth solutions for older adults and has summarized the process based on the 4 HCD steps with additional recommendations. The findings of this review can help designers, developers, and researchers gain an overview of HCD for older adults and implement the framework in their projects. The growing body of literature is encouraging, but more evidence-based results of HCD on creating mHealth for older adults are still needed. Future research should also focus on applying artificial intelligence and machine learning in HCD and utilizing the framework to create novel mHealth solutions for the population.

Acknowledgments

This research project was funded by the Ratchadapisek Sompoch Endowment Fund (2021) under Telehealth Cluster, Chulalongkorn University. The authors also thank the Center of Excellence in Hepatitis and Liver Cancer, Department of Biochemistry, Faculty of Medicine, Chulalongkorn University, and Thailand Research Fund (RTA6280004).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2009 checklist of this systematic review. [PDF File (Adobe PDF File), 96 KB - mhealth_v10i1e29512_app1.pdf ]

References


47. Lund AM. Measuring usability with the USE questionnaire. Usability Interface 2001;8(2):3-6 [FREE Full text]


Edited by G Eysenbach, L Buis; submitted 09.04.21; peer-reviewed by V Mylonopoulou, A Perkusich, V Stara; comments to author 27.04.21; revised version received 24.08.21; accepted 23.11.21; published 14.01.22.

Please cite as:

Nimmanterdwong Z, Boonviriya S, Tangkijvanich P
Human-Centered Design of Mobile Health Apps for Older Adults: Systematic Review and Narrative Synthesis
JMIR Mhealth Uhealth 2022;10(1):e29512
URL: https://mhealth.jmir.org/2022/1/e29512
doi:10.2196/29512
PMID:

©Zethapong Nimmanterdwong, Suchaya Boonviriya, Pisit Tangkijvanich. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 14.01.2022. 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.
Knowledge and Expectations of Hearing Aid Apps Among Smartphone Users and Hearing Professionals: Cross-sectional Survey

Jae Sang Han¹, MD; Yong-Ho Park²,³, MD, PhD; Jae-Jun Song⁴, MD, PhD; Il Joon Moon⁵, MD, PhD; Woojoo Lee⁶, PhD; Yoonjoong Kim⁷, MD; Young Sang Cho⁵, MD; Jae-Hyun Seo¹*, MD, PhD; Moo Kyun Park⁷,⁸*, MD, PhD

¹Department of Otolaryngology-Head and Neck Surgery, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea
²Department of Otolaryngology-Head and Neck Surgery, College of Medicine, Chungnam National University, Daejeon, Republic of Korea
³Brain Research Institute, College of Medicine, Chungnam National University, Daejeon, Republic of Korea
⁴Department of Otorhinolaryngology-Head and Neck Surgery, Korea University College of Medicine, Seoul, Republic of Korea
⁵Department of Otorhinolaryngology-Head and Neck Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
⁶Department of Public Health Sciences, Graduate School of Public Health, Seoul National University, Seoul, Republic of Korea
⁷Department of Otorhinolaryngology, Head & Neck Surgery, Seoul National University Hospital, Seoul National University College of Medicine, Seoul, Republic of Korea
⁸Sensory Organ Research Institute, Seoul National University, Medical Research Center, Seoul, Republic of Korea
*these authors contributed equally

Corresponding Author:
Jae-Hyun Seo, MD, PhD
Department of Otolaryngology-Head and Neck Surgery
College of Medicine
The Catholic University of Korea
222, Banpo-daero, Seocho-gu
Seoul, 06591
Republic of Korea
Phone: 82 2 2258 6210
Fax: 82 2 595 1354
Email: revalseo@catholic.ac.kr

Abstract

Background: Despite the increasing prevalence of hearing loss, the cost and psychological barriers to the use of hearing aids may prevent their use in individuals with hearing loss. Patients with hearing loss can benefit from smartphone-based hearing aid apps (SHAAs), which are smartphone apps that use a mobile device as a sound amplifier.

Objective: The aim of this study is to determine how ear, nose, and throat outpatients perceive SHAAs, analyze the factors that affect their perceptions, and estimate the costs of an annual subscription to an app through a self-administered questionnaire survey of smartphone users and hearing specialists.

Methods: This study used a cross-sectional, multicenter survey of both ear, nose, and throat outpatients and hearing specialists. The questionnaire was designed to collect personal information about the respondents and their responses to 18 questions concerning SHAAs in five domains: knowledge, needs, cost, expectations, and information. Perception questions were rated on a scale of 1 (strongly disagree) to 5 (strongly agree). Questions about the expected cost of SHAAs were included in the questionnaire distributed to hearing experts.

Results: Among the 219 smartphone users and 42 hearing specialists, only 8 (3.7%) respondents recognized SHAAs, whereas 18% (47/261) of respondents reported considering the use of an assistive device to improve their hearing capacity. The average perception score was 2.81 (SD 1.22). Among the factors that shaped perceptions of SHAAs, the needs category received the lowest scores (2.02, SD 1.42), whereas the cost category received the highest scores (3.29, SD 1.14). Age was correlated with the information domain (P<.001), and an increased level of hearing impairment resulted in significantly higher points in the needs category (P<.001). Patients expected the cost of an annual app subscription to an SHAA to be approximately US $86, and the
predicted cost was associated with economic status ($P=.02$) and was higher than the prices expected by hearing specialists ($P<.001$).

**Conclusions:** Outpatients expected SHAAs to cost more than hearing specialists. However, the perception of the SHAA was relatively low. In this regard, enhanced awareness is required to popularize SHAAs.

**(JMIR Mhealth Uhealth 2022;10(1):e27809)** doi:10.2196/27809

**KEYWORDS**

smartphone; hearing aids; app; perception; survey; hearing loss; mobile phone

**Introduction**

Hearing loss is one of the most common health care problems worldwide. When the World Health Organization started reporting hearing loss in 1985, the number of people with moderate-to-profound hearing impairment was estimated to be 42 million. Furthermore, the number of people with disabling hearing loss reached 466 million in 2018 and is projected to reach approximately 630 million by 2030 [1]. Hearing aids (HAs) are standard hearing intervention methods [2], and the adequate use of HAs improves hearing-specific and general health–related quality of life in adults with mild to moderate hearing loss [3].

Nevertheless, HA adoption rates are extremely low. Globally, only 17% of those who need appropriate hearing rehabilitation use HAs [4]. In addition, a large South Korean cohort study reported that among participants who had minimal hearing loss (mild bilateral hearing loss, unilateral hearing loss, and high frequency hearing loss), only 0.47% of those with subjective symptoms used HAs [5]. Failure to achieve early rehabilitation can accelerate the development of hearing loss and, ultimately, incur enormous social costs [6]. The price of HAs is an important barrier to use [7,8]. When hearing health care is subsidized by the government, HA penetration rates slightly increase [9]. Therefore, other factors, such as social stigma, denial of hearing loss, reduced self-efficacy, and limited access to hearing services should be considered [8]. A prolonged time from the onset of hearing loss to HA intervention has negative effects on quality of life [10]. To address these barriers, alternatives such as over-the-counter (OTC) HAs, personal sound amplification products (PSAPs), and smartphone-based HA apps (SHAAs) have been previously evaluated [11-14]. Moreover, the US Food and Drug Administration announced in its 2016 nonbinding guidance document that medical assessment is no longer required for OTC HAs for individuals aged ≥18 years [15].

SHAAs were originally developed to mimic conventional HA devices. SHAAs refer only to the software installed on a mobile device for hearing support, which is different from the traditional HA hardware–software complex. SHAAs require wired or Bluetooth headsets or headphones instead of hardware resources. Many free or low-price HA apps are available on the web. Although they enhance hearing capabilities through sound amplification, SHAAs were previously far less sophisticated because they could not exactly fit an individual’s prescribed target gain as could HAs fitted using real-ear measurement [16]. Some SHAAs have separate channels and advanced functions, such as noise reduction and acoustic feedback suppression [17].

Until now, it was not clear whether SHAAs were clinically effective and could be an alternative device to traditional HAs [18]. In addition, their level of patient satisfaction is generally lower than that with conventional HAs [19]. However, the performance of SHAAs is likely to improve with the development of smartphone hardware and apps, and SHAAs have great potential to contribute to hearing rehabilitation [20].

Easy accessibility is a notable advantage of SHAAs. Users can simply download the app on their smartphones and prepare headsets or headphones for use. SHAAs may particularly help overcome psychological resistance to the use of HAs. Trials with SHAAs showed a reduction in the degree of anxiety and personal distress and increased self-esteem. In addition, reduced stigma or body image of HA users can be expected because of the growing number of individuals who wear headphones with their smartphones [21]. Maidment et al [22] demonstrated that the use of smartphone-connected listening devices in adults with hearing loss could address issues surrounding stigma because smartphones are ubiquitous in everyday life. In addition, the price of SHAAs is lower than that of conventional HAs or PSAPs [23]. Dozens of SHAAs have been released in the App Store (iPhone operating system) and Google Play (Android). Moreover, a new SHAA called Sound Amplifier was introduced by Google [24,25].

The mobile app market is rapidly growing. As of 2019, about 61% of the global population was able to access the internet from mobile devices, and this number is projected to increase to approximately 79% by 2025 [26]. Furthermore, as an increasing number of older adults (>65 years) are using mobile internet via their smartphones, smartphones are expected to exert a greater influence on hearing health care, and SHAAs will expand accordingly [20,27]. Nevertheless, no previous studies have focused on how SHAAs are perceived and the factors affecting the perception of SHAAs. Thus, in this study, we assessed the current awareness of SHAAs and analyzed the associated factors through questionnaires. This information will serve as a baseline for further research on hearing rehabilitation using SHAAs.

**Methods**

**Participants**

We performed a multicenter survey of 5 general hospital outpatients who use smartphones and hearing specialists, including otology specialists, audiologists, and HA researchers. Before gaining access to a questionnaire, the potential participants were informed about the survey, and those who

https://mhealth.jmir.org/2022/11/e27809
agreed to participate were asked to fill out the questionnaire under the direction of a health care provider.

This study was carried out in accordance with the Declaration of Helsinki on biomedical research for human participants, and the study protocol was approved by the institutional review board of each participating hospital (Seoul St Mary’s Hospital, KC20QDI0526; Chungnam National University Hospital, 2020-06-092; Korea University Hospital, 2020GR0020; Samsung Medical Center, 2020-05-056; and Seoul National University Hospital, D-2003-028-1109).

**Questionnaire**

A survey on the perception of HAs published by Park et al [28] was modified for use in this study because there is no standardized questionnaire available to assess perceptions of HAs, including SHAAs. Park et al developed a questionnaire that contained 19 questions with an appropriate level of reliability and validity (Cronbach $\alpha=.76$). To evaluate the consistency of questionnaire items, Hotelling T-square test was used. The items had significant reliability, with $F=28.5$, $P<.001$ [28]. One question (“I know that different types of HAs can be worn depending on the degree of hearing loss”) was excluded from the questionnaire by Park et al because it was not suitable for the SHAQ questionnaires. In addition, hearing aids was replaced with smartphone-based hearing aid apps. A total of 18 questions in the questionnaire were reviewed by 42 hearing rehabilitation specialists who participated in the opinion survey, and Cronbach $\alpha$ for each question was recalculated (Multimedia Appendix 1). The language used in the questionnaire was Korean. To prevent any possible confusion, respondents were fully informed that SHAAs are independent substitutes for HAs and do not require an additional device other than a smartphone and headphone or headset.

The questionnaire was divided into three sections: (1) sociodemographic characteristics, including age, gender, residence, educational background, economic status, and occupation; (2) clinical characteristics, including the recognition of hearing loss and inconvenience level, the presence of tinnitus and inconvenience level, previous experience with PSAPs or SHAAs by the respondent or their family member, respondent’s willingness to use PSAPs or SHAAs, and expected cost of the app; and (3) perception status. In the clinical characteristics section, respondents with hearing loss or tinnitus were asked to assess the degree of their symptoms using a visual analogue scale (VAS). In the perception status section, they were asked to rate 18 questions in 5 categories on a scale from 1 (strongly disagree) to 5 (strongly agree); lower scores indicate poorer awareness. The questions were grouped into 5 categories by similar objectives, which were reviewed by the hearing specialists, allowing the analysis to be simpler and clearer. Questions 1-4 were grouped in the knowledge category, which aimed to evaluate whether respondents were aware of SHAAs as hearing rehabilitation options and how they differed from conventional HAs. Questions 5-6 in the needs category were designed to evaluate whether respondents thought that SHAAs were necessary for hearing discomfort. Questions 7-9 in the cost category were used to identify the influence of price on the decision to purchase, and questions 10-13 were used to evaluate respondents’ expectations regarding the ability of SHAAs to improve hearing capabilities. Finally, questions 14-18 in the information category attempted to determine whether participants had accurate information about how to use SHAAs (Multimedia Appendix 1).

The opinion survey for hearing specialists contained questions to determine demographic information such as employment history, educational background, and professional experience (length of career) as well as the expected annual subscription rate for an SHAA and the main selection criteria for HA devices (Multimedia Appendix 2).

Korean Won was used as the standard currency in the questionnaire and was converted into US $ in this report (US $1=1082.50).

**Statistical Analysis**

Age, gender, education background, and economic status data were treated as categorical variables. Reference variables were 20-39 years for age, male for gender, middle school graduate for educational background, and 1 for economic status. VAS score of hearing loss and VAS score of tinnitus were regarded as continuous variables. Linear regression models with the perception level and the expected annual subscription rate as response variables were applied. Robust variance estimation was used for SEs and CIs. Age, gender, educational background, economic status, VAS score of hearing loss, and VAS score of tinnitus were used as explanatory variables in the regression models. Bonferroni-corrected $P$ values $<.05$ were considered statistically significant. The 2-sample 2-tailed $t$ test was used to compare the expected costs between the hearing specialists and potential users. All statistical analyses were performed using R version 3.6.0 (R Foundation for Statistical Computing).

**Results**

**Clinical Characteristics of Enrolled Participants**

A total of 98.6% (219/222) of respondents’ answers were analyzed after the survey responses of 3 participants with a survey completion rate $<50\%$ were excluded.

The clinical characteristics of the participants are presented in Table 1. The mean age of participants was 52.02 (SD 15.44) years, and male respondents slightly outnumbered female respondents at 59.8% (131/219 male respondents) to 40.2% (88/219 female respondents). Most respondents (138/219, 63%) were college graduates or higher education, and about half (114/219, 52.1%) of the participants estimated themselves as having an intermediate economic status. A total of 44.3% (97/219) of respondents answered that they had subjective hearing loss. The average VAS score of respondents with hearing loss was 2.51 (SD 3.29). In addition, 40.2% (88/219) of respondents had tinnitus, and their average VAS score was 2.39 (SD 3.32). Owing to the multicenter nature of the study, the locations of the participants’ residences varied widely. Most participants lived in urban areas (130/219, 59.3%), followed by suburban areas (59/219, 26.9%) and rural areas (30/219, 13.7%). Only 0.9% (2/219) of the enrolled participants had been using HAs at the time of the survey, so wearing HAs was not used in the analysis.
When asked about SHAAs, 21.5% (47/219) of respondents stated that they had considered using an assistive device for hearing loss, but only 3.7% (8/219) respondents knew the difference between traditional HAs and SHAAs. Only 0.9% (2/219) of respondents had experience with an SHAA. However, 26.5% (58/219) of respondents expressed a willingness to use an SHAA in the future.

A total of 42 responses were received from the hearing specialist group, which comprised 29 (69%) otologists and 13 (31%) audiologists. The average number of years working in this profession was 11.83 (SD 7.81) years. A total of 40% (17/42) of respondents had a bachelor’s degree, followed by 33% (14/42) of respondents with a master’s degree and 26% (11/42) of respondents with a doctorate degree.

Table 1. Participant characteristics (N=219).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>52.0 (15.4)</td>
</tr>
<tr>
<td>Sex, n (%)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>131 (59.8)</td>
</tr>
<tr>
<td>Female</td>
<td>88 (40.2)</td>
</tr>
<tr>
<td>Education level, n (%)</td>
<td></td>
</tr>
<tr>
<td>Junior high graduate or less</td>
<td>29 (13.2)</td>
</tr>
<tr>
<td>High school graduate</td>
<td>53 (24.2)</td>
</tr>
<tr>
<td>College graduate or higher</td>
<td>137 (62.6)</td>
</tr>
<tr>
<td>Economic status, n (%)</td>
<td></td>
</tr>
<tr>
<td>A (very low)</td>
<td>12 (5.5)</td>
</tr>
<tr>
<td>B (low)</td>
<td>29 (13.2)</td>
</tr>
<tr>
<td>C (middle)</td>
<td>113 (51.6)</td>
</tr>
<tr>
<td>D (high)</td>
<td>52 (23.7)</td>
</tr>
<tr>
<td>E (very high)</td>
<td>13 (5.9)</td>
</tr>
<tr>
<td>Subjective hearing loss, n (%)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>97 (44.3)</td>
</tr>
<tr>
<td>No</td>
<td>122 (55.7)</td>
</tr>
<tr>
<td>If hearing loss “yes,” VAS(^a) score (1-10)(^b), mean (SD)</td>
<td>2.5 (3.3)</td>
</tr>
<tr>
<td>Tinnitus, n (%)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>88 (40.2)</td>
</tr>
<tr>
<td>No</td>
<td>131 (59.8)</td>
</tr>
<tr>
<td>If tinnitus “yes,” VAS score (1-10)(^c), mean (SD)</td>
<td>2.4 (3.3)</td>
</tr>
</tbody>
</table>

\(^a\)VAS: visual analogue scale.
\(^b\)Visual analogue scale (VAS) 1=very minimal problem; VAS 10=very serious problem. VAS 0 was considered to indicate no subjective hearing loss.
\(^c\)Visual analogue scale (VAS) 1=very minimal problem; VAS 10=very serious problem. VAS 0 was considered no subjective tinnitus.

Overall Awareness of SHAAs

The overall score of awareness of SHAAs of the 219 respondents was 2.81 (SD 1.21). Among the 5 categories, the needs category received the lowest score of 2.02 (SD 1.42), whereas the cost category ranked first with a score of 3.29 (SD 1.14; Figure 1).

In the opinion survey of hearing specialists, the main consideration factor for recommending an SHAA was basic performance (30/42, 71%), followed by price (8/42, 19%) and additional functions (2/42, 5%). In addition, noise reduction and the number of channels were mentioned by 1 respondent each.
Factors Affecting Awareness Scores

Age, Gender, and Area of Residence

Respondents were divided into three age groups: 20-39 years, 40-59 years, and ≥60 years. Compared with the reference age group of 20-39 years, there was no association between age and perception scores in the 40-59-year group, although there was a marginally significant positive correlation between age and information score ($P=.05$) in this age group. Meanwhile, in the ≥60-year group, there was a remarkable positive correlation between the total SHAA perception score and age in comparison with the 20-39-year group ($P=.002$), and there were also strong associations between the information and perception scores among the 5 categories ($P<.001$; Table 2).

To analyze whether gender affected SHAA perception, male respondents were used as the reference group. There were no significant correlations between gender and SHAA perception (Table 3). The area of residence was also not significantly correlated with SHAA perception.
Table 2. Relationship between age and perception scores on smartphone-based hearing aid apps. The reference age group was the 20-39–year group.

<table>
<thead>
<tr>
<th>Response</th>
<th>Coefficient (SE; 95% CI)</th>
<th>P value</th>
<th>Adjusted P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>40-59–year group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge</td>
<td>0.089 (0.095; −0.097 to 0.276)</td>
<td>.35</td>
<td>N/A</td>
</tr>
<tr>
<td>Needs</td>
<td>−0.155 (0.131; −0.411 to 0.102)</td>
<td>.24</td>
<td>.99</td>
</tr>
<tr>
<td>Cost</td>
<td>0.209 (0.155; −0.094 to 0.511)</td>
<td>.18</td>
<td>.89</td>
</tr>
<tr>
<td>Expectation</td>
<td>0.007 (0.168; −0.323 to 0.336)</td>
<td>.97</td>
<td>.99</td>
</tr>
<tr>
<td>Information</td>
<td>0.301 (0.118; 0.069 to 0.533)</td>
<td>.01</td>
<td>.05</td>
</tr>
<tr>
<td>≥60–year group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge</td>
<td>0.314 (0.102; 0.113 to 0.514)</td>
<td>.002</td>
<td>N/A</td>
</tr>
<tr>
<td>Needs</td>
<td>0.010 (0.151; −0.286 to 0.307)</td>
<td>.95</td>
<td>.99</td>
</tr>
<tr>
<td>Cost</td>
<td>0.386 (0.213; −0.031 to 0.803)</td>
<td>.07</td>
<td>.35</td>
</tr>
<tr>
<td>Expectation</td>
<td>0.286 (0.214; −0.134 to 0.705)</td>
<td>.18</td>
<td>.91</td>
</tr>
<tr>
<td>Information</td>
<td>0.563 (0.130; 0.308 to 0.819)</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

aN/A: not applicable.
bP<.05.
cP<.001.

Table 3. Relationship between gender and perception scores regarding smartphone-based hearing aid apps. The reference group was male respondents.

<table>
<thead>
<tr>
<th>Response</th>
<th>Coefficient (SE; 95% CI)</th>
<th>P value</th>
<th>Adjusted P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge</td>
<td>−0.053 (0.102; −0.253 to 0.147)</td>
<td>.61</td>
<td>.99</td>
</tr>
<tr>
<td>Needs</td>
<td>0.160 (0.170; −0.173 to 0.494)</td>
<td>.35</td>
<td>.99</td>
</tr>
<tr>
<td>Cost</td>
<td>0.129 (0.137; −0.140 to 0.397)</td>
<td>.35</td>
<td>.99</td>
</tr>
<tr>
<td>Expectation</td>
<td>−0.189 (0.120; −0.424 to 0.046)</td>
<td>.12</td>
<td>.58</td>
</tr>
<tr>
<td>Information</td>
<td>−0.027 (0.103; −0.229 to 0.175)</td>
<td>.79</td>
<td>.99</td>
</tr>
<tr>
<td>Total</td>
<td>−0.031 (0.079; −0.186 to 0.124)</td>
<td>.70</td>
<td>N/A</td>
</tr>
</tbody>
</table>

aN/A: not applicable.

**Hearing Loss and Tinnitus**

We next evaluated whether subjective hearing loss or tinnitus influenced the perception of SHAA. There were significant correlations between hearing loss and the total perception score (P=.001). The presence of hearing loss was strongly associated with the needs category (P<.001), but there were no significant associations with the other categories. The degree of hearing loss indicated by the VAS score was closely related to the total scores (P=.001) and needs (P<.001; Table 4).

Although the presence of tinnitus did not show a significant association with total scores, it was positively correlated with the needs category (P=.003). The VAS score for tinnitus did have significant associations with SHAA perception (Table 5).
### Table 4. Relationship between subjective hearing loss and perception scores regarding smartphone-based hearing aid apps.

<table>
<thead>
<tr>
<th>Response</th>
<th>Coefficient (SE; 95% CI)</th>
<th>P value</th>
<th>Adjusted P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subjective hearing loss (yes or no)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge</td>
<td>0.049 (0.014; 0.021 to 0.076)</td>
<td>.001&lt;sup&gt;a&lt;/sup&gt;</td>
<td>N/A&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Needs</td>
<td>0.021 (0.023; −0.024 to 0.066)</td>
<td>.36</td>
<td>.99</td>
</tr>
<tr>
<td>Cost</td>
<td>0.266 (0.030; 0.208 to 0.324)</td>
<td>&lt;.001&lt;sup&gt;c&lt;/sup&gt;</td>
<td>&lt;.001&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Expectation</td>
<td>0.030 (0.027; −0.024 to 0.083)</td>
<td>.28</td>
<td>.99</td>
</tr>
<tr>
<td>Information</td>
<td>0.017 (0.015; −0.012 to 0.046)</td>
<td>.24</td>
<td>.99</td>
</tr>
<tr>
<td><strong>Visual analogue scale score of hearing loss</strong></td>
<td>0.079 (0.024; 0.031 to 0.126)</td>
<td>.001&lt;sup&gt;a&lt;/sup&gt;</td>
<td>N/A</td>
</tr>
<tr>
<td>Knowledge</td>
<td>0.045 (0.037; −0.028 to 0.117)</td>
<td>.23</td>
<td>.99</td>
</tr>
<tr>
<td>Needs</td>
<td>0.304 (0.066; 0.175 to 0.434)</td>
<td>&lt;.001&lt;sup&gt;c&lt;/sup&gt;</td>
<td>&lt;.001&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Cost</td>
<td>0.024 (0.040; −0.054 to 0.102)</td>
<td>.55</td>
<td>.99</td>
</tr>
<tr>
<td>Expectation</td>
<td>0.059 (0.034; −0.008 to 0.127)</td>
<td>.08</td>
<td>.42</td>
</tr>
<tr>
<td>Information</td>
<td>0.057 (0.026; 0.006 to 0.107)</td>
<td>.03&lt;sup&gt;d&lt;/sup&gt;</td>
<td>.14</td>
</tr>
</tbody>
</table>

<sup>a</sup>P < .01.
<sup>b</sup>N/A: not applicable.
<sup>c</sup>P < .001.
<sup>d</sup>P < .05.

### Table 5. Relationship between subjective tinnitus and perception scores regarding smartphone-based hearing aid apps.

<table>
<thead>
<tr>
<th>Response</th>
<th>Coefficient (SE; 95% CI)</th>
<th>P value</th>
<th>Adjusted P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subjective tinnitus (yes or no)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge</td>
<td>0.003 (0.012; −0.022 to 0.027)</td>
<td>.82</td>
<td>N/A&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Needs</td>
<td>0.049 (0.020; 0.011 to 0.088)</td>
<td>.01&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.06</td>
</tr>
<tr>
<td>Cost</td>
<td>−0.094 (0.027; −0.147 to −0.041)</td>
<td>.001&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.003&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Expectation</td>
<td>−0.001 (0.024; −0.047 to 0.046)</td>
<td>.98</td>
<td>.99</td>
</tr>
<tr>
<td>Information</td>
<td>0.022 (0.020; −0.018 to 0.061)</td>
<td>.28</td>
<td>.99</td>
</tr>
<tr>
<td><strong>Visual analogue scale score of tinnitus</strong></td>
<td>0.021 (0.022; −0.022 to 0.064)</td>
<td>.34</td>
<td>N/A</td>
</tr>
<tr>
<td>Knowledge</td>
<td>0.068 (0.031; 0.007 to 0.128)</td>
<td>.03&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.14</td>
</tr>
<tr>
<td>Needs</td>
<td>−0.089 (0.056; −0.200 to 0.021)</td>
<td>.11</td>
<td>.57</td>
</tr>
<tr>
<td>Cost</td>
<td>0.049 (0.044; −0.037 to 0.135)</td>
<td>.26</td>
<td>.99</td>
</tr>
<tr>
<td>Expectation</td>
<td>0.052 (0.035; −0.018 to 0.121)</td>
<td>.15</td>
<td>.72</td>
</tr>
<tr>
<td>Information</td>
<td>−0.003 (0.026; −0.054 to 0.049)</td>
<td>.92</td>
<td>.99</td>
</tr>
</tbody>
</table>

<sup>a</sup>N/A: not applicable.
<sup>b</sup>P < .05.
<sup>c</sup>P < .01.

### Expected Price of SHAAs

The average expected cost for an annual subscription to an SHAA was US $84.43 (95% CI 75.66-93.21). Analyzed by age group, the average expected prices were US $97.37 (95% CI 75.10-119.54), US $78.98 (95% CI 63.46-94.41), and US $86.47 (95% CI 62.91-109.93) in the 20-39-year group, the 40-59-year group, and in the ≥60-year group, respectively (Figure 2). The expected cost was significantly correlated with economic status (P=.02), whereas it was not significantly associated with other categories (Table 6).

The experts’ average expected cost for an annual subscription to a premium version app was US $32.48 (95% CI 17.81-47.24), and 33% (14/42) of respondents answered that the app should be available at no cost. As for an entry-version app, the expected cost for an annual subscription was US $9.69 (95% CI...
2.68–16.70) on average, and 71% (30/42) of respondents expected this app to be provided free of charge.

The average cost for an annual subscription expected by potential users was markedly higher than that expected by hearing specialists (based on the premium version; \( P < .001 \)). A total of 45% (19/42) of respondents among the potential users were not willing to pay for the app, which was much lower than the percentage of the hearing specialist group who thought the SHAA should be provided for free (\( P < .001 \)).

Table 6. Factors affecting the expected price of smartphone-based hearing aid apps.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient (SE; 95% CI)</th>
<th>( P ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40-59</td>
<td>−1.991 (1.348; −4.633 to 0.651)</td>
<td>.14</td>
</tr>
<tr>
<td>≥60</td>
<td>−1.180 (1.509; −4.137 to 1.778)</td>
<td>.43</td>
</tr>
<tr>
<td>Sex (female)</td>
<td>0.971 (0.983; −0.955 to 2.898)</td>
<td>.32</td>
</tr>
<tr>
<td><strong>Education level</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school graduate</td>
<td>−1.156 (1.875; −4.831 to 2.518)</td>
<td>.54</td>
</tr>
<tr>
<td>University graduate</td>
<td>0.122 (1.997; −3.793 to 4.037)</td>
<td>.95</td>
</tr>
<tr>
<td>Economic status</td>
<td>−1.474 (0.638; −2.723 to −0.224)</td>
<td>.02</td>
</tr>
<tr>
<td>Hearing loss visual analogue scale score</td>
<td>0.261 (0.209; −0.148 to 0.671)</td>
<td>.21</td>
</tr>
<tr>
<td>Tinnitus visual analogue scale score</td>
<td>−0.207 (0.179; −0.559 to 0.145)</td>
<td>.25</td>
</tr>
</tbody>
</table>

\( ^a P < .05. \)
Discussion

Principal Findings

This cross-sectional study recruited 261 participants, consisting of 219 outpatients and 42 hearing specialists from multiple locations in South Korea, to avoid regional bias. In addition, as non–smartphone users are not potential candidates for SHAAs use, this study targeted people who own and use smartphones. Overall, only a limited number of participants had heard about SHAAs, and only 4% (9/219) of respondents were aware of the differences between SHAAs and conventional HAs. In addition, only 0.9% (2/219) of the respondents had experience using an SHAAs. These results indicated an extremely low level of perception regarding SHAAs. However, it is noteworthy that 26% (57/219) of the respondents stated their intention to consider using SHAAs after they obtained information about SHAAs during the survey. This suggests that increased awareness of SHAAs may lead to their use by more individuals with hearing loss.

As the perception of SHAAs was more meaningful to people with hearing problems than to the general population, the survey was conducted for ear, nose, and throat outpatients and resulted in a relatively high proportion of participants with hearing loss or tinnitus. There was also a strong association between the perception of SHAAs and the age and degree of hearing loss. The amount of information increased with age, whereas gender showed no relationship with the information. Although the degree of hearing loss influenced the purchase of an SHAAs, the level of tinnitus was not related to the perception of SHAAs. These findings suggest that the demand for hearing rehabilitation devices increases with age and the development of hearing loss, indicating the necessity of providing further relevant information to elderly individuals with hearing impairment.

The expected cost was associated with the economic status. The prices that respondents were willing to pay for SHAAs were relatively high in the 20-39–year group and the >60-year group. This is perhaps because the younger generation group would like to improve their own or their parents’ hearing capacity, and the older adult group faces more inconvenience from hearing loss. In addition, price was regarded as one of the most crucial factors determining whether or not to purchase an SHAAs. Respondents anticipated more advanced features with an increase in price. The average expected cost for an annual subscription to an SHAAs was higher than that expected by the hearing specialists. We assumed that those with hearing impairment were willing to pay a higher price than expected by specialists because of the effect of hearing loss on their quality of life. Furthermore, the expected cost was higher than the actual price of Petrelax (once-off annual cost: US $59.99 for iPhone operating system), one of the most expensive SHAAs on the market [20,29]. It is notable that only 5.9% (13/219) of respondents expected the app to be free. This suggests that potential users are willing to pay a certain amount for an SHAAs with the expectation of efficacy. Nevertheless, the expected cost is substantially lower than that of commercially available HAs or PSAPs. The cost of HA fitting for a single device was US $2336 in the United States [30,31]. Moreover, OTC HAs range in price from approximately US $600 to US $1000 [32], and lower-priced PSAPs range from US $250 to US $350 [33]. The life expectancy of HAs or PSAPs is approximately 5 years. A 5-year subscription to an SHAAs would be approximately US $430, which is much lower than the price of HAs and similar to that of premium PSAPs. Thus, SHAAs are likely to compete with PSAPs for market share in the future.

Smartphone-based mobile health is widely used for diagnostics and therapy [34] and also supports hearing rehabilitation. Pagliaonga et al [25] investigated 200 hearing health care apps available on the market. Among these apps, the largest proportion (28%) comprised sound enhancement apps [25].

SHAAs have several advantages. First, SHAAs range in price from free to US $70, and are therefore cheaper than conventional HAs overall. SHAAs are therefore likely to substitute for traditional HAs [9]. Second, patients with hearing loss can receive a call and perform HA fitting directly with their smartphones [35]. Third, because of the convenience offered by smartphones in our daily lives, SHAAs may allow patients with hearing loss to feel free from the stigma of using HAs [22]. Finally, the advantages mentioned enable SHAAs to act as gateway products to more sophisticated devices, such as conventional HAs [36].

South Korea’s gross domestic product per capita is US $32,310, ranking South Korea 28th across the globe [37]. In particular, South Korea has one of the highest smartphone penetration rates, with the smallest gap among all ages (percentage of adults who own a smartphone in South Korea in 2018: 18-34–year group, 99% and >50-year group, 91%) [38]. Given that smartphone use is skyrocketing worldwide, awareness of SHAAs can increase global accessibility to HA interventions.

However, it should be noted that the effectiveness of SHAAs has not been fully proven. Amlani et al [23] recommended that SHAAs be used only as a temporary means of assistance by patients using HA. Medwetsky et al [39] reported that SHAAs improved listening performance, but test participants had only mild to moderate high frequency hearing loss. As the effectiveness of SHAAs in patients with moderate to severe hearing loss is yet to be determined, it is essential to carry out a series of well-designed studies to determine the efficacy of SHAAs in hearing rehabilitation.

Comparison With Previous Work

Previous studies have demonstrated that SHAAs can improve hearing performance in patients with and without hearing loss [20,40]. Most previous studies compared auditory performance with conventional HAs in patients with hearing loss. They evaluated the self-reported benefits and satisfaction in a small case series in a single center [20,21]. Performance and satisfaction show wide variations according to app, operation system, and type of headphones [18]. However, previous studies did not comprehensively evaluate the awareness and associated factors of SHAAs in a cross-sectional multicenter survey. Most of the participants in our study were non-HA users. As less than half of them had subjective hearing loss, we think they could be potential candidates for SHAAs. In addition, this study showed the expectation and expected cost of HAs in smartphone

https://mhealth.jmir.org/2022/1/e27809

JmIR Mhealth Uhealth 2022 | vol. 10 | iss. 1 | e27809 | p.32 (page number not for citation purposes)
users and hearing professionals. Our data showed that the price of SHAAs is underestimated and suggested an expected cost, which is useful information for mobile app users and developers.

**Strengths and Limitations**

One strength of this study is that it is the first study to measure the perception of SHAAs. These findings are expected to pave the way for more surveys regarding awareness of other hearing rehabilitation devices. In addition, because of the multicenter nature of this study, our findings are generalizable to a broad population. According to a survey conducted by the government, the urban population of South Korea is approximately 90% of the total population, which is similar to the population distribution in our results (189/219, 86.3%) [41].

One limitation of this study is that because no standardized questionnaires are available to evaluate perceptions of hearing rehabilitation devices, we modified the questionnaire of a preceding study that investigated awareness about HAs [28]. As this questionnaire was not originally designed or validated to measure perceptions of SHAAs, our findings should be interpreted with caution. A well-validated survey on the perception of hearing assistant devices such as SHAAs should be developed in the future.

In addition, the participants in this study were younger than those of known typical HA seekers [10]. In addition, we did not investigate the experience of HAs because of the small number of HA users in this study. HA users may not actively seek alternative devices, and these HA users could have altered the results of the survey.

Furthermore, our findings do not provide insight into the efficacy of SHAAs in remediating hearing loss; our focus was primarily on the perception of SHAAs. Thus, clinical validation of the effectiveness of the SHAA is required [21].

**Conclusions**

SHAAs are an alternative hearing rehabilitation option for smartphone users with hearing loss who have no access to appropriate hearing rehabilitation devices because of their high costs. However, the perception of SHAAs was very low. Age and degree of hearing loss were correlated with perception scores. Potential users estimated the cost of an SHAA as approximately US $86 for a 1-year subscription. Those with hearing loss and requiring hearing rehabilitation were willing to pay a higher price than what the hearing specialists expected the price to be. In addition, a higher economic status was associated with an increased willingness to pay higher prices. Considering that a large portion of respondents showed interest in SHAAs after obtaining information from the survey, enhancement of perception of SHAAs is likely crucial to expand their market base.

**Acknowledgments**

This research was supported by a grant from the Korea Health Technology Research and Development Project through the Korea Health Industry Development Institute, funded by the Ministry of Health and Welfare, Republic of Korea (grant HC19C0128).

**Conflicts of Interest**

None declared.

**Multimedia Appendix 1**
Survey on the perception of hearing aid apps.
[PDF File (Adobe PDF File), 322 KB - mhealth_v10i1e27809_app1.pdf ]

**Multimedia Appendix 2**
Hearing specialist’s opinion for hearing aid apps.
[PDF File (Adobe PDF File), 207 KB - mhealth_v10i1e27809_app2.pdf ]

**References**


26. The mobile economy 2019. GSM Association. 2019. URL: [https://www.gsmaintelligence.com/research/?file=9a9a6e6202ee1df5f787cfebb95d3639c5&download](https://www.gsmaintelligence.com/research/?file=9a9a6e6202ee1df5f787cfebb95d3639c5&download) [accessed 2021-12-14]


Abbreviations

HA: hearing aid
OTC: over-the-counter
PSAP: personal sound amplification product
SHAA: smartphone-based hearing aid app
VAS: visual analogue scale

©Jae Sang Han, Yong-Ho Park, Jae-Jun Song, Il Joon Moon, Woojoo Lee, Yoonjoong Kim, Young Sang Cho, Jae-Hyun Seo, Moo Kyun Park. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 07.01.2022. 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 Elderly User Preference for Telehealth Solutions in China: Exploratory Quantitative Study

Nuoya Chen1, PhD; Pengqi Liu2, MSc

1Faculty of Global Studies, Justice and Rights, University of Macerata, Macerata, Italy
2Sino-Danish College, University of Chinese Academy of Sciences, Beijing, China

Corresponding Author:
Nuoya Chen, PhD
Faculty of Global Studies, Justice and Rights
University of Macerata
Crescimbeni 30-32
Macerata, 62100
Italy
Phone: 39 0733 2582418
Email: cn0824@gmail.com

Abstract

Background: In the next 15 to 20 years, the Chinese population will reach a plateau and start to decline. With the changing family structure and rushed urbanization policies, there will be greater demand for high-quality medical resources at urban centers and home-based elderly care driven by telehealth solutions. This paper describes an exploratory study regarding elderly users’ preference for telehealth solutions in the next 5 to 10 years in 4 cities, Shenzhen, Hangzhou, Wuhan, and Yichang.

Objective: The goal is to analyze why users choose telehealth solutions over traditional health solutions based on a questionnaire study involving 4 age groups (50-60, 61-70, 71-80, and 80+) in 4 cities (Shenzhen, Hangzhou, Wuhan, and Yichang) in the next 10 to 20 years. The legal retirement age for female workers in China is 50 to 55 years and 60 years for male workers. To simulate reality in terms of elderly care in China, the authors use the Chinese definition of elderly for employees, defined as being 50 to 60 years old rather than 65 years, as defined by the World Health Organization.

Methods: The questionnaires were collected from Shenzhen, Hangzhou, Wuhan, and Yichang randomly with 390 valid data samples. The questionnaire consists of 31 questions distributed offline on tablet devices by local investigators. Subsequently, Stata 16.0 and SPSS 24.0 were used to analyze the data. O-logit ordered regression and principal component analysis (PCA) were the main theoretical models used. The study is currently in the exploratory stage and therefore does not seek generalization of the results.

Results: Approximately 71.09% (280/390) of the respondents reported having at least 1 type of chronic disease. We started with PCA and categorized all Likert scale variables into 3 factors. The influence of demographic variables on Factors 1, 2, and 3 was verified using analysis of variance (ANOVA) and t tests. The ordered logit regression results suggest that health-related motivations are positively related to the willingness to use telehealth solutions, and trust on data collected from telehealth solutions is negatively correlated with the willingness to use telehealth solutions.

Conclusions: The findings suggest that there is a need to address the gap in community health care and ensure health care continuity between different levels of health care institutions in China by providing telehealth solutions. Meanwhile, telehealth solution providers must focus on improving users’ health awareness and lower health risk for chronic diseases by addressing lifestyle changes such as regular exercise and social activity. The interoperability between the electronic health record system and telehealth solutions remains a hurdle for telehealth solutions to add value in health care. The hurdle is that doctors neither adjust health care plans nor diagnose based on data collected by telehealth solutions.

(JMIR Mhealth Uhealth 2022;10(1):e27272) doi:10.2196/27272

KEYWORDS

telehealth solutions; preference; motivation; elderly user; China
Introduction

Background
COVID-19 has severe effects on the elderly population with multiple chronic diseases such as hypertension, diabetes, and cardiovascular diseases than the healthy subpopulation. Mortality rate analysis shows higher death rates caused by COVID-19 associated with those aged above 50 years [1]. Population projections suggest that the Chinese population will peak from 2025 to 2030 [2], thereby leading to surging demands for high-quality medical services including telehealth solutions. Research on the preference of the elderly for telehealth solutions must be conducted. The legal retirement ages for female and male employees in China are 55 and 60 years respectively, ranking as one of the lowest in the world [3]. Therefore, to study elderly user preferences, the authors chose to start from those aged 50 years to simulate reality at the best possible level in this study.

Telehealth refers to the use of telecommunication tools for health care continuum in the prevention, treatment, diagnosis, recovery, and home care processes. The use of wearables and apps for health management and online hospitals for health consultation has become increasingly popular with the wide use of smartphones. COVID-19 has accelerated the digitization of the health care system at a pace unimaginable a few years ago. The use of telemedicine services has increased by more than 1000% in March and more than 4000% in April [4]. Spending on the use of telehealth solutions also increased starting from March by more than 1000 % [5].

There is a need to explore user preference for telehealth solutions, particularly for the future generations of elderly (aged 50 years and above) who will become 70 years old by 2030. The Chinese society is facing challenges posed by a rapidly aging population and rising chronic disease trends caused by lifestyle changes owing to the urbanization process. Based on projections, the Chinese population will peak between 2026 and 2030 [2]. Thus, researching whether and how telehealth solutions can generate more value for the elderly in the next 15 to 20 years is important. Owing to the lack of high-quality medical resources and trained clinicians, there is an urgent need to look for alternative solutions such as telehealth solutions. The implementation of telehealth solutions faces challenges among elderly users given their lack of experience with technology and thereby the lack of trust in telehealth solutions. Other factors such as the household income, education level, and health status of the users may also play a role.

The rest of the paper is structured as follows. The next section provides the literature review on research methodologies used to study users’ willingness for using telehealth solutions. We then present the methodology and research design used for the analysis, followed by the description of the qualitative and quantitative analyses of the questionnaire revealing why users choose telehealth solutions over traditional health solutions. Finally, we summarize the main findings and implications of this study.

Literature Review
To analyze the state of the art of the research methodology regarding user and physician preferences for telehealth solutions, a thorough literature review was conducted.

There have been several empirical studies on patients in all age groups and clinician perceptions regarding telehealth solutions. The multinominal logit regression model has become popular for statistical analyses in health economics and marketing science [6]. The paired t test has also been also used for comparing the preference for traditional health visits with telehealth consultations or the presence of telephysicians.

Direct-to-consumer telehealth solutions roughly comprise 3 categories [6]. The first category covers solutions provided by the same doctor from whom the patients obtain primary care services. As the health care service is provided by the doctor with whom patients have established a relationship, telehealth solutions can ensure convenience to patients while maintaining care continuity. The second category incorporates solutions provided by doctors from the same institution where patients receive health care services but not the same doctor with whom the patients have an established relationship. This allows the patients’ records to be updated by the doctor from the same care institution while maintaining the connection with the care home. Meanwhile, patients can receive care during and after working hours. The third category consists of telehealth solutions provided by doctors who have no previous relationship with the patients or the patients’ primary care service providers. Many newly emerging telehealth solution providers belong to the third category. Patients can pay for the services provided out of their pockets or by claiming insurance.

One study [7] used SurveyMonkey to send out a questionnaire for conducting a nation-wide survey in the United States. In total, there were 4345 patients covering different ethnicities, age groups, and income groups with various education backgrounds and insurance coverages. The survey aimed to determine the willingness of the participants to use telehealth solutions and their comfort level with telehealth solutions belonging to the aforementioned 3 categories of solutions. Results from the generalized estimation equation model showed that patients were more willing to use category I solutions. The willingness to use telehealth solutions declined if the provider had no relationship with patients before or if the services were provided by other doctors from the same care institutions. More than half of the patients were willing or very willing to use telehealth solutions involving their own doctor. One-third of all participants were willing to use telehealth solutions provided by other doctors from the same care institution. Less than 20% of all participants were willing to use telehealth solutions provided by doctors with whom they had no previous relationship. Patients’ comfort in using telehealth solutions grows with the attachment to their original care institution.

In another study [8], the authors have tried to analyze patient preferences and satisfaction rates with the telehealth program, CVS Minute Clinics. Minute Clinics offer patients video consultation with doctors at collaboration clinics while assisting nurses in performing on-site diagnostic tests and using tools such as otoscopes, telephonic stethoscopes, and digital video...
laryngoscopes to assist doctors in making diagnoses by reading the image or data on the screen. Such treatment costs US $59 on average with life insurance [9].

The survey participants were over 18 years old and agreed to use telemedicine service when on-site doctors were busy. The study used the logistic regression model to assess the preferences of 1734 users of the Minute Clinics services. Among these participants, 94% to 99% reported high satisfaction with telehealth solutions. One-third of all participants preferred telehealth solutions to traditional health solutions. The authors suggest that the lack of medical insurance, gender (female) of the users, self-satisfaction with the understanding of telehealth solutions, service quality, and convenience can predict user preference for telehealth visits [10]. Patients’ satisfaction with on-site nurses has an inverse relationship with the preference for telehealth solutions. The possible explanations are that the more satisfied patients are with on-site nurses, the more they are reminded of the benefits of in-person interactions. Moreover, patients may get the false impression that on-site nurses alone can perform the necessary diagnosis and therefore ignore the fact that on-site nurses do not have the license to practice alone.

One study [11] has analyzed factors associated with clinicians’ perceptions regarding telehealth solutions and examined if these factors affected their decision to continue using telehealth solutions after COVID-19. Doctors from different disciplines, including pediatricians and doctors focusing on adult patients, surgical and nonsurgical doctors, outpatient and inpatient doctors, and doctors who focus on both categories have been covered [10]. The 220 full responses also covered doctors with and without previous telehealth experiences. The study disseminated a Likert scale questionnaire and used logistic regression to analyze the odds of different factors at a significance level of 95%.

Results [10] suggest that ease of use for patients is the most important feature followed by ease of use for clinicians. Physicians’ overall satisfaction [11] and perceived ease of use [12] also directly affect perceived usefulness and the intention to use telemedicine. Meanwhile, the quality of care, ease of physical examination, and beliefs on whether adaptability is an important feature followed by ease of use for clinicians. The study conducted by Miner [10] suggests that clinicians play a significant role in adapting to the digital health trends. Training may prove necessary to help clinicians continue their telehealth practices after COVID-19.

Researchers [13] have studied outpatients’ use of the internet to search for orthopedic information. They used a questionnaire consisting of 12 questions that was distributed by doctors to outpatients during office visits. A total of 1161 complete responses were collected and analyzed with a multivariable binomial logistic regression model. Regression results show that younger age groups are primarily associated with increased use of the internet for obtaining health and orthopedic information. Younger patients are also more likely to find the search results related to their current orthopedic problems “very helpful” and “somewhat helpful.” Google is a more popular search engine than Yahoo and Bing. Patients who visited sports medicine clinics were less likely to use WebMD to search for answers to their orthopedic questions. Other than this, the type of clinic did not have a significant effect on patients’ use of the internet. Males were more likely to find information from the internet very useful than female patients; besides this, gender does not have a significant impact on patients’ internet usage. The study suggests that patients seem to conduct research on the internet with search engines more than on the website of the institution where they are being treated.

Another study [14] confirms that using the internet for searching information along with telehealth solutions, and doctors’ suggestions in clinics and hospitals shall address the problem where patients rely on search engines to search answers to medical problems because of the lack of reliable medical information sources online. Chatbots can offer an alternative for such a problem.

The study [14] compared the accuracy of traditional nurse triages and physician telepresence at an emergency pediatric department. The study used paired t tests to analyze the triage time and accuracy (triage utility) differences between traditional nurse triages and physician telepresence. In total, data on 100 families were collected in this study, which took place at a large, tertiary care children’s hospital with 65,000 emergency department visits occurring annually. Physician telepresence was achieved using the RP-7i robot, with a built-in stethoscope after the patients went through the traditional nurse triage. The questionnaire consists of 9 5-point Likert scale questions and 1 yes/no question to assess the overall experience of using the robot.

At P=.10, there is no difference in the triage time between the traditional nurse triage and physician telepresence. There are statistically significant differences between the triage accuracy of traditional nurse triages and physician telepresence (P=.03). The triage accuracy score of the traditional nurse triage is at 71% whereas the physician telepresence score is at 95%. Parents and children have preference scores for physician telepresence and indicate that they would choose physician telepresence during their next pediatric emergency department visit [14].

Another study [15] focused on children who were 5.99 years old on average. These children preferred new technology. In the emergency department, time is everything whereas it may be tricky for nurses to make accurate judgments without enough physicians in the emergency room. The robotic experience has significantly improved triage accuracy by avoiding missing values on the triage form, which consists of 27 items. This suggests that in an overwhelmed emergency room, having physician telepresence may help ease stress and avoid mistakes. The study [15] also analyzed the impact of the integrated health care buddy project with patients having chronic disease conditions in the United States. The study is a collaboration study between 2 clinics at Washington and Oregon, Robert
Bosch Healthcare and American with 2 groups of patients (an intervention group and a control group), each comprising 1767 patients with chronic obstructive pulmonary disease, congestive heart failure, or diabetes. The health buddy program gives a free handheld device for patients to use at home and a large screen. The device connects patients with care managers and allows patients to interact with their care managers about vital signs, symptoms, health-related knowledge, and behavior. Insurance claim data were used to analyze the cost for managing chronic disease and mortality rates.

Moreover, the study confirms the effectiveness of harnessing telehealth assistants for chronically ill patients. Telehealth solutions not only lowered the mortality rates by 2.7% in the intervention group over 2 years but also saved costs between 7.7% and 13.3% per patient per quarter (US $312-542). The study used multivariate regression to predict the cost reduction for patients who engaged more with the program and patients who do not engage otherwise. The prediction showed cost savings of US $1009 per congestive heart failure patient per quarter (P<.001). For patients engaged in the program, the cost saving is US $968 per patient per quarter (P<.001). For patients who did not engage with the program, the cost saving is not significant. For hospital admission rates, the study suggests that telehealth intervention lowers inpatient admission by 3.4% (P<.001).

The paper highlights the need to recognize the value of integrated telehealth solutions for high-risk patients with chronic diseases who incur high costs. Having a device at home for allowing patients to interact with care managers not only allows care managers to capture the deteriorating vital signs and provide interventions in time but also to identify gaps in patients’ health knowledge and behavior [15].

The discrete choice experiment (DCE) has been a popular tool to identify the preference for telehealth solutions and the different attributes related to the preference for such solutions [16]. Researchers tried to identify the preference of elderly people (aged 65 years and above) in Australia. The study [16] analyzed factors such as the distance to the nearest clinic and cost of virtual visits and their influence on the preference level of telehealth solutions. The study indicates that most of the elderly have never used the internet in the past 3 months, the cost of virtual visits and their influence on the preference level of telehealth solutions.

The paper continues with the analysis of factors associated with the preference of elderly users for telehealth solutions. The study used multivariate regression to predict the cost reduction for patients who engaged more with the program and patients who do not engage otherwise. The prediction showed cost savings of US $1009 per congestive heart failure patient per quarter (P<.001). For patients engaged in the program, the cost saving is US $968 per patient per quarter (P<.001). For patients who did not engage with the program, the cost saving is not significant. For hospital admission rates, the study suggests that telehealth intervention lowers inpatient admission by 3.4% (P<.001).

The discrete choice experiment (DCE) has been a popular tool to identify the preference for telehealth solutions and the different attributes related to the preference for such solutions [16]. Researchers tried to identify the preference of elderly people (aged 65 years and above) in Australia. The study [16] analyzed factors such as the distance to the nearest clinic and cost of virtual visits and their influence on the preference level of telehealth solutions. The study indicates that most of the elderly have never used the internet in the past 3 months, indicating a knowledge gap for elderly users in using telehealth solutions. In the study, 330 respondents were recruited with a mean age of 69 years. The study concludes that participants would rather use telehealth solutions only as complementary tools with in-person visits. As the study was conducted in Adelaide, Australia, where the age structure, family structure, and health status of the elderly are different from those in China, there is a need to analyze the preference of elderly users for telehealth in China.

In another study [17], the choice between mobile health and telehealth was studied with the DCE model involving 1403 residents in rural areas. The study suggests that the preference is associated with the gender and setting of the users. The distance (access to health care) to hospitals and their gender determines if the residents would prefer using telehealth solutions.

**Methods**

**Questionnaire Distribution**

With the legal retirement age standing at 55 years for female employees and 60 years for male employees in China, the questionnaire was distributed among the future elderly (aged above 50 years) in Shenzhen, Hangzhou, Wuhan, and Yichang to best simulate reality. This study followed the DCE methodology and comprised 5 stages including designing research questions, interviews with experts, interviews with individual users, the pretest stage, and the pilot test stage. The questionnaire study was conducted with assistance from the University of Chinese Academy of Sciences and Beijing Cinsos Consulting Corporation. We collected 390 valid answers from 50-60, 60-70, 70-80, and 80+ age groups to analyze individual users’ willingness to use telehealth solutions over traditional health solutions.

**Ethical Approval**

Ethical approval was obtained in May 2019 from the committee in University of Macerata.

Based on the ethical approval results and the analytical results from focus group analysis, questionnaires were designed to analyze stakeholders’ attitudes in China toward whether the Internet of Healthcare Things solutions can help reduce the gap in the demands of the current health care system.

Table 1 summarizes the collected data. All data were stored on the “Box” owned by KU Leuven.

<table>
<thead>
<tr>
<th>Data</th>
<th>Format</th>
<th>Transfer</th>
<th>Consent</th>
<th>Pseudonymization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recording with consumers</td>
<td>Windows Media Audio (WMA), MP3, MP4</td>
<td>Data were collected for scientific research purposes and therefore transferred from China to Europe and stored on cloud.</td>
<td>Question 1 in the questionnaire (see Multimedia Appendix 1)</td>
<td>Yes</td>
</tr>
<tr>
<td>Questionnaire collected on tablet devices</td>
<td>Word</td>
<td>Data were collected for scientific research purposes and therefore transferred from China to Europe and stored on cloud.</td>
<td>Question 1 in the questionnaire (see Multimedia Appendix 1)</td>
<td>Yes</td>
</tr>
<tr>
<td>Excel form with summary of data pseudonymized</td>
<td>Excel</td>
<td>Data were collected for scientific research purposes and therefore transferred from China to Europe and stored on cloud.</td>
<td>Question 1 in the questionnaire (see Multimedia Appendix 1)</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Study Design

The main purpose of the survey was to understand the factors affecting the preference of elderly users for telehealth solutions. The DCE model based on the random utility theory to evaluate the preference for telehealth solutions [18] was used. The DCE method is widely used in studying how patients value different attributes of health care services and the potential demand for new services or treatment [19]. The study follows the standard DCE methodology, namely (1) defining research questions to compile evidence, (2) interviewing experts (stakeholders), (3) interviewing individual users (focus group studies), (4) the pretest stage (online questionnaire in Europe, N=31), and (5) the pilot test stage (online questionnaire in Xiangyang, China, N=104). In the pilot test stage, 104 questionnaires were answered, with 55 questionnaires containing usable data (mostly from Hubei Province).

The questionnaire (see Multimedia Appendix 1) consists of 31 questions and 5 parts. The questionnaire starts with a screening question on whether the participant is willing to participate in the survey and share data for scientific research purposes. There are 10 Likert scale questions related to the motivation, 7 questions surrounding the demographic information including participants’ insurance coverage, 6 questions about the usage of telehealth solutions at the time of survey, 4 questions about the health status of survey participants, and 3 questions about whether users want to share data with insurance companies, doctors from community health centers, and doctors from hospitals. The degree of influence of each factor is evaluated with a Likert scale from 1 to 7 (1=no influence, 4=neutral, and 7=with influence). The questionnaire was written in Chinese and then translated in English for easy understanding by the author.

The questionnaire has 5 parts; the first part is about the current situation of telehealth solution usage by the surveyed elderly users.

Telehealth solutions are defined as smartphone apps (such as Alihealth, Ping An Good Doctor, Chun Yu Doctor, Wedoctor, Yue Dong Quan, etc.), wearables (such as Xiaomi Band, Huawei watch and Apple Watch, etc.), health management tools for home use (such as PICOOC smart scale, Mi Home i-Health blood pressure monitor, Mi Home Hi-Pee Smart Pee Monitor, Smart Sleep Monitor, Smart devices to improve sleep quality, etc.). The section consists of 4 questions asking the usage frequency, reasons for starting to use telehealth solutions, if telehealth solutions were used to monitor sleep, and if telehealth solutions were used to monitor nutrition.

The second part of the questionnaire is about the health status of the survey participants (self-evaluated). The third part asks about the potential benefits of telehealth solutions and elderly users’ motivations. The fourth part is designed around the potential risks of telehealth solutions (price, privacy risk, data accuracy risk, brand and design, resistance to technology, and usage experience). The fifth part is designed to gather demographic information, including gender, age, residence, household income, and education in years.

Further, 13 questions were designed focusing on the reasons for preferring telehealth solutions to traditional health solutions. The following questions are related to F2, the perceived benefits of telehealth solutions: monitoring health status (Q13), reducing health risks (Q14), following the doctor's advice (Q15), free devices provided by insurance companies (Q18), and lack of community health care services (Q20).

There were also questions regarding the perceived risk for telehealth solutions (F3), including data accuracy (trust) concerns (Q22), privacy concerns (Q23), financial reasons for the price (Q24), design, popularity, and usage difficulty concerns (Q25).

As some of the reasons for using telehealth solutions pertain to the social image of the individuals, social influence (Q28) is also considered one of the factors that could influence users’ preference.

Data Collection

In China, the legal retirement age for female factory workers is 50 years, 55 years for female employees, and 60 years for male employees. To simulate reality with respect to elderly care in China, elderly is defined as being over 50 years old instead of being 65 years old according to the World Health Organization.

In our study, we intended to compare the participants’ willingness to use telehealth solutions considering different age groups and residents in different cities, with the data collection target set for each age group (50-60, 61-70, and 71-80 years) containing approximately 100 data subjects. Data subjects more than 80 years old were categorized as being in the 50-100 years group because of the health conditions that limited the number of participants.

In the pretest stage of the study, questionnaires in English were distributed on the internet via Microsoft Forms through the Philips intranet portal and Berlin Expat Group on Facebook. We collected 31 questionnaires. In the pilot testing stage, the questionnaire was translated into Chinese and distributed via the internet with Wenjuanxing through WeChat. We collected 104 questionnaires with 55 valid answers. The pretesting stage was designed to test the design of the questionnaire; therefore, the data collected were not analyzed.

In the distribution stage, questionnaires were disseminated on tablet devices by local investigators randomly among residents more than 50 years old in Shenzhen, Hangzhou, Wuhan, and Yichang with the help of Beijing Cinso Consulting Corporation. More than 450 questionnaires were distributed, and 402 answers were collected, with a recovery rate of 89%. Among them, 390 were completely valid questionnaires, accounting for 87% of the questionnaires issued and 97% of all the questionnaires returned. The other 12 questionnaires were not used in data analysis because they did not provide complete information or were deemed to have not been filled carefully.

The data were collected in Chinese language and then summarized in an Excel sheet (Microsoft Corporation) and converted into a pseudonymized value form in Excel. Data were then analyzed using SPSS 24.0 (IBM Corporation) and Stata 16.0 (StataCorp).
The level of urban development differs with Tier 1, 2, 3, and 4 cities; the disposable income of residents in the designated cities and the medical resources accessible (hospitals and doctors) vary as well (see Table 2). This may lead to differences in the preference for telehealth solutions.

Table 2. Disposable income in Shenzhen, Hangzhou, Wuhan, and Yichang (source: CEIC, 2020; National bureau of statistics, 2020).

<table>
<thead>
<tr>
<th>Category</th>
<th>City</th>
<th>GDP(^a) in 2019 (billion US$)</th>
<th>Disposable and discretionary income (US$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tier 1</td>
<td>Shenzhen</td>
<td>422.875</td>
<td>9818.83</td>
</tr>
<tr>
<td>Tier 2</td>
<td>Hangzhou</td>
<td>241.425</td>
<td>10357.65</td>
</tr>
<tr>
<td>Tier 3</td>
<td>Wuhan</td>
<td>254.744</td>
<td>8120.17</td>
</tr>
<tr>
<td>Tier 4</td>
<td>Yichang</td>
<td>70.05</td>
<td>4518.5</td>
</tr>
</tbody>
</table>

\(^a\)GDP: gross domestic product.

Shenzhen was chosen because it is the headquarter of Ping An Technology. Ping An Technology has worked with the government of Shenzhen and other stakeholders to provide electronic medical insurance schemes. Residents in Shenzhen can now use the Ping An Good Doctor app to buy complementary insurance in addition to the basic medical insurance schemes and get refunded online.

Hangzhou was chosen, as it is the city where Alibaba is headquartered. During the interview with Alihealth, the fact that 80% of all primary health care facilities in Zhejiang Province are now equipped with artificial intelligence–assisted image recognition systems was mentioned.

Wuhan was chosen, as it is an important hub in Central China where the population is growing rapidly in recent years. Recently, the Wuhan Municipality has launched several programs promoting the internet + home care initiative for the elderly. There are several exploratory projects running in different districts in Wuhan such as in the Dongxihu and Wuchang districts. There have been several models proposed and tested in Wuhan for elderly care such as the community embedded model, centralization model, and combinations of the proper centralization and decentralization models. Services provided to the elderly focus on assisted food service, assisted cleaning service, assisted nursing and medical service, and long-distance care.

Yichang was chosen, as the level of aging population in Yi Chang is higher than the national average. Aging was measured by the percentage of people over 60 years old in the entire population and the percentage of people over 80 years old in the elderly population. The Yichang municipality is currently developing community-based care centers and rural cooperative elderly care centers. The Yichang municipality established the telehealth solution platform for elderly care in 2019. Using the platform, in 2020, the tele-elderly-care (translated from Chinese) services were expected to reach all townships in Yichang and cover over 50% of all elderly people.

Theoretical Model and Hypothesis

To evaluate users’ willingness to use telehealth solutions, 3 hypotheses were formulated; in addition, the model considers the effects of demographic factors such as age, education background, income, health status, and living habits such as regular social activity and regular exercise. Table 3 describes the theoretical model built to assess users’ willingness to use telehealth solutions, the hypothesis, and the variables involved in the model. The theoretical model consists of 2 parts; the first part assesses the Likert scale factors and their correlation with the users’ willingness; the second part assesses demographic factors and their impact on the 3 factors and the willingness to use telehealth solutions.

Considering that the dependent variable, namely the willingness to use telehealth solutions, is an ordered discrete variable, the ordered logit model is used for regression. The impact of each factor was assessed by designing 4 models.

\[ Y = \beta F_1 + \gamma Z + \varepsilon \quad (1) \]
\[ Y = \beta F > 2 + \gamma Z + \varepsilon \quad (2) \]
\[ Y = \beta F_3 + \gamma Z + \varepsilon \quad (3) \]
\[ Y = \beta_1 F_1 + \beta_2 F_2 + \beta_3 F_3 + \gamma Z + \varepsilon \quad (4) \]

Model (1) is used to test the impact of Factor 1, and models (2) and (3) are used to test the impact of Factors 2 and 3. Model (4) considers the influence of the above 3 factors.

\( Y \) represents the designated value for the willingness of participants to use telehealth solutions. In the original questionnaire, the question assigned the preference level as from 1 to 7 (1=preference for traditional health solutions [face-to-face communication], 4=neutral, and 7=preference for telehealth solutions), Z represents the control variables such as demographic factors, including the living city, age, gender, education level, health condition, income, living situation, and lifestyle variables (regular exercise and regular social activity) of the participants.
<table>
<thead>
<tr>
<th>Factor</th>
<th>Hypothesis</th>
<th>Corresponding question in the questionnaire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor 1</td>
<td>1.1 Social influence (friend and family opinions) has an impact on the willingness to use telehealth solutions.</td>
<td>Q28, Likert scale value: 1-7</td>
</tr>
<tr>
<td>Factor 1</td>
<td>1.2 The price of telehealth solutions has an impact on the willingness to use telehealth solutions.</td>
<td>Q24, Likert scale value: 1-7</td>
</tr>
<tr>
<td>Factor 1</td>
<td>1.3 The brand and design of telehealth solutions have an impact on the willingness to use telehealth solutions.</td>
<td>Q25, Likert scale value: 1-7</td>
</tr>
<tr>
<td>Factor 1</td>
<td>1.4 The privacy risk associated with the use of telehealth solutions has an impact on the willingness to use these solutions.</td>
<td>Q23, Likert scale value: 1-7</td>
</tr>
<tr>
<td>Factor 1</td>
<td>1.5 Private or business insurance plan coverage has an impact on the willingness to use telehealth solutions.</td>
<td>Q18, Likert scale value: 1-7</td>
</tr>
<tr>
<td>Factor 2: Health-related motivation factors</td>
<td>2.1 The belief that telehealth solutions can lower health risk is positively related to the willingness to use telehealth solutions.</td>
<td>Q14, Likert scale value: 1-7</td>
</tr>
<tr>
<td>Factor 2: Health-related motivation factors</td>
<td>2.2 The belief that telehealth solutions can raise health awareness is positively related to the willingness to use telehealth solutions.</td>
<td>Q13, Likert scale value: 1-7</td>
</tr>
<tr>
<td>Factor 2: Health-related motivation factors</td>
<td>2.3 The belief that telehealth solutions can amend the gap in the lack of community health care for patients has an impact on the willingness to use telehealth solutions.</td>
<td>Q22, Likert scale value: 1-7</td>
</tr>
<tr>
<td>Factor 2: Health-related motivation factors</td>
<td>2.4 The belief that telehealth solutions can help improve doctor-patient relationship has an impact on the willingness to use telehealth solutions.</td>
<td>Q15, Likert scale value: 1-7</td>
</tr>
<tr>
<td>Factor 3: Trust</td>
<td>3. Data accuracy</td>
<td>Q22, Likert scale value: 1-7</td>
</tr>
<tr>
<td>Control variables</td>
<td>Residence city</td>
<td>The residence city of the participants has an impact on Factors 1, 2, and 3 and their willingness to use telehealth solutions.</td>
</tr>
<tr>
<td>Control variables</td>
<td>Gender</td>
<td>The gender of the participants has an impact on Factors 1, 2, and 3 and their willingness to use telehealth solutions.</td>
</tr>
<tr>
<td>Control variables</td>
<td>Education</td>
<td>The education level of the participants has an impact on Factors 1, 2, and 3 and their willingness to use telehealth solutions.</td>
</tr>
<tr>
<td>Control variables</td>
<td>Income</td>
<td>The monthly household income of the participants has an impact on Factors 1, 2, and 3 and their willingness to use telehealth solutions.</td>
</tr>
<tr>
<td>Control variables</td>
<td>Health status</td>
<td>The self-reported health status of the participants has an impact on Factors 1, 2, and 3 and their willingness to use telehealth solutions.</td>
</tr>
</tbody>
</table>
Results

Health Status of Survey Participants

Based on the self-identified responses from the subjects, the following categories were created to identify their health status: healthy, suboptimal healthy, with chronic disease, and self-identified healthy. Then, more detailed data, such as the type of chronic disease and the number of chronic diseases of the survey participants, were analyzed.

Among the 390 participants, 117 (30%) reported having 1 chronic disease (30%), 64 (16.4%) reported having 2 chronic diseases (16.4%), and 47 (12.05%) responded as having 3 chronic diseases (12.05%). Further, 17 participants (4.36%) stated having 4 chronic diseases (4.36%), 7 (1.79%) reported having 5 chronic diseases, 6 (1.53%) reported having 6 chronic diseases, and 2 (0.5%) reported having communicable and chronic diseases (0.5%). Furthermore, 110 participants (28%) reported having no chronic diseases.

Descriptive Statistics

In this section, the qualitative analytical results are presented. All the survey participants are over 50 years old because the survey intends to collect information on elderly users’ needs in the next 5 to 10 years, as shown in Table 4. Among the 390 participants providing valid answers, 160 (41.03%) indicate that they are more willing to use traditional health care solutions; 167 (42.82%) indicate that they are willing to use telehealth solutions, whereas 51 (13.07%) show neutral willingness.

Among the 390 participants, 112 (28.7%) are aged 51 to 60 years; another 112 participants (28.7%) are aged 61 to 70 years. Further, 110 participants (28.2%) are aged 71 to 80 years is. There are 43 participants (11.4%) over 80 years old; the number of participants in this group is less than that in the other 3 age groups, as data subject recruitment was limited by the physical conditions of the individuals in this age group.

Moreover, 67.7% (264/390) of the participants often use telehealth solutions to monitor health status. Most survey participants (246/390, 63.1%) received 6 to 12 years of education, followed by 83 participants (21.3%) who went to elementary school. Given the survey candidate recruitment conditions for the elderly aged above 50 years, the education level of the participants is in line with the reality.

The distribution diagram in Figure 1 shows that participants in the age group of 51 to 60 years and those aged above 80 years show a strong willingness or a willingness to use telehealth solutions. This may be because people in the age group of 50 to 60 years are more familiar with technology, whereas those above 80 years cannot physically attend in-person doctor visits.
Table 4. Demographic characteristics of participants (N=390).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>n (% )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>224 (57.4)</td>
</tr>
<tr>
<td>Female</td>
<td>166 (42.6)</td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
</tr>
<tr>
<td>51-60</td>
<td>112 (28.7)</td>
</tr>
<tr>
<td>61-70</td>
<td>112 (28.7)</td>
</tr>
<tr>
<td>71-80</td>
<td>110 (28.2)</td>
</tr>
<tr>
<td>&gt;=80</td>
<td>56 (14.4)</td>
</tr>
<tr>
<td>Residence city</td>
<td></td>
</tr>
<tr>
<td>Shenzhen</td>
<td>97 (24.9)</td>
</tr>
<tr>
<td>Hangzhou</td>
<td>95 (24.4)</td>
</tr>
<tr>
<td>Wuhan</td>
<td>108 (27.7)</td>
</tr>
<tr>
<td>Yichang</td>
<td>90 (23.1)</td>
</tr>
<tr>
<td>Household income (US$, original value in RMB, $1 = 6.37 RMB)</td>
<td></td>
</tr>
<tr>
<td>No fixed monthly income</td>
<td>21 (5.4)</td>
</tr>
<tr>
<td>(&lt;$785.23)</td>
<td>84 (21.5)</td>
</tr>
<tr>
<td>($785.23-1570.45)</td>
<td>186 (47.7)</td>
</tr>
<tr>
<td>($1570.45-4711.35)</td>
<td>88 (22.6)</td>
</tr>
<tr>
<td>(\geq$4711.35)</td>
<td>11 (2.8)</td>
</tr>
<tr>
<td>Frequency of using telehealth solutions</td>
<td></td>
</tr>
<tr>
<td>Often</td>
<td>264 (67.7)</td>
</tr>
<tr>
<td>Occasionally</td>
<td>82 (21)</td>
</tr>
<tr>
<td>Rarely</td>
<td>44 (11.3)</td>
</tr>
<tr>
<td>Education level</td>
<td></td>
</tr>
<tr>
<td>Primary school (1-6 years)</td>
<td>83 (21.3)</td>
</tr>
<tr>
<td>Junior or high school (6-12 years)</td>
<td>246 (63.1)</td>
</tr>
<tr>
<td>Vocational training (12-15 years)</td>
<td>31 (7.9)</td>
</tr>
<tr>
<td>College graduate (15-18 years)</td>
<td>29 (7.4)</td>
</tr>
<tr>
<td>Graduate School ((\geq)18 years)</td>
<td>1 (0.3)</td>
</tr>
<tr>
<td>Health status</td>
<td></td>
</tr>
<tr>
<td>Healthy</td>
<td>145 (37.2)</td>
</tr>
<tr>
<td>Suboptimal healthy</td>
<td>99 (25.4)</td>
</tr>
<tr>
<td>With minor chronic disease</td>
<td>132 (33.8)</td>
</tr>
<tr>
<td>With major chronic disease affecting life quality</td>
<td>14 (3.6)</td>
</tr>
<tr>
<td>Living situation</td>
<td></td>
</tr>
<tr>
<td>Living alone</td>
<td>47 (12.1)</td>
</tr>
<tr>
<td>Living with partner</td>
<td>156 (40)</td>
</tr>
<tr>
<td>Living with children</td>
<td>177 (45.4)</td>
</tr>
<tr>
<td>Living with grandchildren</td>
<td>4 (1)</td>
</tr>
<tr>
<td>Health insurance status</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>8 (2.1)</td>
</tr>
<tr>
<td>Basic resident or employee medical insurance</td>
<td>309 (79.2)</td>
</tr>
</tbody>
</table>
Among the 390 users surveyed, there are 224 males and 166 females, accounting for 57.4% and 42.6% of the total number of participants, respectively; the proportion of male users is higher than that of female users (as observed in Table 4). Figure 2 suggests that female users are willing to use traditional medical solutions, whereas male users are strongly willing to use telehealth solutions.

In accordance with the study design, survey participants are evenly distributed in the 4 cities. The number of data samples obtained in Shenzhen, Hangzhou, Wuhan, and Yichang are 97, 95, 108, and 90, accounting for 24.9%, 24.4%, 27.7%, and 23.1%, respectively, of the 390 data subjects. Figure 3 suggests that in Shenzhen and Wuhan, the percentage of participants showing preference for telehealth solutions is higher than that in Hangzhou and Yichang.

The distribution of income follows the bell curve, with approximately half (186/390, 47.7%) of the sample’s monthly household income falling between US$ 785.23 and US$ 1570.45; the proportions of the sample with household monthly incomes less than or equal to US$ 785.23 and more than or equal to US$ 4711.35 account for only 5.4% (21/390) and 2.8% (11/390), respectively. The willingness to use telehealth solutions increases with the monthly income as well. Figure 4 points out that in the >=US$ 4711.35 income group, the preference is mainly neutral and above neutral. The lower the income, the higher the percentage of the surveyed data subjects showing strong preference for traditional health solutions. This can be observed among the no income and <=US$ 785.23 income groups.

In terms of using telehealth solutions for monitoring sleep and nutrition intake, the percentage of users who are currently using telehealth solutions for sleep monitoring and nutrition monitoring are respectively 23.07% (90/390) and 26.15% (102/390).

Figure 5 indicates that the major reasons for using telehealth devices are self-care, following the doctor’s advice, and the free devices and services offered by insurance companies in China or a mix of these 3 reasons.
Figure 2. Willingness to use telehealth solutions based on gender.

Figure 3. Willingness to use telehealth solutions in Yichang, Wuhan, Hangzhou, and Shenzhen.
The factors affecting the willingness to use telehealth solutions are ranked by the mean of each variable (Likert scale: 1-7, 1=no impact, 4=neutral, and 7=with an impact), as shown in Table 5. Among the 10 factors, 6 have means more than 4, suggesting that these factors influence the preference for telehealth solutions. The top 4 motivations are lowering health risks, raising health care awareness, lack of community medical services, and following the doctor’s advice; these variables comprise Factor 1.

Factor 2 consists of the price, privacy risk, social influence, design and brand of the solution, and the participants’ coverage with insurance plans. The mean value of these variables is close to neutral or less than 4, suggesting that survey participants in general do not believe that these factors influence their willingness to use telehealth solutions.

The accuracy of the data (Factor 3) collected through telehealth solutions is also a key factor. Compared with traditional medical instruments and equipment having the shortcoming of inaccurate
data reading, telehealth solutions collect more accurate health data. However, most doctors and hospitals still do not trust data collected from telehealth solutions and do not use these data sources as the basis for diagnosis or treatment. This makes it difficult for users to trust the devices used for collecting health data and monitoring health status.

**Table 5.** Ranking of factors affecting the willingness to use telehealth solutions among elderly users.

<table>
<thead>
<tr>
<th>Factors</th>
<th>Ranking</th>
<th>Mean</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lowering health risk</td>
<td>1</td>
<td>5.96</td>
<td>1.672</td>
</tr>
<tr>
<td>Raising health awareness</td>
<td>2</td>
<td>5.85</td>
<td>1.676</td>
</tr>
<tr>
<td>Lack of community health care service</td>
<td>3</td>
<td>5.77</td>
<td>1.721</td>
</tr>
<tr>
<td>Following doctors’ prescriptions</td>
<td>4</td>
<td>5.27</td>
<td>2.032</td>
</tr>
<tr>
<td>Price of the solution</td>
<td>5</td>
<td>4.37</td>
<td>2.462</td>
</tr>
<tr>
<td>Data accuracy</td>
<td>6</td>
<td>4.07</td>
<td>2.314</td>
</tr>
<tr>
<td>Design of the solution</td>
<td>7</td>
<td>3.72</td>
<td>2.498</td>
</tr>
<tr>
<td>Privacy risk</td>
<td>8</td>
<td>3.70</td>
<td>2.342</td>
</tr>
<tr>
<td>Social influence</td>
<td>9</td>
<td>3.44</td>
<td>2.540</td>
</tr>
<tr>
<td>Free device offered by insurance companies</td>
<td>10</td>
<td>2.77</td>
<td>2.157</td>
</tr>
</tbody>
</table>

**Modeling Process**

This section presents the quantitative analytical results.

To avoid heterogeneity issues, the Kaiser–Meyer–Olkin (KMO) and Bartlett test was conducted to examine the correlation between the Likert scale variables; the KMO score of 0.796 suggests that the sample is adequate for factor analysis. Then, principal component analysis (PCA) was performed to reduce the dimensions of the model and the correlation between variables. With the factor loading for each factor confirmed, the Likert scale variables were then ranked based on the mean value of each variable. The next step was to test if demographic factors influenced the 3 factors identified by PCA. This was confirmed with analysis of variance (ANOVA) and t tests.

The modeling process started with a correlation matrix (Pearson correlation and Spearman rank correlation) to test if the data have multicollinearity (see Multimedia Appendix 2). Then, the O-logit model was run using Stata 16.0 along with the control variables. During the final modeling step, 10 participants were randomly selected to determine if the prediction preference scores matched the choices made by the participants.

The KMO and Bartlett test was conducted on 10 Likert scale factors related to survey participant preferences. The KMO coefficient is 0.796 (>0.5) with the Sig. value of 0.000 in the Bartlett sphere being less than 0.05, indicating that there is a certain degree of correlation among the 10 factors. Dimension reduction among the 10 factors was deemed necessary for further analysis.

Factor analysis is a commonly used dimensionality reduction method. PCA and the varimax right-angle rotation method were used to extract 3 principal factors. These 3 principal factors could explain 64.149% of the total variance, with the first, second, and third factors explaining 28.364%, 25.196%, and 10.589% of the total variance, respectively (see Tables 6 and 7). The variables selected had a factor loading greater than 0.5 (See Table 6).

Factor 1 consists of the price (0.812), design (0.738), impact of private insurance coverage (0.713), social influence (0.706), and privacy risk (0.612).

Factor 2, involving health-related motivations, consists of lowering health risk (0.864), raising health awareness (0.818), lack of community health care services (0.771), and following the doctor’s advice or prescription (0.701).

Factor 3, related to the trust for telehealth solutions, consists of the data accuracy variable. Users’ trust levels for telehealth solutions are influenced by whether data collected from wearables or medical devices at home are accepted by doctors and hospitals. Therefore, trust is influenced directly by the data accuracy of the solution (0.762).

ANOVA and t tests were conducted for assessing whether the relative importance of the 3 main factors from PCA analysis differ, depending on the city of residence, age, gender, education level, health status, income, living situation, regular social activity, and regular exercise. The results are shown in Table 8.
Table 6. Principal component analysis.

<table>
<thead>
<tr>
<th>Component</th>
<th>Factor loading</th>
<th>Eigenvalue</th>
<th>Variance contribution rate</th>
<th>Cumulative contribution rate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Factor 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Price</td>
<td>0.812</td>
<td>2.836</td>
<td>28.364</td>
<td>28.364</td>
</tr>
<tr>
<td>Brand and design</td>
<td>0.738</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private insurance coverage</td>
<td>0.713</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social influence</td>
<td>0.706</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Privacy risk</td>
<td>0.612</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Factor 2: Health-related motivations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower health risk</td>
<td>0.864</td>
<td>2.520</td>
<td>25.196</td>
<td>53.560</td>
</tr>
<tr>
<td>Raise health awareness</td>
<td>0.818</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lack of community health care service</td>
<td>0.771</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unstable doctor-patient relationship</td>
<td>0.701</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Factor 3: Trust</strong></td>
<td>1.059</td>
<td>10.589</td>
<td>64.149</td>
<td></td>
</tr>
<tr>
<td>Data accuracy</td>
<td>0.762</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7. Total variance explaineda.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Rotation sums of squared loadings</th>
<th>% variance</th>
<th>Cumulative %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>28.364</td>
<td>28.364</td>
</tr>
<tr>
<td>1</td>
<td>2.836</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2.520</td>
<td>25.196</td>
<td>53.560</td>
</tr>
<tr>
<td>3</td>
<td>1.059</td>
<td>10.589</td>
<td>64.149</td>
</tr>
</tbody>
</table>

*aExtraction method: PCA.

First, we assumed that the relative importance of factors varies with the residence city. The results of variance analysis support this hypothesis.

The first factor is related to the price, brand, and design associated with the telehealth solutions; factors such as social influence and private insurance coverage are also included.

Second, we assumed that age plays a significant role in determining user preference. In this study, survey participants were divided into 4 age groups, 51-60, 61-70, 71-80, and over 80 years. Considering that 71.79% (280/390) of all the survey participants have chronic diseases, users from this age group may consider the relevant health benefits such as raising health awareness, lowering health risk, and improving access to health care more than other age groups. ANOVA results suggest that the second factor varies with age (P=0.088). The second factor mainly reflects the belief that telehealth solutions can raise health awareness, lower health risk, improve doctor-patient relationships and amend the gap related to the lack of community health care services.

Gender is also one of the key factors affecting user preference. The hypothesis is that male and female users have a perceived value, perceived risk, and perceived benefit associated with telehealth solutions. Considering the binary factor of gender, the t test could verify our hypothesis. The results show that the trust factor is significant at the level of 1%. This suggests that male and female survey participants differ in their trust on the data accuracy risk related to telehealth solutions.

The survey classifies users’ education levels by years into five categories: primary school (1-6 years), high school (6-12 years), vocational school (12-15 years), college education (15-18 years), and postgraduation (>=18 years). Our hypothesis is that Factors 1, 2, and 3 differ across different education levels. However, the ANOVA results reject our hypothesis. With data suggesting that 84.4% of all survey participants have high school or primary school education, the conclusion is that survey participants with less than 15 years of education show no difference in Factors 1, 2, and 3.

The health status of survey participants is divided into four categories: self-reported healthy, suboptimal health status, with chronic disease (does not affect life quality), and with chronic disease (affects life quality). The variance analysis results support our hypothesis. The third factor, trust over data accuracy regarding telehealth solutions, is statistically significant and is affected by the health status of survey participants.

ANOVA results suggest that household income has a statistically significant effect on Factor 2, consisting of factors pertaining to health-related motivations. Families with high household incomes can bear the cost of using telehealth solutions, thereby benefiting from active self-health management. Users in the lower household income group pay more attention to factors such as the price of telehealth solutions, often ignoring the need for high-quality health care.
for active health management. Factor 2 varies among different income groups.

Trust over data accuracy regarding telehealth solutions (Factor 3) is also affected by whether survey participants live with their children or grandchildren. The survey participants’ living situations are categorized as prefer living alone, prefer living with spouse, prefer living with children, and prefer living with grandchildren. Usually, it is the children and grandchildren living with their parents or grandparents who pay for telehealth solutions and teach their parents and grandparents to use such solutions. The elderly people thus benefit from living with their children or grandchildren and trust the telehealth solutions more than those who live alone or with spouses only.

Table 8. One-way analysis of variance and two-sample t test.

<table>
<thead>
<tr>
<th>Hypothesis testing and variance analysis value</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residence city</td>
<td>5.718 (3, 386)</td>
<td>2.245 (3, 386)</td>
<td>4.075 (3, 386)</td>
</tr>
<tr>
<td>F value (df1, df2)</td>
<td>.001</td>
<td>.083</td>
<td>.007</td>
</tr>
<tr>
<td>P value</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.467 (3, 386)</td>
<td>2.195 (3, 386)</td>
<td>0.172 (3, 386)</td>
</tr>
<tr>
<td>F value (df1, df2)</td>
<td>.71</td>
<td>.088</td>
<td>.92</td>
</tr>
<tr>
<td>P value</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>0.074 (1, 388)</td>
<td>2.128 (1, 388)</td>
<td>7.570 (1, 388)</td>
</tr>
<tr>
<td>F value (df1, df2)</td>
<td>.79</td>
<td>.15</td>
<td>.006</td>
</tr>
<tr>
<td>P value</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>1.186 (4, 385)</td>
<td>0.180 (4, 385)</td>
<td>1.374 (4, 385)</td>
</tr>
<tr>
<td>F value (df1, df2)</td>
<td>.32</td>
<td>.95</td>
<td>.24</td>
</tr>
<tr>
<td>P value</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health condition</td>
<td>1.494 (3, 386)</td>
<td>1.128 (3, 386)</td>
<td>3.468 (3, 386)</td>
</tr>
<tr>
<td>F value (df1, df2)</td>
<td>.22</td>
<td>.34</td>
<td>.02</td>
</tr>
<tr>
<td>P value</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>1.261 (4, 385)</td>
<td>4.109 (4, 385)</td>
<td>1.436 (4, 385)</td>
</tr>
<tr>
<td>F value (df1, df2)</td>
<td>.29</td>
<td>.003</td>
<td>.22</td>
</tr>
<tr>
<td>P value</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Living situation</td>
<td>1.216 (5, 384)</td>
<td>1.136 (5, 384)</td>
<td>2.665 (5, 384)</td>
</tr>
<tr>
<td>F value (df1, df2)</td>
<td>.30</td>
<td>.34</td>
<td>.022</td>
</tr>
<tr>
<td>P value</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regular social activity</td>
<td>5.998 (1, 388)</td>
<td>2.508 (1, 388)</td>
<td>2.083 (1, 388)</td>
</tr>
<tr>
<td>F value (df1, df2)</td>
<td>.015</td>
<td>.11</td>
<td>.15</td>
</tr>
<tr>
<td>P value</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regular exercise</td>
<td>4.726 (1, 388)</td>
<td>3.963 (1, 388)</td>
<td>0.605 (1, 388)</td>
</tr>
<tr>
<td>F value (df1, df2)</td>
<td>.03</td>
<td>.047</td>
<td>.44</td>
</tr>
<tr>
<td>P value</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The t test results suggest that regular social activity has a statistically significant effect on Factor 1. Peer pressure from regular social interaction may encourage users to choose telehealth solutions over social influence, brand and design, and insurance plans. Survey participants with poor physical conditions often lack social activity and are subject to less social influence when it comes to using telehealth solutions.

Factors 1 and 2 also differ in terms of whether users exercise regularly. Survey participants exercising regularly are more health conscious and more willing to spend on telehealth solutions such as wearables and believe that telehealth solutions may raise health awareness, lower health risks, amend the gap in community health care and ensure health continuity by improving unstable doctor-patient relationships.

**Factor 1**

Ordered logit regression results suggest that Factor 1 has no statistically significant impact on the preference for telehealth solutions (See Table 9, rows 1 and 3, columns 1 and 4). Therefore, hypothesis 1 is rejected.
<table>
<thead>
<tr>
<th>Factor 1</th>
<th>(1) Coefficient</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.0227 (P=.81)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.2449 (379)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factor 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.0235 (P=.80)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.2604 (379)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factor 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.0263 (P=.90)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.5109 (379)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Living city</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.0235 (P=.80)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.2604 (379)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0151 (P=.88)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1553 (379)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.2882 (P=.13)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.5150 (379)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edcation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1317 (P=.28)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.0778 (379)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health status</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0743 (P=.48)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.7034 (379)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.3707 (P=.001)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.3153 (379)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Living situation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0545 (P=.66)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.4400 (379)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regualar socialization</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1871 (P=.069)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.8191 (379)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regualar exercise</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.0572 (P=.64)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.4688 (379)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>390</td>
<td>390</td>
<td>390</td>
<td>390</td>
</tr>
<tr>
<td>Pseudo R²</td>
<td>0.0169</td>
<td>0.0322</td>
<td>0.0217</td>
<td>0.0384</td>
</tr>
</tbody>
</table>

*aNot applicable.*
**Factor 2**

Hypothesis 2 suggests the correlation between the willingness to use telehealth solutions and the health-related reasons. The regression coefficients of Factor 2 in Models 2 and 4 in Table 9 are positive at the significance level ($P=.01$). Hence, hypothesis 2 is valid.

**Factor 3 and Preference**

Hypothesis 3 assumes that data accuracy risk has a significant impact on the preference of elderly users. Table 9 suggests that the belief in the accuracy of the data collected by telehealth solutions is negatively related to the preference for telehealth solutions in Models 3 and 4. The cut variable suggests there are 7 categories of the dependant variable, where cut 1 puts the category at the lower end when $y$ equals to 0.

Table 10. Model validation$^a$.

<table>
<thead>
<tr>
<th>No.</th>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
<th>p5</th>
<th>p6</th>
<th>p7</th>
<th>Y</th>
<th>Prediction results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.336</td>
<td>0.153</td>
<td>0.149</td>
<td>0.127</td>
<td>0.107</td>
<td>0.070</td>
<td>0.058</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>0.437</td>
<td>0.158</td>
<td>0.136</td>
<td>0.104</td>
<td>0.079</td>
<td>0.049</td>
<td>0.039</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>3</td>
<td>0.174</td>
<td>0.111</td>
<td>0.139</td>
<td>0.153</td>
<td>0.163</td>
<td>0.131</td>
<td>0.129</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>0.127</td>
<td>0.088</td>
<td>0.121</td>
<td>0.148</td>
<td>0.178</td>
<td>0.161</td>
<td>0.177</td>
<td>7</td>
<td>no</td>
</tr>
<tr>
<td>5</td>
<td>0.135</td>
<td>0.093</td>
<td>0.124</td>
<td>0.150</td>
<td>0.176</td>
<td>0.155</td>
<td>0.167</td>
<td>7</td>
<td>no</td>
</tr>
<tr>
<td>6</td>
<td>0.052</td>
<td>0.042</td>
<td>0.067</td>
<td>0.101</td>
<td>0.164</td>
<td>0.212</td>
<td>0.362</td>
<td>7</td>
<td>yes</td>
</tr>
<tr>
<td>7</td>
<td>0.044</td>
<td>0.036</td>
<td>0.059</td>
<td>0.091</td>
<td>0.154</td>
<td>0.213</td>
<td>0.402</td>
<td>7</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>0.130</td>
<td>0.090</td>
<td>0.122</td>
<td>0.148</td>
<td>0.177</td>
<td>0.159</td>
<td>0.173</td>
<td>5</td>
<td>yes</td>
</tr>
<tr>
<td>9</td>
<td>0.202</td>
<td>0.122</td>
<td>0.145</td>
<td>0.152</td>
<td>0.153</td>
<td>0.117</td>
<td>0.110</td>
<td>1</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td>0.076</td>
<td>0.058</td>
<td>0.088</td>
<td>0.124</td>
<td>0.179</td>
<td>0.199</td>
<td>0.276</td>
<td>7</td>
<td>yes</td>
</tr>
</tbody>
</table>

$^a$p1 to p7: possibilities of the respondents choosing different categories of preferences.

**Discussion**

**Principal Findings**

Digitalization of the health care system has been rapidly accelerated by COVID-19. Because of social distancing and the highly communicable nature of the disease, the use of telehealth solutions grew exponentially, with expenditure on such solutions increasing as well. The high number of COVID-19 patients has consumed hospital and medical resources rapidly, depriving medical care for many patients with chronic diseases. The importance of using telemonitoring and telehealth solutions inside and outside the hospital setting has become more important than ever.

This study analyzed questionnaire data collected on factors related to the preference for telehealth solutions in Shenzhen, Hangzhou, Wuhan, and Yichang. The preference is related to the following factors: F1-perceived value of telehealth solutions related to product price, design, and social influence; F3-perceived risk for telehealth solutions related to trust over telehealth solutions; and F2-perceived benefits for telehealth solutions in self-care and health management. ANOVA and $t$ tests were conducted to verify the influence of demographic variables on Factors 1, 2, and 3. The ordered logit regression results suggest that the perceived value (F1) has no significant impact on the preference for telehealth solutions, indicating the homogeneity of current consumer-facing telehealth solutions. Perceived benefits in self-care and health management (F2) are positively related with the preference for telehealth solutions, and F3 (trust over data accuracy) is negatively correlated with the preference for telehealth solutions.

The preference distribution graph (Figure 2) suggests that females are more conservative than males in their preference for telehealth solutions. This may be related to their income, health education, and tendency to socialize. This will make female users more willing to communicate with doctors face to face and use traditional health solutions.

Reasons behind the differences in the preference for telehealth solutions among residents from different cities can be explained based on infrastructure differences, such as the smart health initiative in Shenzhen and Wuhan and the concentration of hospitals and other medical resources in Wuhan. Shenzhen and Wuhan started early in their big data + health initiative, whereas other cities started later. The big data + health initiative provides the necessary digital infrastructure (hospital information system) for health system digitalization including the digitalization of hospitals and the connection of primary health service clinics.
The preference for telehealth solutions varies by income, and this may exist because communications with doctors are part of social activity; such social activities are strongly related to health education, social influence, and health insurance coverage. The lower income groups have less insurance coverage and less health education; they have a strong preference for seeing doctors in person and spending hours at hospitals because they have fewer other social activities. Higher income groups have better health education, wider health insurance coverage, and considerably less willingness to spend time at hospitals. Therefore, as they can afford telehealth solutions, based on the premise that telehealth solutions are not covered by the social (employee or resident) medical insurance schemes in China, the higher income groups are more willing to use telehealth solutions rather than wait at hospitals.

It is still necessary to address the data interoperability issue between hospitals and the primary health service clinics to ensure patients’ care continuity; this may also help reduce the concentration of patients at hospitals and divert them back to primary care institutions. After all, doctors at level-3 hospitals have no time for helping outpatients to address their lifestyle problems once they leave the hospital. This gap leaves room for community care centers to step in and advise patients and monitor them regularly. The vacuum for community care centers can be filled in by telehealth solutions for patient self-care.

There are more telehealth solution providers in Shenzhen and Hangzhou, as well as high-quality hospitals. Doctors and nurses are more acceptable for telehealth solutions in large cities such as Shenzhen and Hangzhou. With the fast-paced lifestyle in these 2 cities, elderly (aged 50 years and above) users are more willing to use telehealth solutions. Considering the differences in disposable monthly income, it is more likely that residents in Shenzhen and Hangzhou enjoy the coverage of private insurance. Users with private health insurance coverage are more likely to believe that the coverage of telehealth solutions has an impact on the willingness to use telehealth solutions (whether positive or negative).

The homogeneity of DTC telehealth solutions can lead to indifference in users over factors such as the price, design, privacy risk, and brand and design of telehealth solutions. Currently, telehealth solution providers focus on providing heterogeneous solutions at lower prices. This lowers product profits and deters progress made in data interoperability and the acceptance of telehealth solutions by health care service providers. Some solution providers choose to cut core component configurations to reduce costs, thus failing to guarantee the quality of the solution. Some solution providers have actively marketed their products by offering installments and interest-free loans to attract users. Although promotion and marketing remain important, equipment manufacturers may consider improving the competitiveness of their products by promoting the medical value of their solutions, integrating the solution with the electronic health record system, providing noninvasive monitoring equipment, improving data accuracy, and providing privacy protection.

The findings of the study suggest that users do believe that telehealth solutions can improve health awareness, reduce health risk, mend the gap in community health care services, and improve care continuity, thus having a positive impact on the willingness to use telehealth solutions. Meanwhile, doctors’ suggestions and prescriptions play a role in driving users to choose telehealth solutions over traditional health solutions as well. Elderly (aged 50 years and above) users have a strong demand for self-care and health management anytime and anywhere. Telehealth solutions that are easy to operate, carry, and understand can effectively meet the demand of elderly users for self-care and health management. Moreover, with the global need for qualified clinicians [20], the demand for telehealth solutions is expected to increase.

Compared with traditional health management methods, telehealth solutions offer convenient ways for maintaining health records and health care management at home. However, elderly users do not trust data collected through telehealth solutions. For example, elderly patients with hypertension prefer to go to the doctor to check their blood pressure instead of using the Bluetooth-connected blood pressure monitor at home. With the lack of integration of EHR systems, doctors cannot use the discontinuous data, even if the data collected by telehealth devices are relatively more accurate. There are also technical trust challenges regarding whether the algorithms are trained using accurate data representative of the potential user group. Human trust in the usability of the system, and the regulatory trust issues related to the ethical, legal, and social implications of the use of AI in health care are also important aspects to address [21]. With elderly users, it is extremely important to address the system usability issue of telehealth solutions and build human-level trust.

In the stakeholder interview stage, an interviewed doctor stated that elderly people with chronic diseases are more willing to go to the doctor for blood pressure measurement. The willingness to use telehealth solutions is affected by the users’ health condition. For instance, telehealth solutions can provide users with a large amount of real-time personal health data, such as the heart rate, blood pressure, blood sugar, and other health indexes. For users with chronic diseases, although the data collected by telehealth solutions have certain reference values, doctors either have no access to the data or do not trust the data collected at home. With little or no integration with the health care system, elderly users do not trust the data collected through telehealth solutions.

Conclusions

Many existing telehealth solutions have similar features, prices, and designs. However, doctors often fail to use the data collected from telehealth solutions in making diagnosis and treatment decisions. They often use data collected from telehealth solutions as a reference but rarely base their decision on the daily data trends. Faced with increasing amounts of data from patients, resolving the data interoperability challenges between telehealth solution systems and hospital EHR systems seems more urgent than ever. Further research can focus on data interoperability between the EHR systems and telehealth solutions. The medical value of telehealth solutions can improve if doctors could
interpret data collected from telehealth solutions; furthermore, if doctors could diagnose, provide treatments, and adjust health care management plans based on such data, telehealth solutions can be included in insurance packages, making them more accessible.

There are hurdles to building trust for using telehealth solutions and AI in health care. Future research can also be extended to address such challenges by analyzing how to improve the transparency of algorithms by disclosing the data source and how the algorithms were built.

Owing to the limited scale of the questionnaire study (N=390), this paper only serves as a reference for exploring the implementation of telehealth solutions among elderly users in the next 5 to 10 years. The study was carried out in 4 Chinese cities (Shenzhen, Hangzhou, Wuhan, and Yichang), where the public health system differs in many ways from the health care systems in the United States or Europe. Therefore, the study is limited in scale with a sample size of 390 and remains an exploratory stage study. Further work can be done on the preference for telehealth solutions post-COVID and the changes in business models for telehealth solutions.

Acknowledgments
This work was supported by the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie–ITN Industrial Doctorate (grant 766139) and the National Key R&D Program of China (grant 2017YFE0112000). This article reflects only the author’s views, and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains. The study received support from Dr. Gerd Spekowius from Philips Research, Dr. Lv Ping from University of China Academy of Sciences, Dr. Wei Chen from Fudan University, Dr. Francesca Spigarelli from University of Macerata, and Jerry Yu from Philips China.

Conflicts of Interest
None declared.

Multimedia Appendix 1
English version of the questionnaire.
PDF File (Adobe PDF File), 165 KB - mhealth_v10i1e27272_app1.pdf

Multimedia Appendix 2
Pearson test and Spearman rank correlation test.
DOCX File, 42 KB - mhealth_v10i1e27272_app2.docx

References


3. At 54, China’s average retirement age is too low. The Economist. 2021 Jun 25. URL: https://www.economist.com/china/2021/06/22/chinas-average-retirement-age-is-ridiculously-low-54 [accessed 2021-10-08]


Abbreviations
ANOVA: analysis of variance
DCE: discrete choice experiment
KMO: Kaiser–Meyer–Olkin
PCA: principal component analysis

©Nuoya Chen, Pengqi Liu. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 12.01.2022. 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.
Dose–Response Effects of Patient Engagement on Health Outcomes in an mHealth Intervention: Secondary Analysis of a Randomized Controlled Trial

Yiran Li1*, MMSc; Yan Guo1,2,3*, PhD; Y Alicia Hong4, PhD; Yu Zeng1, MMSc; Aliza Monroe-Wise5, MD, MSc; Chengbo Zeng6,7, PhD; Mengting Zhu3, MS; Hanxi Zhang9, PhD; Jiaying Qiao1, MMSc; Zhimeng Xu1, MMSc; Weiping Cai10, MD; Linghua Li10, MD; Cong Liu10, MSN

1Department of Medical Statistics, School of Public Health, Guangzhou, China
2Sun Yat-sen Center for Global Health, Guangzhou, China
3Department of Population and Quantitative Health Sciences, University of Massachusetts Chan Medical School, Worcester, MA, United States
4Department of Health Administration and Policy, College of Health and Human Services, George Mason University, Fairfax, VA, United States
5Department of Global Health, University of Washington, Seattle, WA, United States
6South Carolina SmartState Center for Healthcare Quality, Arnold School of Public Health, University of South Carolina, Columbia, SC, United States
7Department of Health Promotion, Education, and Behavior, Arnold School of Public Health, University of South Carolina, Columbia, SC, United States
8The Jockey Club School of Public Health and Primary Care, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, Hong Kong
9National Center of AIDS/STD Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China
10Department of Infectious Diseases, Guangzhou Eighth People’s Hospital, Guangzhou Medical University, Guangzhou, China

*these authors contributed equally

Corresponding Author:
Yan Guo, PhD
Department of Medical Statistics
School of Public Health
74 Zhongshan 2nd Road
Guangzhou, 510080
China
Phone: 86 020 87333239
Email: Yan.Guo1@umassmed.edu

Abstract

Background: The dose–response relationship between patient engagement and long-term intervention effects in mobile health (mHealth) interventions are understudied. Studies exploring long-term and potentially changing relationships between patient engagement and health outcomes in mHealth interventions are needed.

Objective: This study aims to examine dose–response relationships between patient engagement and 3 psychosocial outcomes in an mHealth intervention, Run4Love, using repeated measurements of outcomes at baseline and 3, 6, and 9 months.

Methods: This study is a secondary analysis using longitudinal data from the Run4Love trial, a randomized controlled trial with 300 people living with HIV and elevated depressive symptoms to examine the effects of a 3-month mHealth intervention on reducing depressive symptoms and improving quality of life (QOL). We examined the relationships between patient engagement and depressive symptoms, QOL, and perceived stress in the intervention group (N=150) using 4–time-point outcome measurements. Patient engagement was assessed using the completion rate of course assignments and frequency of items completed. Cluster analysis was used to categorize patients into high- and low-engagement groups. Generalized linear mixed effects models were conducted to investigate the dose–response relationships between patient engagement and outcomes.

Results: The cluster analysis identified 2 clusters that were distinctively different from each other. The first cluster comprised 72 participants with good compliance to the intervention, completing an average of 74% (53/72) of intervention items (IQR 0.22). The second cluster comprised 78 participants with low compliance to the intervention, completing an average of 15% (11/72) of intervention items (IQR 0.23). Results of the generalized linear mixed effects models showed that, compared with the low-engagement group, the high-engagement group had a significant reduction in more depressive symptoms ($\beta=-1.93; P=.008$).
and perceived stress ($\beta=-1.72; P<.001$) and an improved QOL ($\beta=2.41; P=.01$) over 9 months. From baseline to 3, 6, and 9 months, the differences in depressive symptoms between the 2 engagement groups were 0.8, 1.6, 2.3, and 3.7 points, respectively, indicating widening between-group differences over time. Similarly, between-group differences in QOL and perceived stress increased over time (group differences in QOL: 0.9, 1.9, 4.7, and 5.1 points, respectively; group differences in the Perceived Stress Scale: 0.9, 1.4, 2.3, and 3.0 points, respectively).

**Conclusions:** This study revealed a positive long-term dose–response relationship between patient engagement and 3 psychosocial outcomes among people living with HIV and elevated depressive symptoms in an mHealth intervention over 9 months using 4 time-point repeat measurement data. The high- and low-engagement groups showed significant and widening differences in depressive symptoms, QOL, and perceived stress at the 3-, 6-, and 9-month follow-ups. Future mHealth interventions should improve patient engagement to achieve long-term and sustained intervention effects.

**Trial Registration:** Chinese Clinical Trial Registry ChiCTR-IPR-17012606; https://www.chictr.org.cn/showproj.aspx?proj=21019

**KEYWORDS**

mHealth; patient engagement; dose–response relationship; long-term effect; generalized linear mixed effects model

**Introduction**

**Background**

Mobile health (mHealth) interventions have gained increasing momentum in delivering easily accessible, patient-centered, individually tailored, and potentially cost-effective programs for a range of psychosocial disorders [1-3]. Previous studies have demonstrated the effectiveness of mHealth interventions in improving psychological outcomes [4,5]. For example, internet-based cognitive behavioral therapy (CBT) has been proven to be effective in treating depressive symptoms [4]. Another study found that an mHealth intervention with self-care strategies reduced depressive symptoms in people living with HIV [5]. However, few studies have explored the impact of patient engagement with mHealth interventions on long-term patient outcomes. Patient engagement is defined as the degree to which a patient adheres to an intervention [6]. Examining the impact of patient engagement on intervention effects beyond the termination of the intervention could help better understand the dose–response relationship in interventions.

Limited mHealth studies have examined the dose–response relationship in program evaluation. Of the few such studies, the most used is the pre–post design with short-term follow-up, typically within 3 months of the intervention [7-10]. To the best of our knowledge, there is only 1 mHealth study that aimed to explore the long-term dose–response relationship between patient engagement and mental health outcomes over 9 months [11]. This 3-month computerized CBT intervention found that different measures of patient engagement, such as number of log-ins, total time spent on the program, and number of visits to a mood diary (1 component of the intervention), were significantly associated with reduced depressive symptoms immediately after the intervention. In the long term, the completion rate of homework assignments was a significant predictor of reduced depressive symptoms at 9 months. However, this study only measured outcomes at the 3- and 9-month follow-ups, and the logistic regression used in the study could not reveal the likely changing dose–response relationship over time. Longitudinal studies with repeated measurements ($\geq 3$) may allow us to examine the changing relationship between patient engagement and intervention effects over time. A better understanding of the potential time-varying relationship between patient engagement and mHealth intervention effects is warranted in the long term.

Long-term dose–response relationships have been examined more thoroughly in face-to-face interventions than in mHealth interventions. In face-to-face CBT interventions, homework assignments are considered indispensable to the effect of psychotherapy. Homework is defined as structured, specific, and therapeutic activities that are routinely completed by the patients between sessions. Homework tasks might include time spent on the program [7,14]. Our previous study examined the associations between patient engagement and intervention outcomes at 3 months in the Run4Love program. We found that a higher completion rate and a greater frequency of completed items were associated with fewer depressive symptoms at 3 months [15]. These automated and multi-dimensional patient engagement data may provide important insights into intervention progress tracking and interpretation of intervention mechanisms. Given that mHealth tools allow for the capture of more multi-faceted factors of patient engagement, more studies to evaluate the impacts of patient engagement on health outcomes in mHealth interventions are needed [14,16,17].
Objectives
This study aims to examine the potential time-varying dose–response relationships between patient engagement and intervention effects in a randomized controlled trial of an mHealth intervention, the Run4Love program, and to fill gaps in the literature. The Run4Love trial aimed to reduce stress and depressive symptoms and to improve quality of life (QOL) among people who lived with HIV and were concurrently experiencing elevated depressive symptoms. We estimated the impact of patient engagement on depressive symptoms (the primary outcome of the intervention), QOL, and perceived stress at the 3-, 6-, and 9-month follow-ups. We hypothesize that better patient engagement in an mHealth intervention could lead to better and sustained health outcomes in the long term.

Methods
Overview
This study is a secondary analysis using data from the Run4Love trial, a parallel randomized controlled trial, to examine the effects of a WeChat (Tencent Holdings Limited)-based intervention on reducing depressive symptoms in people living with HIV and elevated depressive symptoms. The study design and primary results of the Run4Love trial have been published elsewhere [18,19]. The trial was registered in the Chinese Clinical Trial Registry (ChiCTR-IPR-17012606). The study protocol was approved by the Institutional Review Board of Sun Yat-sen University and has been published [19].

Participants and Procedure
A total of 300 people living with HIV and elevated depressive symptoms were recruited from the outpatient department of a large hospital designated for HIV treatment in Guangzhou, the third largest city in China, in 2017. The participants were recruited if they (1) were aged ≥18 years, (2) were HIV-seropositive, (3) had elevated depressive symptoms (measured using a Center for Epidemiologic Studies-Depression [CES-D] score of ≥16), (4) were active users of WeChat, and (5) were willing to provide hair samples (to measure cortisol as a biomarker of chronic stress). Participants were excluded if they were (1) currently on psychiatric or psychological treatment, (2) unable to finish the questionnaire, and (3) unable to engage in the intervention (read or listen to the materials on WeChat or engage in physical exercise because of medical or other reasons). The participants who met the eligibility criteria and were willing to participate completed a baseline survey and were randomized into the intervention or waitlist control group. A total of 150 participants in the intervention group received a 3-month Run4Love intervention and a 3-month booster session; the participants in the control group received a brochure on HIV-related nutrition in addition to usual care for HIV treatment. The run4Love intervention was selected and resent to the participants in the intervention group received up to US $2 as financial incentives based on their course content completion via WeChat accounts on a weekly basis.

Run4Love mHealth Intervention
The Run4Love intervention consisted of two components: adapted cognitive behavioral stress management (CBSM) courses and physical activity promotions [20]. We adapted the evidence-based CBSM courses on stress management and coping skills to the local context and modified them into 65 items in multimedia formats, including short articles, audio clips, and posters. These items were sent via a self-developed, enhanced WeChat platform for 3 months. In the booster session, 7 materials that were read or listened to the most during the intervention were selected and resent to the participants in the next 3 months after the intervention. The articles were approximately 1300 words and took approximately 5 minutes to read; the audio clips were 5-10 minutes in length; and the posters were pictures with motivational captions, which took <30 seconds to read. Physical activity promotions consisted of goal-setting and information on the guidance and benefits of regular exercise. The enhanced WeChat platform had added functions of automatic information sending, course completion tracking, and weekly personalized feedback. Participants in the intervention group received up to US $2 as financial incentives based on their course content completion via WeChat accounts on a weekly basis.

Measurement
Overview
This study collected data on individuals’ sociodemographic characteristics, patient engagement, and psychosocial outcomes, including depressive symptoms, QOL, and perceived stress. Psychosocial outcomes were assessed at baseline and 3-, 6-, and 9-month follow-ups, collected by research staff using electronic questionnaires on a tablet. Data on patient engagement were collected automatically using the enhanced WeChat platform. Sociodemographic characteristics included age, gender, marital status, sexual orientation, and educational level.

Patient Engagement
Patient engagement was assessed through the patient’s completion rate of course assignments and frequency of items completed as these 2 measurements were recommended as reliable measures of patient engagement in mHealth interventions targeting psychosocial outcomes [14] and proven in our previous study [15]. The results from our previous study revealed that these 2 measurements were significantly associated with reduced depressive symptoms at 3 months in the Run4Love intervention. Therefore, we grouped the participants according to these 2 measurements of patient engagement using cluster analysis. A total of 72 intervention items in the form of short articles, audio clips, and posters were delivered to the participants, of which 65 (90%) were sent during the 3-month intervention and 7 (10%) were resent 0 to 3 months after the intervention as a booster. The completion rate was calculated as the percentage of items completed out of 72 by a participant. Items that were clicked by the participants were regarded as completed. The frequency of items completed referred to the total number of times the items were read or listened to by a participant during the 3-month intervention and booster session 0-3 months after the intervention. For example, if a participant read 1 item sent during the 3-month intervention twice and the same item in the booster material twice, the frequency of items completed was counted as 4. As the participants were encouraged to practice the skills for stress management from the CBSM courses repeatedly, the frequency of items completed was used to capture the repetition aspect of patient engagement.
Both completion rate and frequency of items completed were automatically tracked by the enhanced mHealth platform. Good reliability of the composite measurement of patient engagement was shown in the study, and the Cronbach $\alpha$ was .97.

**Depressive Symptoms**

Depressive symptoms were measured using the CES-D scale, one of the most widely used self-reported questionnaires on depressive symptoms in China [21-23]. The CES-D scale consists of 20 items, such as I felt depressed and I did not feel like eating; my appetite was poor, and each item is rated on a 4-point Likert scale ranging from 0 (rarely or none of the time) to 3 (most or all of the time). The CES-D scores range from 0 to 60, with higher scores indicating a higher level of depressive symptoms and 16 being the cut-off point for possible clinical depressive symptoms [24]. Scores ranging from 16 to 20, 21 to 25, and 26 to 60 are considered mild, moderate, and severe depressive symptoms, respectively [25]. Good reliability of the CES-D score was shown in the study, and the Cronbach $\alpha$ at baseline and the 3-, 6-, and 9-month follow-ups was .77, .76,.84, and .83, respectively.

**QOL Measurement**

QOL was measured using the World Health Organization Quality of Life HIV short version (WHOQOL-HIV BREF), with 31 items assessing 6 domains: physical, psychological, level of independence, social relationships, environment, and beliefs [26]. Each domain comprises items rated on a 5-point Likert scale. The WHOQOL-HIV BREF scores range from 24 to 120, with higher scores indicating better QOL. The WHOQOL-HIV BREF has been widely used in the Chinese population with HIV and has shown good validity and reliability [27-29]. In this study, the Cronbach $\alpha$ for the WHOQOL-HIV BREF at baseline and the 3-, 6-, and 9-month follow-ups was .84, .91, .94, and .94, respectively.

**Perceived Stress**

Perceived stress was assessed using the 10-item Chinese version of the Perceived Stress Scale (PSS-10) [30]. The PSS-10 is the most widely used validated instrument for assessing the perception of stress in Chinese population [31-33]. It assesses the participants’ feelings and thoughts in the previous month (eg, How often have you been upset because of something that happened unexpectedly?). The PSS-10 scores range from 0 to 40, with higher scores indicating higher levels of perceived stress. Scores ranging from 0 to 13, 14 to 26, and 27 to 40 are considered low, moderate, and high levels of perceived stress, respectively [34]. In this study, the Cronbach $\alpha$ for the PSS-10 at baseline and the 3-, 6-, and 9-month follow-ups was .67, .65, .69, and .65, respectively.

**Statistical Analysis**

First, descriptive statistics of demographic characteristics, baseline depressive symptoms, QOL, and perceived stress were presented. Continuous variables with normal distribution were described using mean and SD, and those with skewed distribution were described using median and IQR. Categorical variables were described using numbers and percentages. Second, hierarchical clustering was used to classify participants into different groups based on the 2 metrics of patient engagement. Cluster analysis is an exploratory classification technique to group participants into different categories based on their similarities in specific metrics. This statistical method helps identify different engagement groups taking different dimensions into account. Hierarchical clustering iteratively merges smaller clusters into larger clusters. The hierarchical clustering procedure was as follows:

1. A similarity distance matrix was constructed by calculating the pairwise distance between different observations. Each observation was assigned to an individual cluster; therefore, each observation represented 1 cluster.
2. The 2 clusters $r$ and $s$ with a minimum distance from each other were identified.
3. Clusters $r$ and $s$ were merged, and $r$ was replaced with the new cluster. Cluster $s$ was deleted, and distances between the new cluster and each of the old clusters were computed.
4. Steps 2 and 3 were repeated until the total number of clusters was 2 [35].

Through clustering, we categorized participants into high- and low-engagement groups based on both measurements of completion rate and frequency of items completed, which could be used to evaluate the effects of different levels of patient engagement on health outcomes. To verify the results of the cluster analysis, we used Wilcoxon rank-sum tests to compare patient engagement between the high- and low-engagement groups along the 2 metrics. In addition, Wilcoxon rank-sum tests were used to examine whether patient outcomes were balanced between the 2 engagement groups at baseline.

Finally, generalized linear mixed effects models (GLMMs) with fixed effects of the engagement groups and time and including time as a random effect were conducted to estimate the trajectories of patient outcomes of depressive symptoms, QOL, and perceived stress, by the 2 engagement groups over time. The random effects of time represented interindividual varying time trends. This allowed for the estimation of variance in the outcomes within and among these time groups. The GLMM allows for the simultaneous analysis of repeated measures in a longitudinal design, thus providing a more accurate estimation of changes in outcomes over time. It also allows for the inclusion of cases with missing data [36,37], making it well-suited for longitudinal data that is likely to have missing values, such as in this study.

In total, 3 GLMMs were conducted to evaluate the relationships between levels of patient engagement and 3 health outcomes over time, adjusting for time and baseline characteristics. The dependent variables were depressive symptoms, QOL, and perceived stress measured repeatedly at baseline and 3, 6, and 9 months, whereas the independent variables were the patient engagement groups (high- and low-engagement group, the latter as reference) and the 4 time points (baseline and 3, 6, and 9 months, with baseline as reference). Baseline characteristics were included in the GLMMs as control variables, including age, gender, marital status, sexual orientation, and education. Only statistically significant characteristics were retained in the final models. The model estimate of the coefficient for
engagement groups represented how the trajectories of health outcomes differed between the 2 groups. For example, in the GLMM with CES-D as the dependent variable, a significant negative coefficient suggested that the high-engagement group had a reduction in more depressive symptoms over 9 months than the low-engagement group. Statistical significance was defined as $P<.05$. All analyses were conducted using SPSS (version 25; IBM).

**Results**

**Baseline Characteristics**

The Run4Love trial recruited 300 participants, and the data from the 150 participants in the intervention group were used.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>28.0 (5.8)</td>
</tr>
<tr>
<td>Gender, n (%)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>142 (94.7)</td>
</tr>
<tr>
<td>Female</td>
<td>8 (5.3)</td>
</tr>
<tr>
<td>Sexual orientation, n (%)</td>
<td></td>
</tr>
<tr>
<td>Heterosexual</td>
<td>20 (13.3)</td>
</tr>
<tr>
<td>Homosexual, bisexual, or uncertain</td>
<td>130 (86.7)</td>
</tr>
<tr>
<td>Education, n (%)</td>
<td></td>
</tr>
<tr>
<td>High school or lower</td>
<td>52 (34.7)</td>
</tr>
<tr>
<td>More than high school</td>
<td>98 (65.3)</td>
</tr>
<tr>
<td>Marital status, n (%)</td>
<td></td>
</tr>
<tr>
<td>Single, divorced, or widowed</td>
<td>132 (88)</td>
</tr>
<tr>
<td>Married</td>
<td>18 (12)</td>
</tr>
<tr>
<td>CES-D$^a$, mean (SD)</td>
<td>23.9 (6.4)</td>
</tr>
<tr>
<td>QOL$^b$, mean (SD)</td>
<td>77.4 (9.0)</td>
</tr>
<tr>
<td>PSS-10$^c$, mean (SD)</td>
<td>20.0 (4.4)</td>
</tr>
</tbody>
</table>

$^a$CES-D: Center for Epidemiologic Studies-Depression.

$^b$QOL: quality of life.

$^c$PSS-10: Perceived Stress Scale.

**Cluster Analysis**

The cluster analysis identified 2 clusters that were distinctively different from each other (Table 2). The first cluster consisted of 72 participants with good compliance to the intervention program, completing an average of 74% (53/72) of intervention items (IQR 0.22) and 82 items (IQR 35.50) when accounting for repeated visits. Specifically, the high-engagement group completed an average of 77% (50/65) of intervention items (IQR 0.23) in the 3-month intervention and 43% (3/7) of items (IQR 0.57) in the booster session. The second cluster consisted of 78 participants with low compliance to the intervention program, completing an average of 15% (11/72) of intervention items (IQR 0.23) in the intervention program and 15 items (IQR 23.25) accounting for repeated visits. Specifically, the low-engagement group completed an average of 17% (11/65) of intervention items (IQR 0.25) in the 3-month intervention and 0 (IQR 1.00) in the booster session. The results of the Wilcoxon rank-sum tests confirmed the significant differences between the 2 cluster groups in both engagement measurements (completion rate and frequency of items completed; $P<.001$), with the high-engagement group having significantly better compliance in both engagement measurements than the low-engagement group, verifying the 2 distinct groups categorized through the cluster analysis. There were no significant group differences in depressive symptoms, QOL, and perceived stress at baseline (CES-D: $P=.54$; QOL: $P=.45$; and PSS-10: $P=.25$), indicating a balance in the outcomes between the 2 engagement groups at baseline.
Table 2. Differences in the engagement measurements between the high- and low-engagement groups (N=150).

<table>
<thead>
<tr>
<th>Engagement metrics</th>
<th>High-engagement group (n=72), median (IQR)</th>
<th>Low-engagement group (n=78), median (IQR)</th>
<th>Wilcoxon rank-sum test</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completion rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-month intervention</td>
<td>0.77 (0.23)</td>
<td>0.17 (0.25)</td>
<td>3091.0</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Booster session</td>
<td>0.43 (0.57)</td>
<td>0.00 (0.14)</td>
<td>3877.5</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Total</td>
<td>0.74 (0.22)</td>
<td>0.15 (0.23)</td>
<td>3091.5</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Frequency of items completed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-month intervention</td>
<td>80 (30.50)</td>
<td>13 (21.25)</td>
<td>3084.0</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Booster session</td>
<td>3 (4.75)</td>
<td>0 (1.00)</td>
<td>3955.5</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Total</td>
<td>82 (35.50)</td>
<td>15 (23.25)</td>
<td>3084.5</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Effects of Patient Engagement on Health Outcomes Over Time

The trajectories of the 3 outcomes at baseline and 3-, 6-, and 9-month follow-ups are shown in Figure 1, and GLMM results examining the effects of patient engagement on the 3 outcomes are presented in Table 3. Of the 150 participants in the intervention group, 139 (92.7%), 132 (88%), and 133 (88.7%) participants completed the follow-up surveys at 3, 6, and 9 months, respectively. These participants were randomly missing as there were no differences in demographic characteristics and outcomes at baseline between those who completed the follow-up surveys and those who did not. All 3 health outcomes in both groups significantly improved at 3 months immediately after the 3-month intervention (Figure 1). The results of the GLMM (Table 3) showed that the β coefficients of the 3 intervention outcomes were all statistically significant, indicating significant between-group differences in these outcomes over time.

Figure 1. Trajectories of depressive symptoms, quality of life (QOL), and perceived stress over time in high- and low-engagement groups. CES-D: Center for Epidemiologic Studies-Depression; PSS: Perceived Stress Scale.
Table 3. Effects of patient engagement on intervention outcomes at the 3-, 6-, and 9-month follow-ups: results from generalized linear mixed effects models.a

<table>
<thead>
<tr>
<th>Variables</th>
<th>β coefficient (SE; 95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CES-D</strong>b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>19.53 (1.86; 15.87 to 23.18)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Engagement group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High vs low</td>
<td>−1.93 (0.72; −3.34 to −0.51)</td>
<td>.008</td>
</tr>
<tr>
<td><strong>Follow-up</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-month follow-up vs baseline</td>
<td>−6.02 (0.94; −7.87 to −4.16)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>6-month follow-up vs baseline</td>
<td>−6.11 (0.99; −8.05 to −4.17)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>9-month follow-up vs baseline</td>
<td>−5.78 (1.08; −7.91 to −3.65)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Age (years)</td>
<td>.16 (0.06; 0.03 to 0.28)</td>
<td>.01</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school or lower vs more than high school</td>
<td>2.83 (0.75; 1.35 to 4.31)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>QOL</strong>c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>78.55 (0.90; 76.77 to 80.33)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Engagement group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High vs low</td>
<td>2.41 (0.93; 0.59 to 4.23)</td>
<td>.01</td>
</tr>
<tr>
<td><strong>Follow-up</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-month follow-up vs baseline</td>
<td>5.05 (1.22; 2.65 to 7.44)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>6-month follow-up vs baseline</td>
<td>6.01 (1.27; 3.51 to 8.50)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>9-month follow-up vs baseline</td>
<td>5.74 (1.34; 3.10 to 8.38)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school or lower vs more than high school</td>
<td>−6.62 (0.97; −8.53 to −4.71)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>PSS-10</strong>d</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>20.28 (0.45; 19.39 to 21.17)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Engagement group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High vs low</td>
<td>−1.72 (0.45; −2.61 to −0.82)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Follow-up</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-month follow-up vs baseline</td>
<td>−4.25 (0.60; −5.44 to −3.07)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>6-month follow-up vs baseline</td>
<td>−3.42 (0.61; −4.63 to −2.22)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>9-month follow-up vs baseline</td>
<td>−3.78 (0.65; −5.05 to −2.50)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school or lower vs more than high school</td>
<td>1.51 (0.48; 0.57 to 2.44)</td>
<td>.002</td>
</tr>
</tbody>
</table>

aGeneralized linear mixed effects models included all participants in the intervention group at all time points. Analyses were adjusted for individual characteristics, including age, gender, marital status, sexual orientation, and education at baseline, and those that did not show significant effects were removed from the final models. Educational level and age remained in the final models.

bCES-D: Center for Epidemiologic Studies-Depression.

cQOL: quality of life.

dPSS-10: Perceived Stress Scale.

The average group difference in CES-D scores between the high- and low-engagement groups was 1.93 (β=−1.93; P=.008) points over the 9 months, with the high-engagement group having lower levels of depressive symptoms than the low-engagement group. After the initial drastic decreases in both groups during the 3-month intervention (high-engagement group from 23.5 to 17.1 points; low-engagement group from 24.3 to 18.7 points), group differences in depressive symptoms increased in 0 to 6 months after the intervention. As shown in the first graph in Figure 1, there was a slowly rising trend in the CES-D scores at 3, 6, and 9 months (18.7, 18.9, and 20.0, respectively), suggesting some rebound in depressive symptoms.
in the low-engagement group. In contrast, CES-D scores continued to decrease in the high-engagement group at 3, 6, and 9 months (17.1 to 16.6 and 16.3, respectively). Thus, between-group differences and standard effect sizes (Cohen $d$) in depressive symptoms (CES-D scores) increased over time, with 1.6, 2.3, and 3.7 points and 0.17, 0.23, and 0.34 points at 3, 6, and 9 months, respectively. The high-engagement group experienced a 14%, 4%, and 11% decrease in the rate of possible clinical depressive symptoms compared with the low-engagement group at 3, 6, and 9 months (50% vs 64%, 50% vs 54%, and 44% vs 55%, respectively).

Similar trends were also observed in the other 2 outcome measures. The average group differences in the QOL and perceived stress scores between the high- and low-engagement group were 2.41 ($\beta=2.41; P=.01$) and 1.72 ($\beta=1.72; P<.001$) points over the 9 months, with the high-engagement group achieving better health outcomes across both measures. After the 3-month intervention, both outcome measures somewhat rebounded in the low-engagement group at 3, 6, and 9 months (QOL: 81.6, 81.0, and 80.5, respectively; PSS-10: 16.4, 17.7, and 17.7, respectively). In contrast, the high-engagement group had either continued improvement or reduced rebound effects at 3, 6, and 9 months in both outcome measures (QOL: 83.5, 85.7, and 85.6, respectively; PSS-10: 15.0, 15.4, and 14.7, respectively). Therefore, similar trends of widening between-group differences in QOL and perceived stress were observed over time at 3, 6, and 9 months (between-group differences in QOL: 1.9, 4.7, and 5.1 points, respectively; between-group differences in PSS-10: 1.4, 2.3, and 3.0 points, respectively). Standard effect sizes (Cohen $d$) in QOL (QOL scores) and perceived stress (PSS-10 scores) both increased over time, with 0.16, 0.37, and 0.39 in QOL and 0.25, 0.39, and 0.51 in PSS-10 at 3, 6, and 9 months, respectively.

Covariates including age, gender, marital status, sexual orientation, and educational level at baseline were adjusted in the 3 GLMMs, and only statistically significant variables were retained in the final models. Educational level was significantly associated with the 3 outcomes, and age was only significantly associated with depressive symptoms. Specifically, compared with those with higher education, participants with lower educational levels (high school or lower) had poorer health outcomes (CES-D: $\beta=2.83, P<.001$; QOL: $\beta=6.62, P<.001$; PSS-10: $\beta=1.51, P=.002$). Compared with younger participants, older participants were more likely to report higher levels of depressive symptoms ($\beta=.16; P=.01$).

**Discussion**

**Principal Findings**

This study was among the first efforts to explore the potential time-varying dose–response relationships between patient engagement and various health outcomes over a span of 9 months using 4-time-point measurement data from the Run4Love mHealth intervention. The main finding was that patient engagement had a positive impact on the health outcomes, including depressive symptoms, QOL, and perceived stress, and such impacts were sustained over 9 months after the baseline. In addition, the dose–response relationship was not only sustained in the long term but also increased over time, as there were widening differences in health outcomes between the high- and low-engagement groups.

Both the high- and low-engagement groups benefited from the intervention, but the high-engagement group benefited more from the intervention consistently as the differences in health outcomes between the 2 groups became more pronounced over time. Such sustained and potentially increasing dose–response relationship in the long term has not been reported in previous mHealth studies. The increased effect sizes in depressive symptoms and QOL at 3, 6, and 9 months and increased effect sizes in perceived stress at 3 and 6 months were small to medium. The effect size of perceived stress at 9 months was medium [38]. In the low-engagement group, the intervention effects on health outcomes from 3 to 9 months are consistent with the findings of previous studies reporting rebound effects in mHealth or CBT interventions [39,40]. In contrast, the high-engagement group did not show similar trends; instead, it showed either sustained improvements or fewer rebound effects in all the health outcomes 0 to 6 months after the intervention, resulting in widening differences in these health outcomes between the 2 engagement groups.

Existing literature suggests that intervention effects tend to decrease or diminish over time after the intervention, with some demonstrating rebound effects [39,40]. For example, a face-to-face study found that women with breast cancer in the CBSM intervention group experienced a significant decrease in depressive symptoms immediately after treatment, but the level of depressive symptoms rebounded to baseline at the 1-month follow-up after the intervention [40].

The reasons for the sustained and widening dose–response relationship in this study are many; evidence-based interventions with rigorous design and good implementation are more likely to have sustained dose–response effects. What is missing in the literature on mHealth interventions is whether the momentum continues over time, for example, in 6- or 9-month follow-ups, and whether engagement level plays a role in this momentum. This study adds new evidence to this gap in the literature. Additional research to understand what factors predict patient engagement is also needed.

Our findings also revealed that education and age were important individual characteristics associated with the effects of the intervention, with participants with lower levels of education and of older age having poorer health outcomes. These findings are consistent with the literature [41-43]. To close the digital gap and bridge health disparities, mHealth interventions should be tailored to the needs of these more vulnerable groups, such as older people and those less educated. For example, mHealth interventions should be designed with easy-to-navigate interfaces, bigger fonts, and plain language with engaging multimedia such as pictures, audios, and videos [44].

**Policy Implications**

Given the sustained positive impacts of patient engagement on health outcomes found in this study, it is critical to improve patients’ intervention adherence and engagement in both the intervention and subsequent booster sessions in mHealth.
interventions. There are some effective ways to improve patient engagement as suggested in the literature and evidenced in our Run4Love trial. First, the intervention content needs to be culturally tailored and personalized, which requires formative research and a pilot study. The Run4Love intervention was developed based on extensive formative research and a pilot study [19,45]. We culturally adapted theory-guided and evidence-based CBSM courses, which have been proven effective in relieving depressive symptoms and improving other health outcomes in people who live with HIV [19,45]. Second, the program needs to have superior usability and user experience. Previous studies have shown that perceived usefulness and user-friendly experience are critical for improving mHealth engagement [46,47]. After many rounds of in-depth interviews with patients and the iterative development process, we designed and tailored the intervention platform and formats to meet their needs, such as delivering more appealing multimedia items [45]. With rigorous design and implementation, the participants reported high levels of satisfaction (92%-97%) at all 3 follow-ups in the Run4Love trial [18].

During the intervention, another way to improve patient engagement is to provide timely and personalized feedback. The enhanced WeChat platform automatically sends weekly feedback on the completion status to each participant. In addition to automatic weekly feedback, the Run4Love program also consisted of 5 phone calls made by the research staff at 1 week and 1, 2, 5, and 8 months to address technical challenges and motivate their participation. In addition, the backend platforms of mHealth interventions and wearable devices allow for the collection of passive data on various dimensions of patient engagement, such as physical activities, sleep hours and quality, log-in times, and time and duration of reading, listening to, or watching the intervention items [48-50]. Taking advantage of the easy-to-track user engagement data available in mHealth interventions is critical for process monitoring and quality control of the trials.

**Limitations**

There are several limitations to this study. First, the participants in this study were mostly from urban areas, young, and well-educated, and most were nonheterosexual men. Therefore, generalization of the results should be applied with caution. Second, measurement biases may exist in patient engagement. This study only measured patient engagement in CBSM courses but not in physical activity promotions as patient engagement data were not recorded or available in physical activities. However, from the qualitative interviews, we found that most of the engaged participants had a higher engagement in both CBSM courses and physical activities. Therefore, a long-term dose–response relationship between patient engagement in physical activities and intervention outcomes may exist, which needs to be further explored in future research. Moreover, the intervention items were considered as completed when clicked; therefore, we were not able to verify the actual completion or quality of completion. Nevertheless, the patient engagement metrics in this study served as a reliable measurement to assess the dose–response relationship, and the effect of such potential overestimation of patient engagement might only have diluted the observed dose–response relationship [15,51]. Finally, patient engagement in different formats of the intervention content, including short articles, audio clips, and posters, was not differentiated as this was beyond the objectives of this study. Future studies could further explore better measurements of patient engagement in terms of different intervention components and types of multimedia materials in mHealth interventions.

**Conclusions**

In conclusion, this study revealed a positive long-term dose–response relationship between patient engagement and 3 psychosocial outcomes in an mHealth intervention using 4–time-point measurement data over 9 months. High- and low-engagement groups showed significant and widening differences in depressive symptoms, QOL, and perceived stress at the 3-, 6-, and 9-month follow-ups in the Run4Love trial. Future mHealth interventions should improve patient engagement to achieve long-term and sustained intervention effects.

**Acknowledgments**

This study was supported by a grant from the National Natural Science Foundation of China (grant 71573290) and the China Medical Board (grant 17-271). The funder had no role in the study design, data collection and analysis, or writing of the manuscript.

**Authors’ Contributions**

YL analyzed the data and drafted the manuscript. YG and YAH contributed to obtaining funding, the study design, and manuscript revision. YZ helped with the study concept. AMW helped with manuscript revision. CZ, MZ, HZ, JQ, and ZX contributed to the clinical trial and data acquisition. WC, LL, and CL provided administrative, technical, and material support for the clinical trial.

**Conflicts of Interest**

None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
References


Abbreviations

- CBSM: cognitive behavioral stress management
- CBT: cognitive behavioral therapy
- CES-D: Center for Epidemiologic Studies-Depression
- GLMM: generalized linear mixed effects model
- mHealth: mobile health
- PSS-10: Perceived Stress Scale
- QOL: quality of life
- WHOQOL-HIV BREF: World Health Organization Quality of Life HIV short version

©Yiran Li, Yan Guo, Y Alicia Hong, Yu Zeng, Aliza Monroe-Wise, Chengbo Zeng, Mengting Zhu, Hanxi Zhang, Jiaying Qiao, Zhimeng Xu, Weiping Cai, Linghua Li, Cong Liu. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 04.01.2022. 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

Effect of an Integrative Mobile Health Intervention in Patients With Hypertension and Diabetes: Crossover Study

Sang Woo Oh¹, MD, PhD; Kyoung-Kon Kim², MD, PhD; Sung Soo Kim³, MD, PhD; Su Kyung Park¹, MSDH; Sangshin Park⁴, DVM, PhD

¹Department of Family Medicine, Dongguk University Ilsan Hospital, Dongguk University College of Medicine, Gyeonggi-do, Republic of Korea
²Department of Family Medicine, Gachon University Gil Medical Center, Gachon University College of Medicine, Incheon, Republic of Korea
³Department of Family Medicine, Chungnam National University Hospital, Chungnam National University College of Medicine, Daejeon, Republic of Korea
⁴Graduate School of Urban Public Health & Department of Urban Big Data Convergence, University of Seoul, Seoul, Republic of Korea

Corresponding Author:
Sang Woo Oh, MD, PhD
Department of Family Medicine
Dongguk University Ilsan Hospital
Dongguk University College of Medicine
27, Dongguk-ro, Ilsandong-gu
Goyang-si
Gyeonggi-do, 10326
Republic of Korea
Phone: 82 31 961 7490
Email: osw6021@naver.com

Abstract

Background: Obesity, hypertension, and type 2 diabetes mellitus (T2DM) are worldwide epidemics that inflict burdens on both public health and health care costs. Self-management plays an important role in the proper management of these 3 chronic diseases, and in this context, mobile health (mHealth) can be a cost-effective self-management tool.

Objective: The aim of this pilot study is to evaluate the effects of an integrative mHealth approach for obesity, hypertension, and T2DM on body fat, blood pressure, and blood glucose levels and demonstrate the clinical outcomes. The participants were patients aged 40 to 70 years who were treated for T2DM (hemoglobin A1c [HbA1c] above 6.0%) without insulin or hypertension and obesity, controlled with pharmacotherapy.

Methods: This pilot study was performed using a controlled, randomized, 3-month, 2-period crossover design. A total of 37 participants were recruited from 2 university hospitals in South Korea. Integrative mHealth comprised 4 parts: self-measuring home devices for monitoring blood glucose and blood pressure; 2 smartphone apps, where one gathered lifestyle data, giving them feedback with health information, and the other provided drug information and reminders of the medication schedule; unmanned kiosks for official measurement of blood pressure and body composition; and web-based access to participants’ health information.

Results: Data from the 32 participants were analyzed. Their mean HbA1c level was 7.5% (SD 0.8, ranging from 6.1% to 9.4%). Approximately 38% (12/32) of the participants had hypertension. BMIs of all participants except 1 were >23 kg/m². The input rates of food intake and exercise to the smartphone app were very low (24.9% and 5.3%, respectively). On the contrary, the input rate of medicine intake was high (84.0%). Moreover, there was no significant difference in the input rate of taking medicine irrespective of whether the mHealth period was before or after the conventional treatment period (80.3% and 87.3%, respectively; P=.06). Among the 3 input functions of food intake, exercise, and medicine intake in smartphone apps, the input of medicine intake was a more helpful, easier to use, and better-designed function than the others. There were no significant differences in changes in body weight (-0.519 kg vs 0 kg), BMI (-0.133 kg/m² vs -0.167 kg/m²), body composition (body fat -0.255% vs 0.172%), blood pressure (systolic -0.226 mm Hg vs -2.839 mm Hg), and HbA1c (-0.269% vs -0.009%) between the integrative mHealth and conventional treatment groups. However, in proportion to the elevation in the input rate of taking medicine, body fat mass (P=.04) and HbA1c (P=.03) were lower in the integrative mHealth group.

https://mhealth.jmir.org/2022/1/e27192

JMIR Mhealth Uhealth 2022 | vol. 10 | iss. 1 | e27192 | p.68

(page number not for citation purposes)
Conclusions: Although smartphone apps can influence body fat and blood glucose levels, they have failed to show clinical improvement. A higher input rate of taking medicine was related to significantly lower body fat mass and HbA1c levels.

(JMIR Mhealth Uhealth 2022;10(1):e27192) doi:10.2196/27192

KEYWORDS
diabetes mellitus type 2; obesity; hypertension; mHealth; mobile phone

Introduction

Background

Obesity is the established main cause of hypertension and type 2 diabetes mellitus (T2DM) and can lead to the development of coronary vascular disease; furthermore, these 3 illnesses have attained the status of global epidemics recently [1-3]. Consequently, they inflict a huge economic burden on public health systems worldwide. Obesity was estimated to account for 0.7% to 2.8% of the total health care expenditure (HCE), and people with obesity had health care costs that were approximately 30% greater than those with normal weight [4]. The incremental medical expenditure ratios for people in Korea with BMIs of 30 kg/m² to 34.99 kg/m² and >35 kg/m² were 34.3% and 38.4%, respectively, as compared that of with people with BMIs of 18.5 kg/m² to 22.99 kg/m² from 2002 to 2013 [5]. The US national health care spending associated with hypertension was estimated to be approximately US $131 billion, averaged over 12 years from 2003 to 2014 [6]. In 2017, the total estimated cost of diagnosed diabetes in the United States was US $327 billion, including medical costs and lost productivity, and provision of care for people with diagnosed diabetes accounted for a quarter of health care costs [7]. Without appropriate management of obesity, hypertension, and T2DM, patients experience disastrous complications, and societies are troubled with HCEs and disease-related productivity losses.

Self-management is crucial for the proper management of the 3 chronic diseases. A structured lifestyle intervention program comprising a healthy diet, physical activity, and behavioral interventions is essential for the treatment of obesity [8]. Lifestyle management and self-management with self-monitoring are also important in the treatment of hypertension and T2DM [9-11]. However, in face-to-face outpatient consultations, health care providers lack time to deliver information and skills for self-management to patients and motivate them to change their lifestyles.

In this context, mobile health (mHealth) can be a cost-effective tool for self-management in the treatment of chronic diseases. The Global Observatory for eHealth defined mHealth as a medical and public health practice supported by mobile devices [12]. It is useful as it can (1) enhance drug adherence through reminders, (2) facilitate self-monitoring coupled with wireless medical peripheral devices, and (3) provide tailored practical information.

A considerable number of clinical trials have been executed to inspect the usefulness of mHealth interventions in the treatment of obesity, hypertension, and T2DM [13,14]. The reviews and meta-analyses of these trials indicate that although mHealth interventions are likely to promote weight loss, lower hemoglobin A1c (HbA1c), and reduce blood pressure, the individual results are mixed [13,14]. In Korea, several groups have reported the clinical usefulness of smartphone-based apps in T2DM patients; however, their glucose-lowering effect is not clear [15-17]. Although numerous trials have investigated the efficacy of mHealth in the management of obesity, hypertension, and T2DM, they have explored its effect with respect to the 3 diseases, separately or in patients with obesity and hypertension or in patients with hypertension and T2DM. As hypertension and T2DM are comorbidities of obesity, and the 3 diseases are important risk factors of coronary vascular disease, integrative lifestyle approaches for the 3 diseases are more appropriate. They should include the feedback system of diet and exercise, medication assistance, and self-monitoring of blood pressure, blood glucose, and body composition.

Objectives

The aim of this 6-month crossover pilot study is to evaluate the clinical effects of integrative mHealth supported by self-monitoring home devices among patients with T2DM or hypertension and obesity. The integrative mHealth used in this pilot study provides a platform to link the out-of-hospital self-monitoring results of diet, exercise, blood pressure, blood glucose, and body composition with web servers for data storage and web portals for the patient and their physician’s data access. Embedded apps in patients’ smartphones are LIBIT (Huraypositive Co) for recording diet and exercise, connecting self-measuring home devices to web servers, and providing feedback and health information to patients; and Mediram (GST Korea) for medication assistance. These smartphone apps have been newly developed for this project.

Methods

Study Participants

The pilot study was conducted with adults aged 40 to 70 years who were treated for T2DM (without insulin) or hypertension in the departments of family medicine and endocrinology at 2 university hospitals and were in stable status for at least the past 4 months. Recent HbA1c levels of participants measured in <4 months were >6.0%. To use the Bluetooth-enabled self-measuring home devices and smartphone apps developed for this pilot study, the participants should have had and been able to use Android smartphones with OS version 4.3 (jellybean) or later. Recruitment was conducted between October 2018 and February 2020 in Incheon and Daejeon, which are 2 metropolitan cities in South Korea. The exclusion criteria comprised a history of malignant diseases, coronary artery obstructive disease, stroke, organ transplantation, drug abuse and alcohol dependence, disability or respiratory disease limiting
exercise, and hospitalization in the past 6 months with major medical conditions.

All participants were informed of the aim and process of the pilot study during the interviews and were requested for their consent to join this study. The study procedures were performed only with participants who provided informed consent. This pilot study was approved by the institutional review boards of Gachon University Gil Medical Center and Chungnam National University Hospital, and the pilot study was performed in accordance with the Declaration of Helsinki and the guidelines of Good Clinical Practice. Deidentified and anonymized data were used in the analyses.

Study Design, Devices, and Smartphone Apps

This pilot study was performed using a controlled, randomized, 3-month, 2-period crossover design to test the efficacy of integrative mHealth coupled with self-monitoring home devices and smartphone apps as opposed to the conventional treatment (CON) of T2DM with or without hypertension. Figure 1 demonstrates the pilot study design. The recruited participants were randomly assigned to 2 groups using computer-generated random numbers: 1 group started with the integrative mHealth service period and switched to the CON period (mHealth-CON group), whereas the other group started with the CON period and switched to the mHealth period (CON-mHealth group). There was no washout period. Measurements of body weight, body composition, blood pressure, and HbA1c were taken at the start of the first treatment period, during treatment transition, and at the end of the second treatment period. A survey on smartphone apps was conducted at the end of the pilot study.

Integrative mHealth supported the participants and physicians through the following 4 components: (1) self-measuring devices; (2) smartphone apps that gathered and transferred data on the participant’s lifestyle and provided feedback, health and drug information, and reminders of their medication schedule; (3) unmanned kiosks for the official measurement of blood pressure and body composition; and (4) web-based access to participants’ health information through which physicians could review participants’ health data at a glance. The entire architecture of the information transmission in this pilot study is shown in Figure 2. For systematic collection and administration of health information data, this pilot study emphasized data security by applying 5 systems: (1) section encoding via secure socket layer (SSL) or transport layer security, (2) encoding critical information, (3) controlling the users’ and administrators’ accessibility to data, (4) restricting the collection of personal identification information, and (5) agreeing to collect and use personal identification information. Information with a high risk of data loss, such as passwords, was saved using the unilateral encoding system of Secure Hash Algorithm 256 in the health information service system. Integrative mHealth was available only for the mHealth period. The devices and smartphone apps were supplied at the commencement of the mHealth period and retrieved at the end of it.
A total of 2 Bluetooth-enabled devices, the blood pressure monitor HEM-9200T (Omron) and the blood glucose monitor CareSense N Premier BLE (i-sense), were used as self-measuring home devices in this pilot study (Multimedia Appendix 1). A total of 2 unmanned digital kiosks (GST Korea) were established at both hospitals for the self-measurement of blood pressure and body composition (Multimedia Appendix 2). The kiosks were equipped with a user screen, a radio-frequency identification card reader, a bioelectric impedance analyzer SC-330 (Tanita), and an automatic upper arm sphygmomanometer BP-210 (Accuniq). For using the kiosks, a radio-frequency identification card was supplied to each participant.

The readings of blood pressure and glucose levels that were measured at home were transmitted to the patient’s Android smartphone app LIBIT through a Bluetooth connection and then transferred to the main server using cellular data. The data on body composition and blood pressure measured at the kiosks were also linked to the main server through http secure based on certificate verification; http secure sent encoded information of the clients to the server using the security protocol of SSL. SSL or transport layer security operates in the same way as a virtual private network, sending security data to the server via a virtual tunnel.

A total of 2 Android apps, LIBIT and Mediram, were developed to gather participants’ lifestyle data, provide feedback and health and drug information, and enhance their adherence to medication in this pilot study. The LIBIT app comprised 4 functions: nutrition care, exercise assessment, transmission of self-monitored blood pressure and glucose level, and health monitoring (Multimedia Appendix 3). The nutrition care component of LIBIT calculated the suggested total calorie and macronutrient ratios of each participant. Users could record their food intake through the smartphone keypad or their voice using embedded voice recognition technology. Analyzing the food intake records, LIBIT estimated the intake of 14 nutrients (total calories, carbohydrates, proteins, fats, calcium, phosphorus, iron, potassium, sodium, vitamin A, thiamine, riboflavin, niacin, and vitamin C) and reported each of them as insufficient, suitable, or excessive for the users. Users could also record the kind and duration of exercise through the exercise assessment function of LIBIT; it subsequently calculated the amount of calorie consumption and reported it to the users. LIBIT received participants’ self-monitored data on blood pressure and glucose from the peripheral devices through a Bluetooth connection and transmitted them to the main server using cellular data. The health monitoring function of LIBIT provided visual feedback of users’ health status by creating trend graphs of body weight, blood pressure, glucose level, and calorie consumption. The Mediram app offered comprehensive medication information to users (Multimedia Appendix 4). Users could easily upload the prescription to Mediram by just scanning the QR codes of their prescriptions. Mediram supplied drug information and notified users of their drug schedule to enhance their adherence to medication.
Physicians could access the participants’ health data using Bluetooth-enabled home devices and unmanned kiosks at the main server through web browsers (Multimedia Appendix 5). The average blood pressure, glucose level, adherence to medication, body composition change, nutrient intake, and calorie consumption through exercise in the given period were displayed on a page, and the physician could monitor changes in a participant’s health status at a glance.

Statistical Analyses
The input rate of food intake was calculated as the total input frequency divided by the product of the days of the mHealth period and frequency of daily food intake. The input rate of taking medicine was calculated similarly as the total input frequency divided by the product of days of the mHealth period and frequency of daily drug intake. The input rate of exercise was calculated as the total input frequency divided by the number of days of the mHealth period.

Student t test, chi-square test, and Fisher exact test were used to determine the differences in baseline characteristics between the mHealth-CON and CON-mHealth groups. The Wilcoxon signed-rank test was conducted to test the difference in the input rates of diet, exercise, and medicine intake between the groups. A paired t test or Wilcoxon signed-rank test was performed to determine the differences in changes between the integrative mHealth and CON periods. Subsequently, according to the input rate of medicine intake, the estimated between-group differences in changes in the variables—obesity, hypertension, and T2DM—were calculated using generalized linear models, after adjusting for treatment (mHealth or CON) in model 1; treatment, group (mHealth-CON or CON-mHealth), and sex in model 2; and treatment, group, sex, and age in model 3. To compare the 3 input functions of the apps, a Kruskal–Wallis rank sum test [18] was conducted.

All statistical analyses were implemented in the R software version 4.0.3 (R Core Team), which is a language and environment for statistical computing. A 2-tailed P<.05 was considered statistically significant.

Results
Baseline Characteristics
A total of 37 participants were enrolled in this pilot study. Among the 37 participants, there were 3 (8%) cases of newly diagnosed T2DM, 1 (3%) case of bioelectrical impedance analysis showing error, and 1 (3%) case of dropout. In the final data set, 32 participants’ data were included, of whom 15 (47%) were allocated to the mHealth-CON group, whereas the remaining 17 (53%) were assigned to the CON-mHealth group.

Among the 32 participants, 23 (72%) were men. The mean age of participants was 56.8 years, ranging from 40 to 69 years. Of the 32 participants, 17 (53%) graduated from college or above. Approximately 38% (12/32) of participants had hypertension. Most participants were overweight and obese. According to the BMI classification of the Korean Society for the Study of Obesity [19], only 3% (1/32) of participants were in the normal range, that is, 20.8 kg/m\(^2\). Of the 32 participants, 7 (22%) were overweight, with BMIs ranging from 23 kg/m\(^2\) to 24.9 kg/m\(^2\); 20 (63%) had class 1 obesity, with a BMI range of 25 kg/m\(^2\) to 29.9 kg/m\(^2\); and BMIs of the remaining 4 (13%) participants were >30 kg/m\(^2\). The mean HbA\(_{1c}\) level was 7.5%, ranging from 6.1% to 9.4%. The baseline demographic and clinical characteristics of the 2 groups, mHealth-CON and CON-mHealth, were similar, except for the frequency of hypertension (Table 1). There were no significant differences in demographic characteristics, body weight and body composition, blood pressure, and HbA\(_{1c}\) levels between the groups.

https://mhealth.jmir.org/2022/1/e27192
Table 1. Participant baseline characteristics (N=32).

<table>
<thead>
<tr>
<th>Group</th>
<th>mHealth^a-CON^b (n=15)</th>
<th>CON-mHealth (n=17)</th>
<th>P value^c</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>9 (60)</td>
<td>14 (82)</td>
<td>.24</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>58.9 (4.9)</td>
<td>55.1 (7.6)</td>
<td>.11</td>
</tr>
<tr>
<td><strong>Education, n (%)</strong></td>
<td></td>
<td></td>
<td>.14</td>
</tr>
<tr>
<td>Elementary school</td>
<td>1 (7)</td>
<td>1 (6)</td>
<td></td>
</tr>
<tr>
<td>Middle school</td>
<td>0 (0)</td>
<td>2 (12)</td>
<td></td>
</tr>
<tr>
<td>High school</td>
<td>8 (53)</td>
<td>3 (18)</td>
<td></td>
</tr>
<tr>
<td>College or above</td>
<td>6 (40)</td>
<td>11 (65)</td>
<td></td>
</tr>
<tr>
<td><strong>Marital status, n (%)</strong></td>
<td></td>
<td></td>
<td>.99</td>
</tr>
<tr>
<td>Divorced</td>
<td>1 (7)</td>
<td>1 (6)</td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>14 (93)</td>
<td>16 (94)</td>
<td>.40</td>
</tr>
<tr>
<td><strong>House, n (%)</strong></td>
<td></td>
<td></td>
<td>.71</td>
</tr>
<tr>
<td>Apartment</td>
<td>9 (60)</td>
<td>10 (63)</td>
<td></td>
</tr>
<tr>
<td>Detached</td>
<td>2 (13)</td>
<td>5 (31)</td>
<td></td>
</tr>
<tr>
<td>Unit</td>
<td>2 (13)</td>
<td>1 (6)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>2 (13)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td><strong>Monthly income^d, mean (SD)</strong></td>
<td></td>
<td></td>
<td>.01^e</td>
</tr>
<tr>
<td>Hypertension, mean (SD)</td>
<td>5.3 (2.3)</td>
<td>5.6 (2.3)</td>
<td></td>
</tr>
<tr>
<td>Body weight (kg), mean (SD)</td>
<td>73.4 (9.4)</td>
<td>80.0 (11.5)</td>
<td>.09</td>
</tr>
<tr>
<td>BMI (kg/m^2), mean (SD)</td>
<td>26.8 (2.3)</td>
<td>27.8 (3.6)</td>
<td>.34</td>
</tr>
<tr>
<td>Fat mass (kg), mean (SD)</td>
<td>22.3 (5.5)</td>
<td>22.8 (6.9)</td>
<td>.83</td>
</tr>
<tr>
<td>Body fat (%), mean (SD)</td>
<td>30.5 (7.1)</td>
<td>28.4 (6.9)</td>
<td>.38</td>
</tr>
<tr>
<td>Fat free mass (kg), mean (SD)</td>
<td>51.1 (9.0)</td>
<td>57.2 (9.5)</td>
<td>.07</td>
</tr>
<tr>
<td>Body water (%), mean (SD)</td>
<td>51.8 (4.1)</td>
<td>53.0 (4.8)</td>
<td>.45</td>
</tr>
<tr>
<td><strong>Blood pressure (mm Hg), mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Systolic</td>
<td>134.8 (11.4)</td>
<td>129.4 (13.2)</td>
<td>.23</td>
</tr>
<tr>
<td>Diastolic</td>
<td>81.5 (7.8)</td>
<td>78.8 (11.8)</td>
<td>.46</td>
</tr>
<tr>
<td><strong>HbA1c^f (%)</strong>, mean (SD)</td>
<td>7.5 (0.7)</td>
<td>7.5 (0.8)</td>
<td>.98</td>
</tr>
</tbody>
</table>

^a mHealth: mobile health.
^b CON: conventional treatment.
^c Calculated using Fisher exact test or Student t test.
^e Calculated using chi-square test.
^f HbA1c: hemoglobin A1c.

**Diet and Exercise**

The input rates of food intake (24.9%) and exercise (5.3%) were very low (Table 2). Moreover, both input rates were significantly low in the CON-mHealth group, which means that there was attrition in food intake and exercise input to the LIBIT app over time during the pilot study. On account of low input rates of food intake and exercise, it was impossible to execute the analysis of data related to nutrient intake and energy consumption.
Table 2. Input rates of diet, exercise, and taking medicine (N=32).

<table>
<thead>
<tr>
<th>Parameters</th>
<th>mHealth&lt;sup&gt;a&lt;/sup&gt;-CON&lt;sup&gt;b&lt;/sup&gt;</th>
<th>CON-mHealth</th>
<th>Total</th>
<th>P value&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diet, mean (SD)</td>
<td>36.6 (39.3)</td>
<td>14.5 (23.7)</td>
<td>24.9 (33.3)</td>
<td>.03</td>
</tr>
<tr>
<td>Exercise, mean (SD)</td>
<td>8.0 (9.3)</td>
<td>2.9 (8.9)</td>
<td>5.3 (9.3)</td>
<td>.002</td>
</tr>
<tr>
<td>Taking medicine, mean (SD)</td>
<td>80.3 (20.0)</td>
<td>87.3 (20.8)</td>
<td>84.0 (20.4)</td>
<td>.06</td>
</tr>
</tbody>
</table>

<sup>a</sup>mHealth: mobile health.
<sup>b</sup>CON: conventional treatment.
<sup>c</sup>Calculated using Wilcoxon rank sum test.

Drug Adherence

The input rate of medicine intake to the Mediram app was high, at 84.0%. Unlike the input rates of food intake and exercise, there was no attrition in the input of medicine intake. There was no significant difference in the input rate of taking medicine irrespective of whether the mHealth period was before or after the CON period (80.3% and 87.3%, respectively; P=.06; Table 2).

Efficacy of mHealth on Body Weight, Body Composition, Blood Pressure, and HbA<sub>1c</sub>

The changes in body weight and body composition, blood pressure, and HbA<sub>1c</sub> between the mHealth and conventional periods are displayed in Table 3. There were no significant differences in the changes in the variables of obesity, hypertension, and T2DM. The individual changes in these variables are shown in Multimedia Appendix 6.

Table 3. Comparison of the changes in body weight, body composition, blood pressure, and HbA<sub>1c</sub><sup>a</sup>.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>N (mHealth&lt;sup&gt;b&lt;/sup&gt;/CON&lt;sup&gt;c&lt;/sup&gt;)</th>
<th>mHealth, mean (SD)</th>
<th>Conventional, mean (SD)</th>
<th>P value&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Body weight (kg)</td>
<td>32/32</td>
<td>-0.519 (1.655)</td>
<td>0.000 (1.832)</td>
<td>.29</td>
</tr>
<tr>
<td>BMI (kg/m&lt;sup&gt;2&lt;/sup&gt;)</td>
<td>32/32</td>
<td>-0.133 (0.640)</td>
<td>-0.167 (0.709)</td>
<td>.86</td>
</tr>
<tr>
<td>Fat mass (kg)</td>
<td>31/30</td>
<td>-0.292 (1.964)</td>
<td>0.153 (2.357)</td>
<td>.56&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
<tr>
<td>Body fat (%)</td>
<td>31/30</td>
<td>-0.255 (2.454)</td>
<td>0.172 (2.574)</td>
<td>.43&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
<tr>
<td>Fat free mass (kg)</td>
<td>31/30</td>
<td>-0.102 (1.703)</td>
<td>-0.173 (1.933)</td>
<td>.78&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
<tr>
<td>Body water (%)</td>
<td>30/30</td>
<td>0.240 (3.173)</td>
<td>-0.360 (2.702)</td>
<td>.21</td>
</tr>
<tr>
<td>Blood pressure (mm Hg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Systolic</td>
<td>31/31</td>
<td>-0.226 (12.328)</td>
<td>-2.839 (11.097)</td>
<td>.51</td>
</tr>
<tr>
<td>Diastolic</td>
<td>31/31</td>
<td>-2.839 (9.000)</td>
<td>-1.097 (9.239)</td>
<td>.51&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
<tr>
<td>HbA&lt;sub&gt;1c&lt;/sub&gt;&lt;sup&gt;f&lt;/sup&gt; (%)</td>
<td>32/32</td>
<td>-0.269 (0.663)</td>
<td>-0.009 (0.693)</td>
<td>.19</td>
</tr>
</tbody>
</table>

<sup>a</sup>Some participants' data are missing.
<sup>b</sup>mHealth: integrative mobile health service.
<sup>c</sup>CON: conventional treatment.
<sup>d</sup>Calculated by paired t test.
<sup>e</sup>Calculated using Wilcoxon signed-rank test.
<sup>f</sup>HbA<sub>1c</sub>: hemoglobin A<sub>1c</sub>.

Effect of the Medication Assistance App on Body Weight, Body Composition, Blood Pressure, and HbA<sub>1c</sub>

To inspect the effect of mHealth apps on body weight and composition, blood pressure, and HbA<sub>1c</sub>, the estimated changes in variables, according to the input rate of medicine intake, were calculated using generalized linear models (Table 4). Group was included as a variable in models 2 and 3 as there were differences in the prevalence of hypertension between the groups mHealth-CON and CON-mHealth. In proportion to the elevation of the input rate of medicine intake, body fat mass and HbA<sub>1c</sub> were lower (Table 4). Owing to low input rates, the effects of food intake and exercise-related app functions on changes in the variables could not be assessed.
Table 4. The estimated changes in variables—obesity, hypertension, and type 2 diabetes mellitus—according to the input rate of taking medicine.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Model 1a Estimate</th>
<th>P value</th>
<th>Model 2b Estimate</th>
<th>P value</th>
<th>Model 3c Estimate</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Body weight (kg)</td>
<td>-0.016</td>
<td>.16</td>
<td>-0.019</td>
<td>.09</td>
<td>-0.019</td>
<td>.11</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>-0.004</td>
<td>.34</td>
<td>-0.005</td>
<td>.25</td>
<td>-0.005</td>
<td>.31</td>
</tr>
<tr>
<td>Fat mass (kg)</td>
<td>-0.027</td>
<td>.08</td>
<td>-0.034</td>
<td>.03</td>
<td>-0.032</td>
<td>.04</td>
</tr>
<tr>
<td>Body fat (%)</td>
<td>-0.030</td>
<td>.10</td>
<td>-0.037</td>
<td>.05</td>
<td>-0.035</td>
<td>.07</td>
</tr>
<tr>
<td>Fat free mass (kg)</td>
<td>0.007</td>
<td>.61</td>
<td>0.010</td>
<td>.49</td>
<td>0.008</td>
<td>.55</td>
</tr>
<tr>
<td>Body water (%)</td>
<td>0.035</td>
<td>.11</td>
<td>0.042</td>
<td>.07</td>
<td>0.040</td>
<td>.08</td>
</tr>
</tbody>
</table>

Blood pressure (mm Hg)

<table>
<thead>
<tr>
<th></th>
<th>P value</th>
<th>Estimate</th>
<th>P value</th>
<th>Estimate</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systolic</td>
<td>.95</td>
<td>-0.016</td>
<td>.85</td>
<td>-0.014</td>
<td>.87</td>
</tr>
<tr>
<td>Diastolic</td>
<td>.80</td>
<td>-0.028</td>
<td>.67</td>
<td>-0.028</td>
<td>.68</td>
</tr>
<tr>
<td>HbA1c (%)</td>
<td>.07</td>
<td>-0.009</td>
<td>.04</td>
<td>-0.010</td>
<td>.03</td>
</tr>
</tbody>
</table>

Survey on the Smartphone Apps

The results of the smartphone app survey at the end of the pilot study are summarized in Tables 5 and 6. Among the 3 input functions of food intake, exercise, and medicine intake in smartphone apps, the input of medicine intake was a more helpful, easier to use, and better-designed function than the others. There were more opinions about improvements in the input of food intake. The 2 most difficult functions were those of recording food intake and finding food items in the provided list. For the input of exercise, multitasking with other apps was highly desired.

Table 5. Survey about the functions of smartphone apps for the input of food intake, exercise, and taking medicine (Likert scale result; N=32).

<table>
<thead>
<tr>
<th>Function of the apps and question</th>
<th>Likert scale (1-5), mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input of food intake</strong></td>
<td></td>
</tr>
<tr>
<td>Helpfula</td>
<td>3.4 (1.0)</td>
</tr>
<tr>
<td>Easyb</td>
<td>2.9 (1.2)</td>
</tr>
<tr>
<td>Well-functionedc</td>
<td>3.1 (0.9)</td>
</tr>
<tr>
<td><strong>Input of exercise</strong></td>
<td></td>
</tr>
<tr>
<td>Helpfula</td>
<td>3.6 (1.0)</td>
</tr>
<tr>
<td>Easyb</td>
<td>3.3 (1.0)</td>
</tr>
<tr>
<td>Well-functionedc</td>
<td>3.3 (0.9)</td>
</tr>
<tr>
<td><strong>Input of taking medicine</strong></td>
<td></td>
</tr>
<tr>
<td>Helpfula</td>
<td>4.1 (0.8)</td>
</tr>
<tr>
<td>Easyb</td>
<td>4.1 (0.7)</td>
</tr>
<tr>
<td>Well-functionedc</td>
<td>4.0 (0.9)</td>
</tr>
</tbody>
</table>

aP=.01 using the Kruskal-Wallis rank sum test.
bP<.001 using the Kruskal-Wallis rank sum test.
cP=.002 using the Kruskal-Wallis rank sum test.
Table 6. Survey about the functions of smartphone apps for the input of food intake, exercise, and taking medicine (opinions; N=32).

<table>
<thead>
<tr>
<th>Opinions</th>
<th>Numbers, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input of food intake</strong></td>
<td></td>
</tr>
<tr>
<td>Have experience of using the diet app</td>
<td>1 (3)</td>
</tr>
<tr>
<td><strong>What needs to be improved?</strong></td>
<td></td>
</tr>
<tr>
<td>Difficult to record food intake</td>
<td>12 (38)</td>
</tr>
<tr>
<td>Errors in voice recognition</td>
<td>4 (13)</td>
</tr>
<tr>
<td>Wish the portion size was broken down</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Wish to input data on my own</td>
<td>2 (6)</td>
</tr>
<tr>
<td>No food in the food list</td>
<td>5 (16)</td>
</tr>
<tr>
<td>Unbelievable calculated calorie</td>
<td>1 (3)</td>
</tr>
<tr>
<td><strong>Input of exercise</strong></td>
<td></td>
</tr>
<tr>
<td><strong>What needs to be improved?</strong></td>
<td></td>
</tr>
<tr>
<td>Wish it worked with other apps simultaneously</td>
<td>7 (22)</td>
</tr>
<tr>
<td>Wish to input data on my own</td>
<td>4 (13)</td>
</tr>
<tr>
<td>Wish it was recorded automatically</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Difficult to input data</td>
<td>2 (6)</td>
</tr>
<tr>
<td></td>
<td>1 (3)</td>
</tr>
<tr>
<td><strong>Input of taking medicine</strong></td>
<td></td>
</tr>
<tr>
<td><strong>What needs to be improved?</strong></td>
<td></td>
</tr>
<tr>
<td>Errors in running the app</td>
<td>5 (16)</td>
</tr>
<tr>
<td>Not easy to use</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Wish to go back to home screen after the input</td>
<td>2 (6)</td>
</tr>
<tr>
<td>Wish to control the medication time</td>
<td>1 (3)</td>
</tr>
</tbody>
</table>

**Discussion**

**Principal Findings**

This pilot study demonstrated that smartphone apps could influence changes in body fat and blood glucose levels. As the input rate of medicine intake increased, body fat mass and HbA1c decreased. Although the improvement in drug adherence for diabetes is expected to enhance the control of blood glucose levels, good adherence to antidiabetes medication is irrelevant to body fat reduction; however, some—not all—antidiabetes medicines can induce body weight loss [20]. This relationship can be better explained by positive behavioral changes and effective self-management skills obtained from the integrative mHealth intervention. These findings imply that mHealth can improve body fat and blood glucose status in patients with T2DM or hypertension; however, it failed to result in clinical improvement in this pilot study. In addition, an app for drug information and reminders is more pleasing to the eyes of the patients than a diet diary and exercise monitor. A larger and long-term clinical trial is needed to determine whether integrative mHealth services help patients with T2DM or hypertension and obesity. The findings of this pilot study are currently being applied to an improved mHealth intervention project for people who have moved into a large, new apartment complex.

A special feature of this pilot study is its crossover design. This pilot study allocated participants to 2 different interventions over two 3-month periods. A crossover design has the following advantages over a parallel design: (1) it may offer more precise estimates of intervention effects as it would remove the differences in participants’ characteristics and methodological variations of open trials, and (2) it requires a smaller number of participants [21,22]. Obesity, hypertension, and T2DM are chronic diseases and are appropriate for a crossover study as the conditions of patients are stable if the prescriptions do not change, and they are not usually curable. In addition, a crossover design was appropriate as integrative mHealth service was an add-on treatment to the CON and not a separate stand-alone treatment.

Although there was a chance of carryover effect in this pilot study, it could not have widened the differences in changes between the treatment groups. A washout period was logically impossible for the transition from CON to the add-on integrative mHealth intervention. There might be a carryover effect in the mHealth-CON group, which could be the reason that integrative mHealth could not induce significant clinical improvement as compared with CON alone. According to previous studies [14,23], mHealth has a clinically positive effect on chronic disease management. In the case of the carryover effect, it would strengthen the positive effect of the second phase of the treatment, that is, CON; subsequently, the differences in the 2
treatment phases would become smaller, which would make the analysis more conservative.

The input rate of taking drugs was high, that is, >80%, even in the CON-mHealth group, where Mediram use was assigned in the second 3-month period. In fact, it was marginally higher in the CON-mHealth group than in the mHealth-CON group. Pharmacological adherence is very important for the successful management of hypertension and diabetes. Among adults with several common chronic diseases, only 40% to 70% of medications are taken properly [24,25]. Poor drug adherence might lead to poor clinical outcomes and increased HCE in chronic diseases [26]. The use of mHealth may increase medication adherence in chronic diseases and coronary heart disease; however, the results are diverse [27-29]. The variation among the previous trials was probably because of the different modules of mHealth that facilitated proper drug intake. It appears that the medication assistance app Mediram independently improved body fat and blood glucose status in our pilot study. Further research is needed to verify the clinical usefulness of the medication assistance app and the features of the app that make it more effective.

On the contrary, the nutrition and exercise care app LIBIT was not popular among most participants. Although LIBIT introduced a voice recognition technique for diet diaries to make food intake input much easier, the mean input rate of food intake was only 24.9%. The self-recording rate of exercise was even lower, at 5.3%. Following the low input rates of food intake and exercise, individualized lifestyle feedback based on the amount of macro- and micronutrient intake and calorie consumption through exercise was not properly reported to the participants. Most of all, for the individualized lifestyle feedback system, the input methods of diet and activity should be easy and simple, as revealed by the participants.

The integrative mHealth service of this pilot study failed to clinically reduce body weight and fat, lower blood pressure, and improve T2DM. The amount of HbA1c reduction in the mHealth period in our pilot study (~0.269%) was smaller than that in previous reports. A total of 2 meta-analyses reported that mHealth interventions improved HbA1c significantly, with mean differences of ~0.39% and ~0.44%, respectively [14,23]. However, for obesity and blood pressure, the results of previous studies were mixed, and there might be attrition of improvement over time. A randomized controlled trial for 3 months in patients with T2DM and hypertension reported that the blood pressure differences between the intervention and control groups were narrowed during the second and third months compared with that of the first month [30]. Wang et al [13] outlined that mHealth induced an average weight loss widely ranging from −1.97 kg in 16 weeks to −7.1 kg in 5 weeks and mentioned that most studies were executed with small samples and in short intervention periods and did not use proper data collection or analytical methods. A reason for the failure of our mHealth intervention in improving body fat, blood pressure, and blood glucose levels might be the unsuccessful individualized lifestyle modification linked with the neglected input of food intake and exercise, considering the high input rate of medicine intake linked with the decrease in body fat and HbA1c.

Diabetes and hypertension are 2 major chronic diseases that incur burdens on public health economically, and the burdens will be bigger in the near future of the aging world. Life expectancy (LE) has increased drastically over the past several decades. Between 1950 and 2017, it increased from 48.1 years to 70.5 years for men and from 52.9 years to 75.6 years for women worldwide [31]. However, there is a big gap between LE and health-adjusted LE (HALE). Although the global average HALE increased from 57.6 years in 1995 to 63.3 years in 2017, the gap between LE and HALE also increased from 8.6 years to 9.7 years during the same period [32]. Population aging has been considered a big source of increase in HCE, and individual health status has been suggested as a main factor of HCE in the aging population [33]. Older adults who are hypertensive patients are more likely to have complications, including congestive heart failure or chronic kidney disease, and these comorbidities induce incremental medical expenditures for adults aged ≥65 years, which is approximately US $2500 more than that for adults aged 18 to 44 years [6]. Similarly, the prevalence of diabetes and the medical costs related to diabetes are primarily increasing among the population aged ≥65 years [7]. Integrative mHealth can be a cost-effective tool to prevent catastrophic complications and increased HCE, which are associated with hypertension and diabetes management in the aging population, by enhancing their self-monitoring skills and drug adherence.

There were some limitations to our pilot study. First, the treatment period was short, at 3 months. Until now, the long-term efficacy of mHealth has been doubtful, and one cannot be sure if its effect would wear down over time. Long-term clinical trials with serial assessments of their effects are necessary for the future. Second, only a few participants were prompt in recording their input of food intake and exercise on the lifestyle app. Consequently, without the availability of individualized lifestyle information, the integrative mHealth service in this pilot study was scaled down to the combination of self-monitoring of blood pressure and glucose levels, medication assistance app, unmanned kiosks, and physicians’ access to participants’ health records through a web browser. In addition to simpler and easier input methods of diet and exercise, more immersive and highly functional mHealth apps should be designed that focus in depth on the content and user experience and can motivate patients. Third, bioelectrical impedance analysis was used for the measurement of body composition. For the assessment of treatment effects on body composition, dual-energy X-ray absorptiometry would be appropriate, as bioelectrical impedance analysis is easily affected by the hydration status of the body. Finally, this pilot study was conducted with relatively young adults and not with older adults. Only 6% (2/32) of participants were aged ≥65 years. As chronic disorders are more common among older adults who may not have good digital literacy and have difficulty in adopting new information technologies, the apps and peripheral medical devices that are designed for self-monitoring should be better accessible for older adults. In addition, voice-based mHealth may be preferred by older adults with limited digital literacy and poor vision [34].
Conclusions
This pilot study illustrated that smartphone apps could influence changes in body fat and blood glucose; however, mHealth failed to result in clinical improvement. A higher input rate of medicine intake was related to a significantly lower body fat mass and HbA1c. This result could possibly be because of positive behavioral changes and effective self-management skills obtained from the integrative mHealth intervention. In addition, the app for drug information and reminders was considered to be more pleasing to the eyes of the patients than a diet diary and exercise monitor.

Acknowledgments
The authors appreciate Huraypositive Co for providing the LIBIT app and GST Korea for supplying the Mediram app and nonhuman kiosks.

This work was supported by a grant (19RERP-B090094-06) from the Residential Environment Research Program funded by the Ministry of Land, Infrastructure, and Transport of the Korean government (NRF-2017R1A6A3A11036413) and a grant from the National Research Foundation of Korea funded by the Korean government.

Authors' Contributions
SWO conceptualized, funded, and administered the pilot study; KKK and SSK executed the clinical trial; KKK and SKP acquired and analyzed the data; KKK and SP interpreted data; KKK drafted the manuscript; and KKK, SSK, SP, and SWO revised the manuscript with critics.

Conflicts of Interest
None declared.

Multimedia Appendix 1
(a) Bluetooth blood pressure monitor and (b) Bluetooth blood glucose monitor.

Multimedia Appendix 2
Nonhuman kiosks comprising a bioelectrical impedance analyzer, a user screen, a mobile health service radio-frequency identification card reader, and an automatic upper arm blood pressure monitor.

Multimedia Appendix 3
Smartphone screen of the LIBIT app for (a) nutrition and (b) exercise assessment.

Multimedia Appendix 4
Smartphone screen of the Mediram app.

Multimedia Appendix 5
Participant’s health data access through web browser.

Multimedia Appendix 6
Line plots of individual changes in (a) body weight, (b) BMI, (c) body fat mass, (d) body fat percentage, (e) fat free mass, (f) body water percentage, (g) systolic blood pressure, (h) diastolic blood pressure, and (i) HbA1c; mHealth: integrative mobile health service; CON: conventional treatment.

References


Abbreviations

CON: conventional treatment
HALE: healthy life expectancy
HbA1c: hemoglobin A1c
HCE: health care expenditure
LE: life expectancy
mHealth: mobile health
SSL: secure socket layer
T2DM: type 2 diabetes mellitus

©Sang Woo Oh, Kyoung-Kon Kim, Sung Soo Kim, Su Kyung Park, Sangshin Park. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 11.01.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
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 Mobile App to Increase Fruit and Vegetable Acceptance Among Finnish and Polish Preschoolers: Randomized Trial

Henna Vepsäläinen¹, PhD; Essi Skaffari¹, MSc; Katarzyna Wojtkowska², PhD; Julia Barlińska², PhD; Satu Kinnunen¹, MSc; Riikka Makkonen¹,³, MSc; Maria Heikkilä¹, PhD; Mikko Lehtovirta¹, MD, PhD; Carola Ray¹,⁵, PhD; Eira Suohonén⁶, PhD; Jaakko Nevalainen⁷, PhD; Nina Sajaniemi¹,³,⁶, PhD; Maijaliisa Erkkola¹, PhD

¹Department of Food and Nutrition, University of Helsinki, Helsinki, Finland
²Faculty of Psychology, University of Warsaw, Warsaw, Poland
³School of Applied Educational Science and Teacher Education, University of Eastern Finland, Joensuu, Finland
⁴Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland
⁵Folkhälsan Research Center, Helsinki, Finland
⁶Department of Education, University of Helsinki, Helsinki, Finland
⁷Faculty of Social Sciences, Tampere University, Tampere, Finland

Corresponding Author:
Henna Vepsäläinen, PhD
Department of Food and Nutrition
University of Helsinki
PO Box 66
Helsinki, 00014
Finland
Phone: 358 443581467
Email: henna.vepsalainen@helsinki.fi

Abstract

Background: Early childhood education and care (ECEC) centers are ideal venues for food education. As smartphones and tablets are becoming increasingly popular in ECEC centers, technology can be used to deliver such pedagogical content. Evidence suggests that video games can affect fruit and vegetable (FV) consumption among 9- to 12-year-old children, but studies among preschoolers are scarce.

Objective: This paper describes the development of the Mole’s Veggie Adventures app and its effectiveness in increasing FV acceptance among Finnish and Polish preschoolers aged 3 to 6 years.

Methods: A multiprofessional team created an app to be used in ECEC centers in groups of 3 to 10 children. The app aimed to increase vegetable acceptance, and it was built using elements that support the development of self-regulation and social skills. Altogether, 7 Finnish and 4 Polish ECEC centers participated in the study. Before randomization, parents reported background factors and their children’s willingness to taste different FVs. The ECEC professionals in the intervention arm were instructed to use the app at least once a week during the 3- to 4-week intervention period. The main outcomes in this unblinded, cluster-randomized study were FV acceptance and relative FV acceptance. The first was calculated as a sum variable describing the children’s willingness to taste 25 different FVs, the second as FV acceptance divided by the number of FVs served. We used analysis of covariance to compare the FV acceptance and relative FV acceptance scores between the intervention and control groups at follow-up.

Results: A total of 221 children were included in the analysis. At follow-up, the intervention group (115/221, 52%) had higher FV acceptance scores (baseline adjusted difference of mean 7.22; 95% CI 1.41-13.03) than the control group (106/221, 48%). The intervention effect was parallel for relative FV acceptance scores (baseline adjusted difference of mean 0.28; 95% CI 0.05-0.52).

Conclusions: The Mole’s Veggie Adventures app has the potential to increase FV acceptance among preschoolers and can be a valuable tool in supporting food education in ECEC centers. Furthermore, the app can be feasibly incorporated into preschool routines in countries with different educational environments.

Trial Registration: ClinicalTrials.gov NCT05173311; https://tinyurl.com/4vfbb283
Introduction

Background

Most European children do not consume the recommended amount of fruit and vegetables (FVs) [1,2]. Among European countries, Poland and Finland face the same challenges. For instance, there seems to be a large proportion of both Polish and Finnish preadolescents who do not eat FVs daily [3]. Studies among Polish preschoolers are scarce, but in a 2011 report, the proportion of Finnish 6- to 8-year-olds consuming the recommended amount of FVs was less than 5% [4]. More recent studies have observed average consumption to be closer to recommendations, but the consumption of vegetables seems to be lower than that of fruit [5,6]. Indeed, children tend to prefer sweet tastes, as observed in fruits, compared with bitter-tasting foods, such as vegetables [7], and need more taste exposures to accept new vegetables [8]. However, as reassuring evidence suggests that repeated taste exposure and even exposure to picture books can help children to learn to enjoy vegetables [7-9], early childhood is a significant phase to support the formation of healthy eating habits among children.

In both Poland and Finland, most 3- to 6-year-olds attend early childhood education and care centers (ECECs) [10,11]. The general aim of the Finnish curriculum for ECEC is to strengthen skills related to children’s well-being [12], such as self-regulation skills, which refer to the ability to monitor and manage emotions and behaviors. Reinforcing self-regulation skills in childhood is important because they are associated with health outcomes later in life [13-15]. Moreover, self-regulation skills are also linked to health behaviors because, for instance, eating is regulated according to internal cues of hunger and fullness [16]. Both the Polish and Finnish recommendations for ECEC encourage ECEC professionals to support the development of self-regulation in eating and to promote a positive attitude toward food and eating [17,18]. Hence, food education is part of the pedagogically guided activities and holistic learning about well-being. However, ECEC professionals lack concrete, age-appropriate, effective, yet appealing tools for food education.

Objective

Mobile devices, such as smartphones and tablets, are ubiquitous and increasingly used at ECEC centers [19]; thus, food education can be delivered using technology. As ECEC centers should provide children with equal possibilities to familiarize themselves with technology and to practice responsible use of digital devices [12], the use of technology per se is beneficial. Furthermore, video games can trigger feelings of joy, intense participation, social interaction, and pleasure [20,21], and their educational use is considered promising [22]. Vast numbers of educational games and apps are already available through the digital distribution platforms Google Play and the App Store, and some of these even focus on food education. Studies reporting on the use of digital games and food-related outcomes are scarce and mostly concentrate on negative outcomes, such as an increase in fast food consumption due to advergaming exposure [23,24]. However, evidence from the United States suggests that video games can positively affect FV intake among 9- to 12-year-olds [25-27]. Moreover, a mobile app including vegetable-based activities has been shown to increase liking and consumption of vegetables among 3- to 6-year-old children in the United Kingdom [28]; however, educational games targeting ECEC environments are lacking. To fill this gap in knowledge, this paper describes the development of the Mole’s Veggie Adventures app and its effectiveness in increasing FV acceptance among Finnish and Polish preschoolers.

Methods

App Development

As part of the European Institution of Innovation & Technology (EIT) Food School Network project and together with a software development company specialized in designing, developing, and implementing serious games and gamified solutions (NordicEdu Oy), we designed and pilot-tested the Mole’s Veggie Adventures mobile app to increase vegetable acceptance among preschoolers. The University of Helsinki team was in charge of the app development, and the educational content of the app was designed by experts in nutrition science, food education, and ECEC. The content was first created in Finnish and later adapted and translated into English and Polish. The design of the app is described in detail in Multimedia Appendix 1. Briefly, the app was designed to be used in ECEC centers in groups of 3-10 children, but the ECEC professionals were encouraged to adjust the contents to fit the current situation in their group. The primary purpose of the app is to familiarize children with FV and increase FV acceptance. Unlike traditional mobile apps, Mole’s Veggie Adventures was built using elements that support the development of self-regulation and social skills. The app consists of 4 seasons, each of which includes 6 FVs. At the time of the intervention, the app listed 6 tasks for each of the vegetables and fruits: (1) Learn, (2) Color, (3) Shape, (4) Taste, (5) Pretend, and (6) Play, and the current version was numbered 0.4.5.0 (7b57516). An updated version of the Mole’s Veggie Adventures app is free of charge and available for download in the App Store and Google Play, and Multimedia Appendix 2 provides an overview of the most important sections of the app. All changes were made after the intervention. Table 1 summarizes the main characteristics of the app.
Table 1. Short description (descriptive table modified from the original form [29]) of the Mole’s Veggie Adventures app.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health topics covered</td>
<td>Food behavior (especially acceptability and consumption of FVs(^a)); food education</td>
</tr>
<tr>
<td>Targeted age group and environment</td>
<td>3- to 6-year-olds; preschool groups</td>
</tr>
<tr>
<td>Short description of the game idea</td>
<td>The main character in the game is the Mole, who moves around in a vegetable patch. The game is divided into seasons, each of which contains 6 FVs. The children can familiarize themselves with the FVs by completing different tasks in a group. For each FV, there are adult-led tasks to be completed in groups. In addition, the game includes a Taste Bank, which can be used to record the number of FVs tasted by the group, and Mini-Games, which can be played individually or in pairs. Multimedia Appendices 1 and 2 describe the contents of the game in more detail</td>
</tr>
<tr>
<td>Target players</td>
<td>• Individual</td>
</tr>
<tr>
<td></td>
<td>• Dyad</td>
</tr>
<tr>
<td></td>
<td>• Small group</td>
</tr>
<tr>
<td>Guiding knowledge or behavior change theories, models, or conceptual frameworks</td>
<td>Interactive tailoring (the ECEC(^b) professionals can adjust the tasks to be suitable for their group); role-playing (the players can learn from each other and from the ECEC professionals); goal setting and social cognitive theory (the group can decide to taste new vegetables together); learning through play (social interaction, motor skills, self-regulation etc.)</td>
</tr>
<tr>
<td>Intended health behavior change</td>
<td>Increase in FV acceptance</td>
</tr>
<tr>
<td>Clinical or parental support needed?</td>
<td>ECEC professionals or parents needed in the adult-led sections; mini-games can be played without adults</td>
</tr>
<tr>
<td>Data shared with parent or clinician?</td>
<td>No</td>
</tr>
<tr>
<td>Type of game</td>
<td>• Active</td>
</tr>
<tr>
<td></td>
<td>• Role-playing</td>
</tr>
<tr>
<td></td>
<td>• Educational</td>
</tr>
<tr>
<td>Game platforms needed to play the game</td>
<td>• Smartphone</td>
</tr>
<tr>
<td></td>
<td>• Tablet</td>
</tr>
<tr>
<td>Recommended play time</td>
<td>30 min at a time; minimum of 1-2 times a week</td>
</tr>
</tbody>
</table>

\(^a\)FV: fruit and vegetable.  
\(^b\)ECEC: early childhood education and care.

Recruitment

To test the effectiveness of the app, we conducted a feasibility study in 2 countries, Finland and Poland, between September and November 2019. On the basis of the literature regarding pilot trial sample size estimation [30], we aimed to recruit 100 children from both countries. Owing to differences in the ECEC systems and cultural environment, we describe the recruitment separately for the 2 countries. In Helsinki, Finland, 12 ECEC center directors were contacted and asked to participate in the study. Of these, 33% (4/12) declined (2 ECEC centers had a busy schedule, 1 did not have enough resources to participate, and in 1 ECEC center, the ECEC professionals were not enthusiastic about the study). In addition, one director could not be reached by email or phone. Thus, 14 groups from 7 public ECEC centers (58% of those invited) agreed to participate in the study. From the consenting groups, we invited all children to participate in the study. Informed consent was requested from legal guardians (later referred to as parents) via the ECEC groups, and the parents of 56% (130/232) of children invited provided their consent to participate in the study. The study was approved by the Education Division of the City of Helsinki, and the University of Helsinki Ethical Review Board in Humanities and Social and Behavioral Sciences deemed the study to be ethically acceptable (Statement 35/2019).

In Poland, the heads of 4 ECEC centers agreed to participate in the study. The study was carried out in 1 public ECEC center in the countryside (Wilczyn) and 2 public (Międzylesie-Warsaw, Piaseczno-Warsaw) and 1 private (Kobyłka-Warsaw) ECEC center in large urban agglomerations. The University of Warsaw’s research team organized meetings for parents in each ECEC center. The aims of the meetings were to introduce the goals of the study and to present the educational content of the intervention to the ECEC professionals and parents of the participating children. The parents received informed consent forms in the meetings, and of 213 who were invited, the parents of 196 (92%) children provided their consent to participate in the study. The study procedure was evaluated and approved by the Ethics Committee of the Faculty of Psychology at the University of Warsaw.

Background Characteristics

At baseline, the parents of the participating children completed questionnaires regarding background factors and their children’s FV acceptance. They reported the child’s gender and birthdate as well as the number of children living in the same household. The number of children living in the same household was categorized into 3 groups: 1 child, 2 children, and 3 or more children. In addition, the parents indicated whether the child had any vegetable- or fruit-related food allergies. The parents reported their highest educational level using 6 predefined
response options (comprehensive school, upper secondary school, vocational school, bachelor’s degree, master’s degree, and licentiate or doctoral degree), which were categorized into 3 groups: low (comprehensive, upper secondary, or vocational school), middle (bachelor’s degree), and high (master’s degree or higher).

Outcomes
The parents filled in a questionnaire listing 25 vegetables and fruits and inquiring whether these had been offered to the child during the past 4 weeks and how the child reacted to those that had been served. All the listed vegetables and fruits were introduced in the app. The answer options were 0 = was not offered during the past four weeks, 1 = refused to touch food, 2 = touched food but did not put in/near mouth, 3 = put food to lips but not in mouth, 4 = put food in mouth but spat out/did not eat, and 5 = ate food. A similar questionnaire was used earlier in a UK study examining toddlers’ willingness to taste different foods [8]. For each participant, we calculated an FV acceptance score by summing the answers to each of the 25 vegetable and fruit items, with higher scores indicating a higher FV acceptance (theoretical range 0-125). We also calculated the number of FVs served during the past 4 weeks (range 0-25) and used this information to create a relative FV acceptance score (range 0-5) by dividing the FV acceptance score by the number of FVs served.

Intervention and Control Arms
After the parents of the participating children had filled in the baseline questionnaires, the participating ECEC centers (in Finland) or groups within the ECEC centers (in Poland) were randomly allocated into intervention and control arms. In Finland, 4 ECEC centers with 7 groups were randomized into the intervention arm, whereas 7 groups from 3 ECEC centers were in the control arm. In Poland, groups within each ECEC center were evenly randomized into the intervention (5 groups from 4 ECEC centers) and control arms (5 groups from 4 ECEC centers). The study was not blinded. Researchers visited the intervention arm groups and introduced the app to the ECEC professionals. The ECEC professionals received a printed guide, which contained instructions and information about the app, and a PDF version of the guide was also available through the app. The ECEC professionals were instructed to use the app with a tablet computer at least one to two times a week during the intervention period (3-4 weeks) and to record the number of tasks completed by their group in a logbook. In addition, we recommended that each group focus on at least six vegetables or fruits during the intervention period. The ECEC professionals also provided quantitative and qualitative feedback using a feedback form. Written feedback was used to update the app after the study period. The control arm groups were instructed to continue their normal routines during the intervention period. They were instructed to refrain from introducing any novel food education methods during the intervention period. After the intervention period, follow-up questionnaires inquiring about the children’s FV acceptance were distributed to the parents of the participating children. To treat the intervention and control groups democratically, the app was introduced to the control ECECs after the study period. The trial was not registered because the study did not assess health outcomes and was thus not a clinical trial (World Health Organization defines a clinical trial as “any research study that prospectively assigns human participants or groups of humans to one or more health-related interventions to evaluate the effects on health outcomes” [31]).

Statistical Analysis
We used an analysis of covariance-type linear model to compare the FV acceptance and relative FV acceptance scores at follow-up. These models were adjusted with baseline FV acceptance score categories (missing, lower than median, and median or higher). To investigate the sensitivity of the results to the choice of the number of baseline categories, we also used baseline FV acceptance scores in tenths (unadjusted FV acceptance score) and sevenths (relative FV acceptance score) in the models.

Results
Altogether, 67.8% (221/326) of children had data on FV acceptance and relative FV acceptance scores at follow-up and were included in the analyses. The participating children were on average aged 5.0 years (SD 1.2 years). About half of the participants were girls, and slightly more children participated from the Polish than from the Finnish ECEC centers (Table 2). In 55.7% (123/221) of the participating families, at least one of the parents had a master’s degree or higher education, and most of the respondents (the person who filled in the questionnaires on behalf of the child) were mothers.

Table 3 shows the FV acceptance and relative FV acceptance scores at baseline and at follow-up. At follow-up, the FV acceptance score was 78.5 in the intervention group and 72.4 in the control group, whereas the values for relative FV acceptance scores were 3.97 and 3.75, respectively (Table 3). A score of approximately 3 means that on average, the children put the FVs on the lips but not in the mouth, whereas a score of approximately 4 implies that, on average, the children put the FVs in the mouth but did not eat them. When adjusted for baseline FV acceptance score category, participants in the intervention arm scored higher than control participants (Δ estimate 7.22; 95% CI 1.41-13.03). This corresponds to approximately a 10% improvement because of the intervention compared with a control group participant with the same baseline score. Similarly, relative FV acceptance scores at follow-up were, on average, 0.28 higher (+7%) in the intervention arm group than in the control arm group (95% CI 0.05-0.52). Regarding FV acceptance scores, the sensitivity analyses (data not shown) showed a similar and consistently significant intervention effect (Δ estimate 6.38; 95% CI 0.69-12.07), whereas a smaller and borderline significant intervention effect (Δ estimate 0.19; 95% CI -0.03-0.41) was detected for the relative FV acceptance score.

On average, the intervention arm groups used the app 1.9 times/week during the intervention period. The frequency of app use was missing from one group, but the number of completed tasks was as recommended, suggesting sufficient compliance. Furthermore, 17% (2/12) of groups did not use the app as instructed, that is, they used the app less than once a
week and did not complete tasks related to at least six FVs. Both
groups were from Finnish ECEC centers. The app was typically
used in a group of 2-10 children in the Finnish ECEC centers,
whereas the usual group size in the Polish ECEC centers was
24-25 children. The most popular FVs chosen by the
intervention arm groups were blueberries (11/12, 92% of the
groups completed related tasks); lettuce (10/12, 83%);
mushrooms (9/12, 75%); kidney, brown, and black beans (9/12,
75%); beetroot (8/12, 67%); and squash (8/12, 67%).

Table 2. Description of the study population (N=221).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Total (N=221), n (%)</th>
<th>Intervention (n=115), n (%)</th>
<th>Control (n=106), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Girls</td>
<td>120 (54.3)</td>
<td>67 (58.3)</td>
<td>53 (50.0)</td>
</tr>
<tr>
<td>Boys</td>
<td>100 (45.2)</td>
<td>47 (40.9)</td>
<td>53 (50.0)</td>
</tr>
<tr>
<td>Missing</td>
<td>1 (0.5)</td>
<td>1 (0.9)</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>Country</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finland</td>
<td>95 (43.0)</td>
<td>50 (43.5)</td>
<td>45 (42.5)</td>
</tr>
<tr>
<td>Poland</td>
<td>126 (57.0)</td>
<td>65 (56.5)</td>
<td>61 (57.5)</td>
</tr>
<tr>
<td>Vegetable or fruit allergy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>208 (94.1)</td>
<td>108 (93.9)</td>
<td>100 (94.3)</td>
</tr>
<tr>
<td>Yes</td>
<td>12 (5.4)</td>
<td>7 (6.1)</td>
<td>5 (4.7)</td>
</tr>
<tr>
<td>Missing</td>
<td>1 (0.5)</td>
<td>0 (0.0)</td>
<td>1 (0.9)</td>
</tr>
<tr>
<td>Number of children living in the same household</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One</td>
<td>56 (25.3)</td>
<td>27 (23.5)</td>
<td>29 (27.4)</td>
</tr>
<tr>
<td>Two</td>
<td>124 (56.1)</td>
<td>62 (53.9)</td>
<td>62 (58.5)</td>
</tr>
<tr>
<td>Three or more</td>
<td>36 (16.3)</td>
<td>22 (19.1)</td>
<td>14 (13.2)</td>
</tr>
<tr>
<td>Missing</td>
<td>5 (2.3)</td>
<td>4 (3.5)</td>
<td>1 (0.9)</td>
</tr>
<tr>
<td>Respondent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Father</td>
<td>27 (12.2)</td>
<td>15 (13.0)</td>
<td>12 (11.3)</td>
</tr>
<tr>
<td>Mother</td>
<td>193 (87.3)</td>
<td>100 (87.0)</td>
<td>93 (87.7)</td>
</tr>
<tr>
<td>Missing</td>
<td>1 (0.5)</td>
<td>0 (0.0)</td>
<td>1 (0.9)</td>
</tr>
<tr>
<td>Parental educational level</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upper secondary school or lower</td>
<td>55 (24.9)</td>
<td>32 (27.8)</td>
<td>23 (21.7)</td>
</tr>
<tr>
<td>Bachelor’s degree or equivalent</td>
<td>39 (17.6)</td>
<td>21 (18.3)</td>
<td>18 (17.0)</td>
</tr>
<tr>
<td>Master’s degree or higher</td>
<td>123 (55.7)</td>
<td>60 (52.2)</td>
<td>63 (59.4)</td>
</tr>
<tr>
<td>Missing</td>
<td>4 (1.8)</td>
<td>2 (1.7)</td>
<td>2 (1.9)</td>
</tr>
</tbody>
</table>

Table 3. FV<sup>a</sup> acceptance and relative FV acceptance scores in the intervention (n=82-115) and control (n=79-106) groups at baseline and at follow-up.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Baseline, mean (SD)</th>
<th>Follow-up, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FV acceptance score&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intervention group</td>
<td>70.6 (25.5)</td>
<td>78.5 (30.6)</td>
</tr>
<tr>
<td>Control group</td>
<td>70.2 (25.0)</td>
<td>72.4 (26.2)</td>
</tr>
<tr>
<td>Relative FV acceptance score&lt;sup&gt;c&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intervention group</td>
<td>3.84 (1.06)</td>
<td>3.97 (1.03)</td>
</tr>
<tr>
<td>Control group</td>
<td>3.77 (1.10)</td>
<td>3.75 (1.01)</td>
</tr>
</tbody>
</table>

<sup>a</sup>FV: fruit and vegetable.

<sup>b</sup>FV acceptance score: sum variable describing willingness to taste the 25 FVs listed; higher score indicates higher FV acceptance (theoretical range 0-125).

<sup>c</sup>FV acceptance score: FV acceptance score divided by the number of FVs served (range 0-5).
Discussion

Principal Findings

This paper describes the design and pilot-testing of the Mole’s Veggie Adventures app, which aimed to increase FV acceptance among 3- to 6-year-old preschoolers in Finland and Poland. Our pilot study showed a favorable and meaningful intervention effect; compared with the control arm participants, the participants in the intervention arm had higher FV acceptance scores at follow-up 3-4 weeks after baseline. Thus, the app can be considered an effective food education tool in an early education environment. Earlier studies have shown that video games are effective in increasing FV consumption among 9- to 12-year-old children [25-27], and promising results among preschool aged participants have also been obtained [28]. To the best of our knowledge, the current food education tool is the first to be used routinely in a preschool environment. As healthy food behaviors, such as frequent and diverse FV consumption, are typically adopted in childhood and may track into adulthood [32-34], early childhood is a crucial time to intervene.

Comparison With Earlier Work

Some serious games emphasize increased knowledge [35], whereas others aim to incorporate multiple theory-driven behavior change techniques, such as tailoring, goal setting, problem solving, and feedback, into a fun and attractive game [36]. In Mole’s Veggie Adventures, no specific behavior change technique was selected, but several of them were included in the game mechanics. For instance, the ECEC professionals were instructed to select those FVs for discussion that they deemed most important for their group. Moreover, the game includes Mini-Games with educational and knowledge-enhancing content. In addition, advergame researchers have described multiple methods that can potentially influence player behavior [37]. Advertising in games can appear at different levels of the game and in many forms, for example, as product placement, background presentation, and engagement via interactivity. In the case of the Mole’s Veggie Adventures app, all the aforementioned expositions of FVs are present, as the product, FVs, appears both in the background and is subject to manipulation itself, potentially enhancing FV acceptance among children. In addition, because of the various types of tasks and activities in the game, the children do not familiarize themselves with FVs only virtually but also in reality. The diverse stimuli with FVs as the main characters may encourage children to become familiar with them. However, the extent to which such an intervention could realistically affect behavior warrants further research.

The Mole’s Veggie Adventures app includes a strong social aspect. It has been shown that ECEC professionals’ opinions may contribute to children’s food consumption [38]. In addition, peers may also act as role models for preschoolers [39]. As our game was used in the ECEC centers in a group of preschoolers, it may have offered opportunities to model—for better or worse—the early educators as well as other children and motivated children to try new FVs. Social interaction provides opportunities for problem solving and peer engagement, which in turn, can cultivate useful skills such as negotiation and cooperation [40]. In addition, approval from the early educator, supporting comments from the group, and the opportunity to boast and present the results of tasks in the group may have been rewarding. Previous serious game research has also suggested that engaging parents—gatekeepers of the home environment [41]—may be critical in changing child behavior [26,27]. Bearing this in mind, we updated the Mole’s Veggie Adventures app after the intervention to better fit both the preschool and home environments (Multimedia Appendices 1 and 2). We also encourage future game designers to consider including a parental component to ensure adult support in all environments relevant to the child.

Fun is an essential part of playing games and can produce intrinsic motivation in players [42]. However, it remains unknown how certain target groups (eg, preschoolers) comprehend and experience fun or how to use fun to design games to bring about larger or more consistent changes in health behaviors [43]. Baranowski et al [42] contemplated the building blocks of fun and suggested that fun in games is probably a combination of interaction, overcoming challenges, making choices, and detecting their consequences without risking oneself, receiving feedback, increasing difficulty through levels of game play, and using personally relevant stories and characteristics in meaningful situations. To ensure that the Mole’s Veggie Adventures app would be perceived as fun, preschoolers participated in the development process (see Multimedia Appendix 1 for details). In addition, the game incorporated elements, such as physical play, invented stories, and adult-led activities, which have been identified as occasions for fun and shared humor among preschoolers [44]. A growing consensus describes play as an intrinsically motivated activity that results in joyful discovery [40]; thus, it is presumable that the app, by covering various forms of play (ie, active physical and pretend play), was indeed perceived as fun by the preschoolers.

Although games can deliver food education in an enticing way, not all behaviors encouraged by games are beneficial. Excessive gaming can evoke negative psychosocial effects [20] and even cause addictive behaviors [45]. Possible adverse effects include increased impulsivity [46], and as impulse control is an element in self-regulation, games can impair the development of self-regulation skills. Poor self-regulation skills in childhood have been linked to diminished social and cognitive outcomes later in life [47] and may also be associated with adverse health outcomes such as overweight and obesity or increased screen time [48-52]. To avoid these pitfalls, the Mole’s Veggie Adventures app was designed to support the development of self-regulation skills. For instance, the app includes elements that require peaceful action and waiting. Moreover, the app is mostly intended to be used with an ECEC professional, whose role as a coregulator is significant in strengthening self-regulation skills [53]. Subdued, mild colors and delicate music allow the child to focus on the educational content and could potentially prevent impulsivity.
Study Strengths and Limitations

The strengths of the study include testing in 2 countries, Finland and Poland. The 2-country setting allowed us to recruit a larger sample, which in turn, enabled the detection of the intervention effect. Another strength is the random allocation of ECEC centers into the intervention and control arms. Therefore, it is unlikely that the outcome was confounded by uncontrolled variables. Moreover, we examined one specific outcome (FV acceptance) instead of testing for multiple outcomes, which could have led to type 1 error [54]. The app development process was extensive and included cocreation with the target group and a multidisciplinary research team as well as prepiloting of the demo version in Finnish preschools (see Multimedia Appendix 1 for details). Most intervention arm groups used the app as instructed, suggesting moderate feasibility.

Although the study was able to demonstrate favorable changes in FV acceptance, it also had some limitations that should be addressed. First, our study was not blinded, and the participating early educators in the intervention group knew that the children’s FV acceptance was being measured. The parents of the participating children were also aware of the intervention. However, the app was used in the ECEC centers, whereas parents reported FV acceptance, and thus, the parents did not know exactly how much their children had used the app. In Finland, randomization was conducted at the preschool level to avoid contamination. Owing to nonexistent between-group communication among parents, contamination was not considered probable in Poland. Second, only 72.9% (161/221) of participants had data at both baseline and follow-up. To use data from as many participants as possible, we categorized participants into 3 groups based on their baseline data: missing, lower than median, and median or higher. To determine how the categorization affected the results, we ran multiple sensitivity analyses, which yielded parallel results. Third, as the app was used in a group, we did not know which individual children in the intervention arm groups participated in the game sessions. In addition, as the degree of implementation varied between the ECEC groups, this could have attenuated the observed effects. Fourth, our sample was relatively highly educated, and thus, the results may not be generalizable to socioeconomically disadvantaged groups. Owing to differences in cultural environment, the recruitment process was carried out differently in the 2 countries, which probably resulted in differing participation rates (56% in Finland vs 92% in Poland). Thus, the Polish sample might have been more representative of the target population than the Finnish sample. Furthermore, because of the limited time frame set by the funding period, the intervention was relatively short. As children need repeated taste exposures, preferably integrated with sensory learning strategies as well as nutrition education to get used to different vegetables [55], it is possible that a longer intervention period would have been needed to achieve more prominent and permanent results. However, we realistically aimed to increase FV acceptability, not FV consumption, which would probably require more time. Future studies should include postintervention follow-up to examine the stability of the intervention effects.

Conclusions

In summary, the Mole’s Veggie Adventures app has the potential to increase FV acceptance among preschoolers. The app can support food education and be incorporated into the preschool curriculum in countries with different educational environments, such as Finland and Poland. When designing serious games for preschoolers, game designers should consider including both home- and preschool-based components to ensure adult endorsement in all relevant environments, which could result in even stronger effects. Future studies should aim to identify the game mechanisms that best support children in making behavior changes.

Acknowledgments

The authors thank the participating children, families, and municipalities, the early childhood education and care centers, and the early childhood education and care professionals. The authors are also grateful to Dr. Elena Santa Cruz, Dr. Alan Roberts, Professor Kate Harvey, and Professor Tom Baranowski for their collaboration in this project. This work was funded by EIT Food (The EIT Food School Network: Integrating solutions to improve eating habits and reduce food wastage # 18145, # 19057, # 20129). EIT Food is the Innovation Community on Food of the EIT, a body of the European Union, under Horizon 2020, the EU Framework Programme for Research and Innovation. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. Open access was funded by the Helsinki University Library.

Authors’ Contributions

HV, E Skaffari, ML, CR, E Suhonen, NS, and ME participated in the development process of the app; KW and JB translated the app into Polish and adapted it to the Polish cultural context; HV, E Skaffari, KW, JB, and ME designed the study; JB and ME were responsible for funding acquisition; ME was responsible for leadership in research activity planning and execution; HV and KW managed and coordinated research activities; HV, E Skaffari, KW, JB, SK, RM, and MH collected the data and participated in data curation; HV, JN, and ME designed the statistical analyses; HV analyzed the data and wrote the manuscript. All authors reviewed and approved the final manuscript.

Conflicts of Interest

The authors are the developers of the Mole’s Veggie Adventures mobile app. The authors have no other relationships or activities that could potentially be construed as a conflict of interest with the present work.
Editorial Notice

This randomized study was only retrospectively registered. The authors explained that their study "did not assess health outcomes and was not a clinical trial". However, readers are advised to carefully assess the validity of any potential explicit or implicit claims related to primary outcomes or effectiveness, as retrospective registration does not prevent authors from changing their outcome measures retrospectively.

Multimedia Appendix 1
Detailed description of the application development process.
[DOCX File, 175 KB - mhealth_v10i1e30352_app1.docx]

Multimedia Appendix 2
Video demonstrating the most important sections of the current version of the Mole’s Veggie Adventures application. Note that the version used in the current study differed slightly from the current version.
[MP4 File (MP4 Video), 31691 KB - mhealth_v10i1e30352_app2.mp4]

Multimedia Appendix 3
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 2541 KB - mhealth_v10i1e30352_app3.pdf]

References


**Abbreviations**

ECEC: early childhood education and care  
EIT: European Institution of Innovation & Technology  
FV: fruit and vegetable
A Mobile App to Increase Fruit and Vegetable Acceptance Among Finnish and Polish Preschoolers: Randomized Trial

JMIR Mhealth Uhealth 2022;10(1):e30352

URL: https://mhealth.jmir.org/2022/1/e30352
doi:10.2196/30352
PMID:34982718

©Henna Vepsälläinen, Essi Skaffari, Katarzyna Wojtkowska, Julia Barlińska, Satu Kinnunen, Riikka Makkonen, Maria Heikkilä, Mikko Lehtovirta, Carola Ray, Eira Suhonen, Jaakko Nevalainen, Nina Sajaniemi, Maijaliisa Erkkola. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 04.01.2022. 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.
Multipurpose Mobile Apps for Mental Health in Chinese App Stores: Content Analysis and Quality Evaluation

Xiaoqian Wu, MS; Lin Xu, MS; PengFei Li, PhD; TingTing Tang, MS; Cheng Huang, MD

1College of Medical Informatics, Chongqing Medical University, Chongqing, China
2Medical Data Science Academy, Chongqing Medical University, Chongqing, China
3School of Public Health, Weifang Medical University, Weifang, China
4The Children’s Hospital of Chongqing Medical University, Chongqing, China

Corresponding Author:
Cheng Huang, MD
College of Medical Informatics
Chongqing Medical University
No.1 Yixueyuan Road, Yuzhong District
Chongqing, 400016
China
Phone: 86 023 6848 0060
Email: huangcheng@cqmu.edu.cn

Abstract

Background: Mental disorders impose varying degrees of burden on patients and their surroundings. However, people are reluctant to take the initiative to seek mental health services because of the uneven distribution of resources and stigmatization. Thus, mobile apps are considered an effective way to eliminate these obstacles and improve mental health awareness.

Objective: This study aims to evaluate the quality, function, privacy measures, and evidence-based and professional background of multipurpose mental health apps in Chinese commercial app stores.

Methods: A systematic search was conducted on iOS and Android platforms in China to identify multipurpose mental health apps. Two independent reviewers evaluated the identified mobile apps using the Mobile App Rating Scale (MARS). Each app was downloaded, and the general characteristics, privacy and security measures, development background, and functional characteristics of each app were evaluated.

Results: A total of 40 apps were analyzed, of which 35 (87.5%) were developed by companies and 33 (82.5%) provided links to access the privacy policy; 21 (52.5%) apps did not mention the involvement of relevant professionals or the guidance of a scientific basis in the app development process. The main built-in functions of these apps include psychological education (38/40, 95%), self-assessment (34/40, 85%), and counseling (33/40, 82.5%). The overall quality average MARS score of the 40 apps was 3.54 (SD 0.39), and the total score was between 2.96 and 4.30. The total MARS score was significantly positively correlated with the scores of each subscale (r=0.62-0.88, P<.001). However, the user score of the app market was not significantly correlated with the total MARS score (r=0.17, P=.33).

Conclusions: The quality of multipurpose mental health apps in China’s main app market is generally good. However, health professionals are less involved in the development of these apps, and the privacy protection policy of the apps also needs to be described in more detail. This study provides a reference for the development of multipurpose mental health apps.

(JMIR Mhealth Uhealth 2022;10(1):e34054) doi:10.2196/34054

KEYWORDS
mobile apps; app; mental health; mHealth; content analysis

Introduction

Anxiety, depression, stress, and other mental health conditions are increasing worldwide, involving family, study, work, social intercourse, and other aspects. Nearly 1 billion people worldwide have been found to have mental disorders [1]. In China, the weighted lifetime prevalence of mental diseases, except dementia, among the population over the age of 18 years is as high as 16.57% [2]. Although mental disorders will impose varying degrees of burden on patients and their surroundings,
only 15.7% of individuals with lifelong mental disorders in China seek help [3]. This situation is still obvious in cities with high economic development. A study in Shanghai, China, revealed that approximately 21.4% of the subjects reported depressive symptoms but only 4.7% sought mental health services [4]. The general underutilization of mental health services is worrying. The main reasons for this are the shortage of resources, the limited number of mental service professionals, and the uneven geographical distribution of mental health services in China [5]. Additionally, stigmatization of mental health services [6], low perceived demand [7], and economic constraints make people reluctant to take the initiative to accept mental health services [8].

Mobile apps are considered an effective way to eliminate these obstacles, improve mental health awareness [9], and contribute to symptom tracking and self-management [10]. Mobile apps are not limited to a particular time and place and realize the large-scale provision of cost-effective medical services, especially for people who find it difficult to receive traditional psychological services, for example, the population in rural areas with relatively low economic development and those who find it difficult to receive services face-to-face because of special reasons. During the COVID-19 outbreak, the spread and uncertainty of the pandemic caused a pessimistic mood of anxiety and fear for some patients, medical workers, and the general public [11]. This led to a sudden increase in mental health problems and their higher incidence rate, and mental health needs worldwide [12,13]. However, social and interpersonal networks were relatively closed because of epidemic prevention and control. Thus, traditional mental health services were difficult to obtain. Mental health apps can overcome the limitation of distance and expand the scope of psychological counseling for people. This also highlights the potential of digital health in improving the coverage of mental health services [14,15].

The advantages of mobile apps and the growth of mental health demand are making mental health and adaptive mobile health (mHealth) apps increasingly popular. Furthermore, there is an urgent need for more research to promote the formulation of better mental health service recommendations, especially in China’s huge untapped market [16]. The main categories of mental health apps are assessment, tracking or monitoring, treatment, and multipurpose [17]. However, the multipurpose mental health apps integrate evaluation, monitoring, treatment, and other mental health services into 1 platform to provide users with one-stop services. They are the most popular apps for all ages [17]. However, there is currently no specific evaluation for multipurpose mental health apps. One study referred to multipurpose mental health apps, but there is a lack of standardized measures to evaluate and compare the quality of apps [18]. A study searched and evaluated China’s mental health apps [19]. However, some features of the apps, such as the professional background of app development, the theme distribution of the function, and the user privacy protection policy are still unclear. The privacy protection of applications is an important reference for people to choose mental health apps [20], and the lack of a professional background in the process of app development may reduce users’ confidence in the apps. Simultaneously, we found that great changes have taken place in China’s app market with the development of the internet and cell phone manufacturers. The app market developed by cell phone manufacturers replaced third-party app markets, such as 360 and Baidu, and occupied the main share of China’s app market, together with Tencent My App [21].

Therefore, this study aims to investigate the characteristics of multipurpose mental health apps in China, evaluate their quality, and describe the main functions, user privacy protection, and professional background of the development process of multipurpose mental health apps in the current market in order to help users make more informed choices and provide reliable evidence for app developers.

**Methods**

**Systematic Search Strategy**

This study featured a systematic search and content analysis of multipurpose apps on mental health in Chinese app stores on December 17, 2020. Huawei cell phones occupy the first place in the smartphone market share in mainland China [22], and the Huawei AppGallery has become the largest Android app store in China [21]. Moreover, Tencent My App, provided by the Chinese internet giant Tencent, is the second-largest Android app store in China [21]. Thus, we searched the Apple App Store (for iOS apps), Tencent My App, and Huawei AppGallery (for Android apps).

By preliminary test searches, the following keywords were determined: psychology, psychological counseling, psychological intervention, emotion, stress relief, anxiety, and depression. These keywords were searched anonymously using Chinese language terms in the app stores not logged into any user accounts. All search results were collected to ensure that all potentially relevant apps were captured. If an app exists in both iOS and Android and has the same design and content, the Android version was evaluated.

**Eligibility Criteria**

After removing duplicates, each potentially suitable app was reviewed by 2 independent researchers based on the app name, screenshots, and description. In this round, apps were included if they (1) provide multiple mental health services, (2) focus on individual mental health consumers seeking professional help, and (3) are in the Chinese language. Apps were excluded if they (1) focus on content unrelated to mental health services, such as social and communication apps, e-books, heartbeat measurement, and pulse measurement; (2) target mental health service providers, such as doctors, nurses, or counselors; and (3) provide a single function only. All apps that met the inclusion criteria were downloaded onto test devices. Of these, apps were excluded if they are not usable because of technical errors or require special authentication (such as an enterprise or a school).

**Data Extraction**

The relevant information provided by the app market was extracted to evaluate the descriptive features of the apps. The general characteristics of the apps, including platform,
developer, target user, update time, star rating, and downloads, were recorded. Additionally, combined with the researchers’ use of the apps, the characteristics of the apps with regard to personal privacy protection and professional development background were extracted (Table 1). Furthermore, through a literature review and group discussion, we divided the main services provided by mental health apps in China into 6 categories: psychoeducation, counseling, self-assessment, question-and-answer (Q&A) module, stress relief, self-monitoring, and management. The characteristics of the apps and the main services provided were recorded by 2 independent researchers. All differences were resolved through discussion until the researchers agreed upon the results.

### Table 1. General information collected for each app.

<table>
<thead>
<tr>
<th>Assessment measure</th>
<th>Definition and values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform</td>
<td>Apple App Store, Huawei AppGallery, Tencent My App</td>
</tr>
<tr>
<td>Developer</td>
<td>Unknown, commercial, psychological service organization, individual developer</td>
</tr>
<tr>
<td>Target user</td>
<td>Children or adolescents, general public, specific for women and patients with mental health problems</td>
</tr>
<tr>
<td>Update time</td>
<td>Days from the retrieval date to the last update</td>
</tr>
<tr>
<td>Star rating</td>
<td>Star rating (out of 5) left by users in the app store</td>
</tr>
<tr>
<td>Downloads</td>
<td>Number of app downloads in the app store</td>
</tr>
<tr>
<td>Privacy protection</td>
<td>Does the description of the app claim to provide privacy protection? (Yes or no)</td>
</tr>
<tr>
<td></td>
<td>Is there an obvious privacy protection logo in the process of using apps? (Yes or no)</td>
</tr>
<tr>
<td></td>
<td>Does the app report relevant privacy protection regulations? (Yes or no)</td>
</tr>
<tr>
<td>Evidence-based and professional background</td>
<td>Does the app claim to be designed based on proven psychotherapy theory or opinions of mental health service professionals (such as clinicians and psychotherapists) or whether the usability of the app has been proved by peer-reviewed academic research? (Yes or no)</td>
</tr>
</tbody>
</table>

### Quality Appraisal of Apps

To evaluated the quality of the apps, we used the Mobile App Rating Scale (MARS), a validated scoring tool for assessing the quality of mHealth apps [23]. MARS has been used to evaluate the quality of different apps, such as apps for mental disorders [24,25], nutrition [26], drug-drug interaction checks [27,28], and chronic disease management [29-31]. MARS contains 23 items, including 4 objective quality subscales of engagement, functionality, aesthetics, and information quality and 1 subjective quality subscale. All items were rated on a 5-point Likert scale from 1 (inadequate) to 5 (excellent). We emphasized the objective quality of the apps, so the subjective quality subscale was excluded from the study. Before formal scoring, all reviewers evaluated the apps that provided only a single mental health service (excluded from the analysis) and discussed inconsistencies in regard to the results to ensure a unified understanding of MARS projects and standards. To fully experience the service provided by the apps, 2 independent reviewers downloaded and used each app for at least 15 min. The score of each subscale is calculated as the mean of the items in that subscale, and the total score is the mean of each subscale, which describes the overall quality of the app.

### Statistical Analysis

The quantitative variables of the MARS score in quality evaluation are described by the mean and SD. Classification variables, such as app characteristics, are described by frequency and percentage. To ensure the reliability of the quality assessment of 2 independent observers, the intragroup correlation coefficients (2-way random, mean measurement, and absolute consistency) were used to evaluate the consistency of commentators at the subscale and overall score level [32]. Pearson correlation coefficients were used to compare (1) the MARS total score and each subscale score, (2) the MARS total score and the user rating, and (3) the user rating and each subscale score. All statistical analyses were conducted using SPSS Statistics 25.

### Results

#### App Selection

A total of 1674 apps were identified through keyword retrieval (711 [42.47%] apps from Apple App Store, 770 [46.00%] apps from Huawei AppGallery, and 193 [11.53%] apps from Tencent My App). Combining the search results of the 3 app stores, 144 (8.6%) duplicate apps were excluded. A total of 1440 (86.02%) apps were excluded on the basis of the exclusion-inclusion criteria. The remaining 90 (5.38%) apps were downloaded onto the evaluation device for further evaluation. Of these 90 apps, 45 (50%) were excluded because of technical reasons (unable to download or use normally because of major technical reasons) and 5 (5.6%) were excluded because of the need to provide special authentication (employees/students). Finally, 40 (44.4%) apps were included in this study (Figure 1).
General Characteristics

Of the 40 apps included, 31 (77.5%) are from Huawei AppGallery, 6 (15%) from Apple App Store, and 3 (7.5%) from Tencent My App. Most of the apps (35/40, 87.5%) are developed by companies (4 of them are companies mainly engaged in mental health services), 2 (5%) apps are from professional psychological counseling centers, and 3 (7.5%) apps are from individual developers. In addition, 2 of the 40 (5%) apps are specifically designed for adolescents or children. Furthermore, 21 (52.5%) were updated more than 1 month and less than 1 year ago, 15 (37.5%) were maintained within 1 month, and 4 (10%) were updated more than 1 year ago (Table 2).
Table 2. Flowchart for the systematic search and selection of apps (N=40).

<table>
<thead>
<tr>
<th>Assessment measure</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Platform</strong></td>
<td></td>
</tr>
<tr>
<td>Apple App Store</td>
<td>6 (15)</td>
</tr>
<tr>
<td>Huawei AppGallery</td>
<td>31 (77.5)</td>
</tr>
<tr>
<td>Tencent My App</td>
<td>3 (7.5)</td>
</tr>
<tr>
<td><strong>Developer</strong></td>
<td></td>
</tr>
<tr>
<td>Commercial</td>
<td>35 (87.5)</td>
</tr>
<tr>
<td>Psychological service organization</td>
<td>2 (5)</td>
</tr>
<tr>
<td>Individual developer</td>
<td>3 (7.5)</td>
</tr>
<tr>
<td><strong>Target user</strong></td>
<td></td>
</tr>
<tr>
<td>Children or adolescents</td>
<td>2 (5)</td>
</tr>
<tr>
<td>General public</td>
<td>38 (95)</td>
</tr>
<tr>
<td>Specific for women</td>
<td>0</td>
</tr>
<tr>
<td>Patients with mental health problems</td>
<td>0</td>
</tr>
<tr>
<td><strong>Update time</strong></td>
<td></td>
</tr>
<tr>
<td>&gt;1 year</td>
<td>4 (10)</td>
</tr>
<tr>
<td>&gt;1 month and &lt;1 year</td>
<td>21 (52.5)</td>
</tr>
<tr>
<td>&lt;1 month</td>
<td>15 (37.5)</td>
</tr>
<tr>
<td><strong>Privacy protection</strong></td>
<td></td>
</tr>
<tr>
<td>Privacy protection mentioned in the description</td>
<td>17 (42.5)</td>
</tr>
<tr>
<td>An obvious privacy protection logo</td>
<td>23 (57.5)</td>
</tr>
<tr>
<td>Relevant privacy protection regulations</td>
<td>33 (82.5)</td>
</tr>
<tr>
<td><strong>Evidence-based and professional background</strong></td>
<td></td>
</tr>
<tr>
<td>Proven psychotherapy theory</td>
<td>5 (12.5)</td>
</tr>
<tr>
<td>The opinions of mental health service professionals</td>
<td>14 (35)</td>
</tr>
<tr>
<td>Peer-reviewed academic research</td>
<td>1 (2.5)</td>
</tr>
<tr>
<td>Not mentioned</td>
<td>21 (52.5)</td>
</tr>
</tbody>
</table>

*Some apps claim to be designed based on one or more scientific foundations.

**Privacy Protection**

Of the 40 apps, 17 (42.5%) mentioned the protection of user privacy in the descriptive content of the app store. More than half of the apps (23/40, 57.5%) provide obvious identification during app use to remind users of privacy protection. Almost all apps (33/40, 82.5%) provide links in the app market interface or within the app to access the privacy policy. However, less than two-thirds of these 33 apps (n=21, 63.6%) display privacy policies before users log in to their accounts. All privacy policies explain how to collect user data, and almost all privacy policies (30/33, 90.9%) express how to share, transfer, and publicly disclose user data. Less than half of the apps explain how to store information (16/33, 48.5%) and how to use cookies and other similar technologies (15/33, 45.5%). Of the 33 privacy policies, 14 (42.4%) list the rights of users to manage personal data. For example, users have the right to delete or correct their data. Slightly more than half of the policies (19/33, 57.6%) explain the processing method of minor personal information.

Finally, 18 of the 33 (54.5%) apps provide the contact information of the data protection officer that users can access when they have questions or opinions about the content of the privacy policy.

**Evidence-Based and Professional Background**

In the description of the app market, 21 of the 40 (52.5%) apps do not mention that the app design specifies a relevant scientific basis. The remaining 19 (47.5%) apps claim to be designed on the basis of 1 or more scientific foundations: 14 (73.7%) apps are described as being designed according to the opinions of mental health professionals, such as clinicians, psychologists, and psychotherapists, while 5 (26.3%) apps claim to be designed using proven psychotherapy theories, such as cognitive behavioral therapy; of these 5 apps, only 1 (20%) claims that its usability has also been confirmed by peer-reviewed academic research.
Functionality Review

Psychoeducation (38/40, 95%), counseling (33/40, 82.5%), and self-assessment (34/40, 85%) occur in more than three-quarters of the apps, the Q&A community (27/40, 67.5%) appears in about two-thirds of the apps, and stress relief modules (9/40, 22.5%) exist in less than a quarter of the apps (Figure 2). The most common combination in multipurpose apps is the combination of psychoeducation, counseling, self-assessment, and the Q&A community (14/40, 35%). The second is the combination of psychoeducation, counseling, and self-assessment (7/40, 17.5%).

Figure 2. Functional review results of the multipurpose mental health apps. Q&A: question and answer.

Most of the applied psychological education, psychological counseling, and self-assessment function modules set up a topic classification in navigation. To understand the distribution of topics provided in the functional modules of the current apps, we created a heatmap, as shown in Figure 3. The most common themes are love and marriage emotion, parent-child education, and emotion management. The next most common themes are mental disorders in career development, interpersonal relationships, and personal growth.

Figure 3. A heatmap of the topic type (top) for each function and the app function type. Note: numbers in white refer to the frequency of topic types involved in app functions. Warmer colors indicate higher counts.

Psychoeducation

Almost all apps (38/40, 95%) provide psychological education intervention. The mental health education part of the apps is reflected by reading articles related to mental health (32/38, 84.2%), learning relevant courses (28/38, 73.7%), and obtaining relevant information through radio stations (8/38, 21.1%) or live broadcast (5/38, 13.2%).

Among them, the most common way is reading articles related to mental health. Most of these articles are originally created by the platform or psychological counselors, which have guiding and educational significance for the public. More than half of the apps (19/32, 59.4%) have classified the topics of articles to set columns. The main content of the columns includes love and marriage emotion (18/19, 94.7%), parent-child education (15/19, 78.9%), career development (14/19, 73.7%), and emotion management (11/19, 57.9%). Additionally, it also includes themes, such as interpersonal relationships (9/19, 47.4%), personal growth (9/19, 47.4%), gender psychology (4/19, 21.1%), and stress relief (6/19, 31.6%). However, only 2 of the 32 (6.3%) apps have set up columns for specific disorders, such as depression and anxiety, and 1 (3.1%) app has a column for sleep disorders. In addition, 2 (6.3%) apps have columns for students or teenagers, and 1 (3.1%) app set up a column with the rehabilitation story of psychological disorders as the main content.

Mental health–related courses record videos in the form of online education and invite mental health service professionals to teach according to the needs of the public. The most popular courses are mental health and emotion management (14/19, 73.7%), career development (14/19, 73.7%), and mental health and daily life (13/19, 68.4%).
to present their professional knowledge, which covers common mental health knowledge, such as emotion management, love emotion, and psychological knowledge. Some of the app courses are free, while most charge a specific fee, ranging from as low as RMB 1.9 (US $0.3) to more than RMB 10,000 (US $1596). Only 6 of 28 (21.4%) apps offer completely free courses. Most other apps (22/28, 78.6%) provide free and paid courses at the same time, and users can choose based on their situation.

Radio and live broadcast are considered less psychological education ways. Mental health education provided using the radio station is mainly manifested in showing users past cases of mental health disorder adjustment or sharing common methods of mental health disorder adjustment. Live broadcast makes up for the shortcomings of the radio form. Users can directly contact consultants through a live broadcast, which strengthens the interaction between users and consultants.

Counseling

Mental health counseling services are provided by mental health service providers, who help solve psychological problems and “heal the soul” through online listening and answering of questions. This functionality is provided in 33 of 40 (82.5%) of the apps. Particularly, the Xiaoxin Psychology app applies artificial intelligence technology to mental health services and provides online counseling services through intelligent robots. Additionally, almost all apps provide consulting services by psychological counselors, and users can select an appropriate provider by viewing the basic information about the psychological counseling provider or modifying the label. Common labels include professional qualification (31/33, 93.9%), areas of expertise (32/33, 97.0%), user evaluation (24/33, 72.7%), and service person-times (23/33, 69.7%). The professional qualification of psychological counselors is the key factor for users to choose from. However, only 19 of 33 (57.6%) apps clearly express the authenticity of professional qualification; 4 of these 19 (21.1%) apps provide evidence of professional psychological counselor qualification, such as a certificate photo or a certificate number. In addition, 15 of the 19 (78.9%) apps are guaranteed by the platform to ensure the authenticity of counselor data. All app downloads are free, but users are charged a specific consulting fee. The consultation cost varies depending on the time or number of times. Almost all consultants (30/33, 90.9%) use voice chat to communicate with consumers. Others provide consultation using text and pictures (21/33, 63.6%), video communication (20/33, 60.6%), and offline face-to-face consultation (16/33, 48.5%).

Self-Assessment

The most common function in the apps is psychological testing, accounting for 34 of 40 (85%) of the total. The apps provide some evidence-based or entertainment scales, and users can understand their mental health status through self-assessment of scale problems. Most apps (26/34, 76.5%) provide evidence-based scales, the most common of which include the Self-Rating Anxiety Scale (24/26, 92.3%) [33], the Self-Rating Depression Scale (22/26, 84.6%) [34], and the Symptom Checklist-90-Revised (14/26, 53.8%) [35]. Almost all apps (32/34, 94.1%) provide a scale for the nature of entertainment to attract users’ attention. These scales are developed by the app team or formed by a scale with unclear origin to evaluate the user’s emotion, personality, ability, sleep status, professional interest, and interpersonal status.

Q&A Module

The Q&A community embodies the great advantages of mental health apps compared with traditional mental health services. Of the 40 apps, 27 (67.5%) provide this module. The Q&A community gathers other users and consultants on the same platform for rapid communication of mental health problems between users and between users and consultants, which is difficult to achieve by traditional psychological services. Users express their troubles, puzzles, or problems in the Q&A community, discuss and communicate with other users and consultants through Q&A feedback, solve problems, and gain knowledge.

Stress Relief

The stress relief module regulates the user’s mood, improves the sleep state, and relieves the user’s psychological pressure through proven ways, such as meditation and audio decompression. Of the 40 apps, 9 (22.5%) provide functional modules for stress relief. Meditation is considered a popular intervention method to relieve stress. Of these 9 apps, 7 (77.8%) provide functional modules to assist meditation. The modules guide stress relief training based on mindfulness or breathing technology (7/7, 100%) and cognitive behavioral therapy (1/7, 14.3%) in the form of audio or video. Audio is the main medium to help users relieve pressure. In addition to the audio used for meditation, other types of audio include light music (4/9, 44.4%), autonomous sensor meridian response audio (1/9, 11.1%), and nature recording (4/9, 44.4%). Additionally, 4 (44.4%) apps specifically provide audio to help users sleep.

MARS Evaluation

The overall MARS score showed high interreviewer reliability (intraclass correlation coefficient [ICC] 0.95, 95% CI 0.858-0.960). Simultaneously, all subscales also showed good consistency: engagement ICC 0.93 (95% CI 0.865-0.963), functionality ICC 0.71 (95% CI 0.462-0.847), aesthetics ICC 0.91 (95% CI 0.833-0.953), and information ICC 0.85 (95% CI 0.727-0.923).

The total average MARS score of all apps was 3.54 (SD 0.39), and the total score ranged from 2.96 (Emmasa Psychology) to 4.30 (Yi Psychology). The MARS score of 7 of 40 (17.5%) apps was ≥4. Furthermore, 30 of 40 (75%) apps had MARS scores ranging from 3.0 to 3.99. MARS scores of 3 of 40 (7.5%) apps ranged from 2.0 to 2.99. There were no apps with a score of <2.

The average scores of each subscale were as follows: information quality score=3.29 (SD 0.41), engagement quality score=3.37 (SD 0.51), aesthetic quality score=3.50 (SD 0.61), and functional quality score=3.97 (SD 0.37). The aesthetic quality score showed the largest span, with a minimum of 2.33 and a maximum of 4.50. The information quality part was the lowest, ranging from 2.50 to 4.00. The rating distribution of overall quality and 4 subscale dimensions is shown in Figure 4.
Figure 4. Graphical representation of the distribution of the MARS overall and subscale score. The median, the interquartile distance, and the range were given (N = 40). MARS: Mobile App Rating Scale.

The overall MARS score was significantly positively correlated with the scores of each subscale (r=0.62-0.88, P<.001). However, the user rating of the app market was not significantly correlated with the total MARS score (r=0.17, P=.33) and the scores of various scales (r=0.05-0.22, P=.21-.77; Table 3).

Combined with the professional background of app development, the total average MARS score of the app described as designed according to the opinions of mental health professionals is 3.66. The total average score of apps claiming to use proven psychotherapy theory is 3.77. The only app that clearly states that its usability is confirmed by peer-reviewed academic research has a total MARS score of 3.95. The quality score for all of the above cases is higher than the overall average score for all apps. However, the total average score of apps without reference to app design involving relevant scientific basis is 3.40.

Table 3. Correlation between the Mobile App Rating Scale subscale and the overall score and the user star score.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Engagement, correlation (P value)</th>
<th>Functionality, correlation (P value)</th>
<th>Aesthetics, correlation (P value)</th>
<th>Information, correlation (P value)</th>
<th>Overall rating, correlation (P value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engagement</td>
<td>a</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Functionality</td>
<td>0.39 (.01)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aesthetics</td>
<td>0.67 (&lt;.001)</td>
<td>0.31 (.049)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information</td>
<td>0.73 (&lt;.001)</td>
<td>0.51 (.001)</td>
<td>0.65 (&lt;.001)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall rating</td>
<td>0.88 (&lt;.001)</td>
<td>0.62 (&lt;.001)</td>
<td>0.86 (&lt;.001)</td>
<td>0.88 (&lt;.001)</td>
<td></td>
</tr>
<tr>
<td>User rating</td>
<td>0.22 (.21)</td>
<td>0.21 (.23)</td>
<td>0.08 (.64)</td>
<td>0.05 (.77)</td>
<td>0.17 (.33)</td>
</tr>
</tbody>
</table>

aNot applicable.
bApps with zero user star ratings were excluded.

Discussion

Principal Findings

This study identified 40 multipurpose mental health apps, understood their main functional distribution, and evaluated their content and quality.

The evaluation of the professional background of app development provides an opportunity for app developers to improve the scientificity and accuracy of these apps. Among all included apps, only 2 (5%) are developed by professional organizations engaged in mental health services, while most of the apps (38/40, 95%) are developed by commercial companies or individuals without mental health–related backgrounds. Additionally, more than half of the app development background lacks the participation of professionals or scientific theories. In the description of the apps, only 1 (2.5%) clearly stated that its availability had been confirmed by peer-reviewed academic research. The absence of the development process in the app description may raise questions about the credibility of the apps [36]. Simultaneously, the lack of a professional development background may lead to an inappropriate final app, which is considered a potential threat to users. Previous evidence also emphasizes that the development process should include the participation of health care professionals and target users with regard to apps that provide health content and collect health data [37]. Therefore, these findings emphasize the need to take action to ensure the scientific quality of mental health apps, which will improve the reliability and quality of the content provided by the apps.
The in-depth analysis revealed that the most common functional combination of multipurpose mental health apps is psychological education, counseling, self-assessment, and the Q&A community. Only 1 (2.5%) app contains all functional modules, and its quality score is also good. In all apps, psychological education, self-assessment, and counseling occupy the main positions. However, the online consultation function of most apps is provided by professional mental health care personnel. Therefore, the problem of insufficient mental health service personnel still exists. However, we found in this search an app, Xiaoxin Psychology, that provides the possibility of solving this problem. Xiaoxin Psychology combines artificial intelligence with mental health to replace mental health service personnel. Although we have not found a test of the effectiveness of the app, studies have confirmed the effectiveness of evidence-based computerized interventions in alleviating anxiety and depression in adults [38].

Among the 40 apps identified, except “grape heart,” which is an app for children with autism, no app for a specific mental disorder was found. This is different from foreign mental health apps [39-41]. However, we found that the included apps classify the service theme and set up navigation in the menu for quick access. We observed that the obstacles of love and marriage, emotion, and parent-child education are the most common. This may be related to the traditional Chinese concept of paying attention to family emotion. COVID-19 has increased the demand for mental health services, but only 3 (7.5%) of all assessment apps have added psychological aid plates. This may be due to an untimely update of the current apps. App developers set content classification modules on the basis of current events and hot spots, which is a good way to attract new users and stabilize old users.

The 40 apps’ choice of target population also has specific characteristics. Only 2 (5%) of the apps are designed specifically for teenagers or children, and there are no apps designed for women. App developers prefer ordinary adult users, which may be because it is easier to obtain users and maintain the stability of users. However, adolescents are one of the most vulnerable to mental health problems [42]. They are often reluctant to seek professional help because of their sense of shame and tendency toward self-reliance [8]. The method of getting help based on a network provides a way to overcome these obstacles [43-45].

Although almost all apps provide privacy policies for the collection and use of users’ personal data, they lack detailed information about data storage, user management permissions, and the use of cookies. Moreover, in this evaluation, most apps lacked a description of the endpoint of data sharing, which is consistent with the results of previous studies [46]. The economic benefits of shared data promote the occurrence of such situations and pose a threat to user data disclosure [47]. However, personal health information is highly sensitive, and the disclosure of health information may cause varying degrees of negative effects and even death [48]. Additionally, people may refuse to use mHealth apps because of concerns about health data security and privacy [49,50]. Furthermore, concerns about privacy protection are exacerbated by people’s sense of shame about using mental health services [51-53]. However, the trust between app developers and users may be damaged by the lack of clarity of privacy policies, which results in the loss of potential long-term users [54]. Therefore, there is still a long way to go in terms of the compliance with privacy policy content and the pertinence to special types of apps (such as health).

The overall quality of the multipurpose mental health apps we reviewed is good. The MARS score ranges from 2.96 to 4.30. There is a gap in the quality of the apps, which is similar to that in previous studies [55]. However, the 3 (7.5%) apps with the highest MARS quality score also have the highest download frequency, indicating the attraction of high-quality apps to target users. Additionally, the apps show advantages in functional evaluation rather than the information part. This emphasizes that future app development should focus on improving the information quality of apps. Adding professional mental health care personnel to the app development process may be a feasible way. This will also provide a reference for users to evaluate the degree of expertise involved in the app development process before downloading [56]. In this study, apps with professional development backgrounds also reflected their advantages in quality scoring. The user star rating was not correlated with the total MARS quality score, which may indicate that there is a different structure between the user star rating and the app quality score. Similarly, in previous studies, it has been reported that app quality depends not only on the content but also on the function and design method of the content [57]. Furthermore, the star rating of the app market may involve the early version of the apps and cannot fully represent the current version, which may lead to distortion in the evaluation of the current version [58].

**Contribution**

This study conducted a specific survey on the content and quality of multipurpose mental health apps in China based on a systematic and evidence-based approach. To the best of our knowledge, this is the first attempt to incorporate the professional background of app development and the user privacy protection policy into the evaluation of mental health apps in China. These findings will assist app developers in enhancing current apps or design new apps. Additionally, through the analysis of privacy policies, users can better understand the potential risks of providing information to service providers. Furthermore, we found that the combination of artificial intelligence and mental health may provide the possibility of solving the problem of insufficient mental health service personnel. This will be the direction of designing and creating apps in the future.

**Limitations and Future Work**

There are some limitations to this study. We may have missed some apps. Keywords retrieval cannot exhaust all apps. Some apps that met the inclusion criteria were ignored because the title or description did not contain search criteria related to mental health. Moreover, the app market is constantly changing, new apps may be on the shelf at any time, and old apps may be deleted for various reasons. Although the research examines the apps’ emphasis on user privacy protection, we cannot verify whether the apps really implement privacy protection measures. Additionally, this study did not verify the scientificity of the content provided by the apps, and in-depth research will be
continued in the future. Two researchers independently screened the eligibility of the apps, extracted the characteristics of the apps, and used MARS to evaluate the quality of the apps. The rater’s reliability is good or excellent, but if more researchers participate, the results may be more objective. Hence, we will consider increasing the number of researchers in follow-up studies. Furthermore, some of our findings may reveal the direction of such research in the future. We found some behavior change techniques aimed at improving users’ mental health, such as meditation. Future research can further evaluate the quality and characteristics of behavior change techniques in these apps. In addition, this survey also identified some high-quality apps. Before widely recommending these apps, a further randomized controlled trial can be used to determine and compare their effectiveness.

**Conclusion**

This study identified 40 multipurpose mental health apps, analyzed their main functional distribution, and evaluated their content and quality. These findings will assist app developers in enhancing current apps or design new apps. The quality of multipurpose mental health apps in China’s main app markets is generally good. Most apps provide rich functionality and classify the service theme to set up navigation in the menu for quick access. However, the lack of professional background in the app development process raises concerns about the scientificity of the apps. Furthermore, the privacy protection policy of the apps also needs to be described in more detail.

**Conflicts of Interest**

None declared.

**References**


Russian-Language Mobile Apps for Reducing Alcohol Use: Systematic Search and Evaluation

Anna Bunova¹; Veronika Wiemker²; Boris Gornyi¹, PhD; Carina Ferreira-Borges³, PhD; Maria Neufeld³,4,5, MSc

¹National Research Center for Therapy and Preventive Medicine of the Ministry of Health of the Russian Federation, Moscow, Russian Federation
²Heidelberg Institute of Global Health, Medical Faculty and University Hospital, Heidelberg University, Heidelberg, Germany
³WHO European Office for the Prevention and Control of Noncommunicable Diseases, Moscow, Russian Federation
⁴Institute for Mental Health Policy Research, Centre for Addiction and Mental Health, Toronto, ON, Canada
⁵Institute of Clinical Psychology and Psychotherapy, Dresden University of Technology, Dresden, Germany

Corresponding Author:
Anna Bunova
National Research Center for Therapy and Preventive Medicine of the Ministry of Health of the Russian Federation
Petroverigskiy Pereulok 10
Moscow, 101990
Russian Federation
Phone: 7 9151416154
Email: asbunova@gmail.com

Abstract

Background: Personalized prevention tools such as mobile apps designed to reduce alcohol consumption are widespread in mobile app stores accessible in Russia. However, the quality and content of these mobile apps have not been systematically evaluated.

Objective: This study aimed to identify Russian-language mobile apps for reducing alcohol use and to evaluate their quality and potential to change alcohol-related health behavior. It further aimed to identify apps that could facilitate screening and brief interventions in primary health care in Russia.

Methods: A systematic search for mobile apps available in Russia was carried out between April 1 and 15, 2020, December 1 and 15, 2020, and in March 2021 in the iPhone App Store, Google Play Store, and the 4PDA forum. App quality was assessed using the Mobile App Rating Scale (MARS), and structured searches in electronic libraries and bibliographic databases were used to evaluate the apps’ evidence base. The number of features facilitating changes in lifestyle behavior was assessed using the App Behavior Change Scale (ABACUS).

Results: We identified 63 mobile apps for reducing alcohol use. The mean MARS quality ratings were high for the subscales of functionality (3.92 out of 5, SD 0.58) and aesthetics (2.96, SD 0.76) and low for engagement (2.42, SD 0.76) and information (1.65, SD 0.60). Additional searches in electronic libraries and bibliographic databases yielded no studies involving the identified apps. ABACUS scores ranged from 1 to 15 out of 25, with a mean of 5 (SD 3.24). Two of the identified apps might be useful for screening and brief interventions in Russian primary health care after improvements in content and scientific testing.

Conclusions: Russian-language mobile apps for reducing alcohol use are accessible in the app stores. Many of them are aesthetically pleasing, functional, and easy to use. However, information about their scientific trialing or testing is lacking. Most apps contain a low number of features that facilitate changes in lifestyle behavior. Further research should examine the context of Russian-language mobile apps for reducing alcohol use. Our findings underline the need to develop evidence-based apps to mitigate alcohol consumption in Russia and elsewhere.

Trial Registration: PROSPERO International Prospective Register of Systematic Reviews CRD42020167458; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=167458

(JMIR Mhealth Uhealth 2022;10(1):e31058) doi:10.2196/31058

KEYWORDS
alcohol; mHealth; mobile applications; screening and brief intervention; Mobile Application Rating Scale; App Behavior Change Scale; mobile phone
Introduction

Background

Alcohol is one of the leading risk factors contributing to the global burden of disease and mortality [1-3]. The World Health Organization (WHO) estimates that globally, about 3 million deaths are caused by alcohol use each year, almost 1 million of which occur in the WHO European Region as the region with the highest level of per capita alcohol consumption. Drinking alcohol contributes to the development of more than 200 diseases and injuries. It increases the risk of cardiovascular and digestive diseases, neoplasms, mental and behavioral disorders (not limited to alcohol use disorders) as well as violent crimes, suicides, and road traffic accidents [4,5]. The impact of alcohol on mortality in Russia has been well documented, and Russia remains one of the countries with the largest alcohol-attributable burden of diseases worldwide, although substantial improvements were made over the last decade [6-9]. The WHO has recently launched the SAFER initiative to reduce alcohol-related harm, which recommends that health services should provide prevention and treatment interventions to individuals and families at risk of or affected by alcohol use disorders and associated conditions [10]. One of the 5 high-impact interventions of SAFER is the screening and brief intervention (SBI) programs in primary health care (PHC) [10,11].

There is a large body of research supporting the effectiveness of SBI in reducing alcohol consumption and other alcohol-related outcomes [12-15]. Attempts to introduce SBI in the Russian PHC for patients at risk for harmful alcohol use began in 2013 when legislative changes allowed the establishment of SBI as a part of Russia’s dispensarization program within PHC facilities [16]. Dispensarization is a set of standardized measures in PHC that includes preventive medical examination for assessing the state of health and is carried out in relation to certain groups of the population in accordance with the legislation of the Russian Federation. Following the currently established provisions, dispensarization includes an evidence-based 2-step screening procedure aiming to provide early and timely detection of conditions and diseases as well as risk factors for their development, including the nonmedical use of drugs and psychotropic substances [17]. The introduced SBI as within the broader dispensarization framework consists of 2 steps [17]. In the first step, the self-administered 3-item version of the Alcohol Use Disorders Identification Test-Consumption (AUDIT-C) is used to detect patients at risk [17,18]. If the screening result is positive, that is, the score exceeds the sex-specific cut-off, the patient is asked to complete the full 10-item version of the Alcohol Use Disorders Identification Test (AUDIT) as part of an interview with a health care professional. The health care professional then provides a brief intervention depending on the results [17,19].

Despite the effectiveness of SBI, there are several barriers to their widespread implementation in PHC settings in Russia beyond the dispensarization framework [20-22]. The AUDIT is based on the concept of a standard drink, which was introduced in several countries worldwide as a measure of alcohol consumption and to provide information about alcohol consumption to consumers, which was also as part of communicating the number of standard drinks on labels of alcoholic beverages [19]. In practice, using the standard drink concept remains a challenge for PHC professionals. PHC professionals report that the concept is not understandable for patients and difficult to calculate with, especially for patients that engage in heavy episodic drinking [23,24]. Moreover, delivering a brief intervention requires specific skills and knowledge as well as additional time and resources from the PHC professionals [20,21].

The development of electronic systems to deliver or support SBIs can potentially address some of these challenges and support health care workers. For instance, electronic devices such as smartphones and tablets can be used instead of the traditional paper-and-pencil screening tests and facilitate counting standard drinks as part of the risk assessment and support the delivery of brief interventions. Moreover, electronic SBIs are also potentially more flexible and can be adapted to take into account the regional patterns of alcohol consumption and make the assessment more personalized. They can also potentially reach larger audiences beyond the health sector [25,26].

The growing popularity of mobile phones and the active development of mobile internet in all regions of Russia open up great opportunities for using mobile apps as tools to change individual health behavior [27]. Mobile apps can provide an additional resource for preventive interventions catering to at-risk populations. However, to be successful, such interventions require the health care professional to select effective, evidence-based, and field-proven mobile apps [28]. A study by Abroms and colleagues [29] showed that making such a choice is difficult since many mobile apps contain inaccuracies and low-quality information, are not tested in practice, or lack an evidence base. Abroms et al [29] point out the potential dangers of such apps, ranging from misinformation to misleading risk level estimates. A rigid evaluation of apps for reducing alcohol use is therefore of great interest to both alcohol consumers and health care professionals. While former studies have described the features of highly rated Russian-language apps for reducing alcohol use, they did not evaluate the app quality and the potential to change alcohol-related health behavior by using validated instruments [30,31]. By closing this gap and providing researchers and health care professionals with an overview of the currently available evidence-based apps for reducing alcohol use, this study may contribute to facilitating the provision of SBI programs in the Russian PHC.

Objective

The aim of this study was to conduct a systematic search and evaluation of Russian-language mobile apps for reducing alcohol use. The specific objectives were to (1) create an overview and establish a list of relevant apps available in Russia, (2) assess their overall quality and evidence base, and (3) evaluate if any of the available apps could be used to support the provision of AUDIT-based SBI in Russia and its broader implementation in PHC facilities.
Methods

Study Design
The study was performed in 2 steps. In step 1, we conducted a systematic app store search to identify the apps for reducing alcohol use, following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [32]. In step 2, we evaluated the identified apps by using the Mobile App Rating Scale (MARS) [33] and the App Behavior Change Scale (ABACUS) [34]. For specific steps of the rating procedure, please see below. The study protocol was published in PROSPERO (Prospective Register of Systematic Reviews) [35] under registration CRD42020167458 (review ongoing).

Step 1: Systematic App Store Searches

Eligibility Criteria
We defined mobile apps for reducing alcohol use as tools for tablets or smartphones that facilitate behavioral change related to alcohol use. We excluded apps that were clearly not aimed at the reduction of alcohol use, such as games, barcode scanners as part of the Unified State Automated Information System (EGAIS) tracking alcohol distribution and sales under the Russian Federal Service for Alcohol Market Regulation [36], recipes for alcoholic drinks, and wallpapers. Only currently available and working Russian-language apps with at least a basic or trial version free of charge were included.

Search Strategy
Six systematic searches in the iPhone App Store, Google Play Store, and a Russian internet forum of mobile apps, that is, 4PDA [37] were conducted between April 1 and 15, 2020, December 1 and 15, 2020, and in March 2021. Two native Russian speakers conducted these 6 independent searches on different dates to account for additional mobile apps that were created during the COVID-19 pandemic. Keywords included the Russian words for “alcohol,” “alcoholic drinks,” “spirits,” “beer,” “vodka,” “drink alcohol,” “alcohol calculator,” “alcohol tracker,” “sober,” “alcohol monitoring”; and “breathalyzer” (details in PROSPERO protocol [35] and Multimedia Appendix 1) and were entered through the general search bar of the app stores and the forum.

Screening and Selection of Apps

In the first step, we recorded the name, app icon, developer, store, platform, brief description, and URL of all the available alcohol-related mobile apps. Next, duplicates were removed, and app store descriptions were screened against inclusion criteria. We retained only 1 record if identical versions of an app were available for Android and iPhone operating systems (iOSs). All remaining apps were downloaded onto the study devices (Samsung Galaxy Tab A 7.0 SM-T285 8GB/Android version 9, Lenovo Tablet TB-X704L 64G/Android version 7, and iPhone 11/iOS version 14.0.1). Apps that could not be opened on these devices were excluded.

Data Extraction

The following information was extracted for all the included apps: app name, the app’s star rating on the platform, number of installations, developer, current version, number of ratings for current version, last update, existence of a basic version and paid premium versions, and platform. All included apps were available and all data were updated in the last week of March 2021.

Step 2: Evaluation of Mobile Apps

Measures/Rating Tools
We used 2 scales to rate the identified mobile apps. The MARS scale, assessing the quality of the mobile apps, contains 23 items across 5 subscales: engagement, functionality, aesthetics, information, and subjective quality [33]. Each item is rated on a 5-point scale from 1 (lowest quality) to 5 (highest quality). The overall app quality is assessed by calculating the mean scores of the first 4 subscales and the total mean score. The subjective quality score describes the raters’ personal liking of the app and should be reported separately if assessed. Subjective quality was not assessed in this study. The ABACUS scale, evaluating the apps’ potential to facilitate behavior change, contains 21 items across 4 subscales: knowledge and information, goals and planning, feedback and monitoring, and actions [34]. The total score is obtained by counting the number of items answered affirmatively. The MARS and the ABACUS showed good internal consistency and interrater reliability (MARS, $\alpha=0.92$; intraclass correlation coefficient [ICC]=0.85 and ABACUS, $\alpha=0.93$; ICC=0.92) [33,34].

A full evaluation of all the included apps was carried out by a first rater. A second rater independently evaluated a random sample of 30% (19/63) of the apps. Both raters were prepared for their task by completing a MARS video training tutorial [38]. To rate an app’s evidence base, as measured in MARS item 19, raters searched for randomized studies in the electronic libraries and bibliographic databases eLibrary, CyberLeninka, and Google Scholar by using the app’s name as a keyword as suggested by the MARS authors [38].

If the mobile app requested the input of demographic characteristics or consumption data, the following data were used: female gender, 30 years of age, body weight of 60 kg, height of 170 cm, and alcohol consumption on the last occasion as 200 ml of 40% vodka. If required, the maximum legal blood alcohol content was set to 0.3 ppm.

Classification of Apps and Criteria for Potential Use in SBI Programs

The identified apps were classified according to their main features. For this purpose, the following data were extracted: the app’s ability to estimate blood alcohol concentration and sobering time, its ability to record personal alcohol consumption, the presence of SBI elements, the presence of a “soberity counter” to count the time since the last drinking occasion, and the app’s ability to support the reduction of alcohol use in a structured way. To investigate the potential of the available apps for supporting the provision of SBI in Russia, app descriptions and main features were reviewed against 2 criteria: (1) availability of AUDIT (2) whether the app provided any type of brief intervention.
Data Analysis

Statistical analysis and data visualization were carried out in Excel (Microsoft Excel for Office 365) and SPSS Statistics 20 (IBM Corp). Measures of interrater reliability were obtained by calculating the ICCs for all MARS and ABACUS subscales [39], using a 2-way mixed effects and average measures model with absolute agreement [40]. Descriptive analysis included total sample size, percentage, median, mean, and standard deviation.

Results

Systematic App Store Searches

A total of 620 alcohol-related apps were identified through keyword searches in the iOS App Store, Google Play Store, and 4PDA (Figure 1). After removing duplicates, 310 apps were screened against inclusion criteria, leaving 65 apps for reducing alcohol use for further download and evaluation. Among the downloaded apps, 2 had to be excluded as they did not work properly or required connection to a breathalyzer. Finally, 63 apps were included for evaluation, 51 of which were available only in the Google Play Store and 5 only in the iOS App Store. Only 7 apps were available in both stores (Multimedia Appendix 2).

Overview of the Included Apps for Reducing Alcohol Use

The included apps were grouped into 6 categories according to their main feature: apps estimating blood alcohol concentration and sobering time (n=29), apps recording personal alcohol consumption (n=15), apps for SBIs (n=2), apps counting the time since the last drinking occasion (sobriety counters, n=8), apps with structured support to reduce alcohol use (n=4), and other apps for reducing alcohol use (n=5). Most apps were designed for Android systems (n=51); a minority were available for iOS (n=5) or both operation systems (n=7). A total of 19 apps were created by commercial organizations; 1 app was developed by a science center [41]. In 43 cases, no further information on the developer's legal status could be obtained. Most of the apps (53/63, 84%) had last been updated between 2016 and 2021. A minority (n=16) offered a paid upgrade version featuring the removal of advertisements and the use of additional features. According to Google Play statistics, the median number of installations was 10,000. No comparable information was provided in the iOS App Store and 4PDA. The median star rating of all apps in Google Play and iOS App Store was 3.9, based on a median of 81 ratings. Apps with structured support to reduce alcohol consumption were downloaded most often, with a median of 300,000 installations. The median star rating of the apps in this category was 4.8, based on a median of 4418 ratings.

Apps Estimating Blood Alcohol Concentration and Sobering Time

The main feature of this group of apps (n=29) was the estimation of the maximum blood alcohol concentration and sobering time. Most apps (n=14) were based on Widmark’s equation [42], 2 apps used Watson’s equation [43], and 13 apps provided no
information about the method of calculation. Seven apps only allowed to calculate the maximum blood alcohol concentration (n=4) or sobering time (n=3). The other 22 apps provided a combination of both mentioned features (n=16) or offered additional features such as the estimation of “no-driving” time after drinking (n=10), the estimation of the maximum alcohol consumption to sober up by a certain time (n=2), a drinking diary (n=2), or the option to unlock achievements for reducing alcohol consumption (n=1).

**Apps Recording Personal Alcohol Consumption**

Apps in this group (n=15) provided detailed drinking diaries (n=4), consumption calendars allowing users to indicate on what days they drank alcohol (n=9), or both functions combined (n=2), featuring statistics of consumption per day, week, month, year. In some apps (n=4), users could calculate costs related to their alcohol use and see how much money they saved by cutting down their consumption.

**Apps for SBI**

We found only 2 Russian-language apps fulfilling our criteria to potentially facilitate SBI. Both apps allowed users to complete the AUDIT. The first app provided detailed instructions for brief interventions aimed at health care professionals and a standard drink calculator allowing users to choose consumed alcoholic drinks and calculate the number of standard drinks consumed. The second app provided information on the individual level of risk and alcohol-related harm according to AUDIT results.

**Apps Counting the Time Since the Last Drinking Occasion (Sobriety Counters)**

We identified 8 apps in this group. Two of them consisted of a simple timer, counting the time since the last drinking occasion. The other 6 apps included additional motivation components such as a progress bar and achievements to be obtained (n=4) or inspiring citations (n=2). Two apps featured a chat where users could share their experiences. One app allowed users to consult with a medical specialist and to observe positive changes connected to alcohol abstinence in the physical appearance of the visualized avatars.

**Apps With Structured Support to Reduce Alcohol Use**

A total of 4 apps featured structured support to help users quit drinking or to reduce their alcohol consumption. Most of these apps provided a plan with daily tasks (n=3), a sobriety counter (n=4), a drinking diary (n=2), and motivation components. Motivation components included a progress bar and achievements to be obtained (n=3), inspiring articles or citations (n=4), daily notifications (n=4), encouraging pictures or videos (n=2), and a visualization of the positive health consequences of alcohol abstinence (n=2). One app had a community chat where users shared their experience of reducing consumption or quitting alcohol. One app provided a blood alcohol concentration calculator. Three apps allowed users to complete the AUDIT (n=2) or the Michigan Alcohol Screening Test (n=1) [44].

**Other Apps for Reducing Alcohol Use**

Five apps could not be assigned to any of the aforementioned categories. These included an app for audio hypnosis, an app allowing to record withdrawal symptoms, an app featuring notifications about alcohol-related harm, an app allowing to estimate the dose of consumption needed to relax, to get drunk or to have fun, and an app for counting unplanned alcohol drinking occasions after quitting drinking.

**Behavior Change Techniques Featured in the Apps**

The number of behavior change features provided by each app as reflected in ABACUS scores (Multimedia Appendix 2) ranged from 1 to 15 out of 21, with a mean of 5 points (SD 3.24). A great majority (54/63, 86%) of the apps requested individual baseline information and 71% (45/63) of the apps provided (individualized) user feedback. Many apps allowed the user to self-monitor their behavior (36/63, 57%) and customize or personalize certain app features (32/63, 51%). Table 1 shows the frequencies of the 21 behavioral change features evaluated in the apps. All ABACUS scores showed high interrater reliability (2-way mixed ICC=0.96; 95% CI 0.90-0.98).

The largest number of behavior change techniques was found in the categories of apps with structured support to reduce alcohol use and apps counting the time since the last drinking occasion (sobriety counters). Out of the 12 apps in these 2 groups, 9 apps featured more than 7 behavior change techniques. Apps estimating blood alcohol concentration and sobering time provided the lowest number of behavior change techniques, with an average ABACUS score of 2.38 (SD 1.37).
Table 1. Behavioral change features in the apps for reducing alcohol use (N=63).

<table>
<thead>
<tr>
<th>Behavioral change feature</th>
<th>Apps providing the feature, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Knowledge and information</strong></td>
<td></td>
</tr>
<tr>
<td>Ability to customize and personalize features</td>
<td>32 (51)</td>
</tr>
<tr>
<td>Consistency with national guidelines or created with expertise</td>
<td>3 (5)</td>
</tr>
<tr>
<td>Request for baseline information</td>
<td>54 (86)</td>
</tr>
<tr>
<td>Instruction on how to perform the behavior</td>
<td>6 (10)</td>
</tr>
<tr>
<td>Information about the consequences of continuing or discontinuing behavior</td>
<td>22 (35)</td>
</tr>
<tr>
<td><strong>Goals and planning</strong></td>
<td></td>
</tr>
<tr>
<td>Request for willingness for behavior change</td>
<td>6 (10)</td>
</tr>
<tr>
<td>Setting of goals</td>
<td>7 (11)</td>
</tr>
<tr>
<td>Ability to review goals, update, and change when necessary</td>
<td>6 (10)</td>
</tr>
<tr>
<td><strong>Feedback and monitoring</strong></td>
<td></td>
</tr>
<tr>
<td>Ability to quickly and easily understand the difference between current action and future goals</td>
<td>9 (14)</td>
</tr>
<tr>
<td>Ability to allow the user to easily self-monitor behavior</td>
<td>36 (57)</td>
</tr>
<tr>
<td>Ability to share behaviors with others or allow for social comparison</td>
<td>12 (19)</td>
</tr>
<tr>
<td>Ability to give the user feedback—either from a person or automatically</td>
<td>45 (71)</td>
</tr>
<tr>
<td>Ability to export data from app</td>
<td>4 (6)</td>
</tr>
<tr>
<td>Material or social reward or incentive</td>
<td>7 (11)</td>
</tr>
<tr>
<td>General encouragement</td>
<td>12 (19)</td>
</tr>
<tr>
<td><strong>Actions</strong></td>
<td></td>
</tr>
<tr>
<td>Reminders or prompts or cues for activity</td>
<td>13 (21)</td>
</tr>
<tr>
<td>App encourages positive habit formation</td>
<td>7 (11)</td>
</tr>
<tr>
<td>App allows or encourages for practice or rehearsal in addition to daily activities</td>
<td>1 (2)</td>
</tr>
<tr>
<td>Opportunity to plan for barriers</td>
<td>4 (6)</td>
</tr>
<tr>
<td>Assistance with or suggest restructuring the physical or social environment</td>
<td>1 (2)</td>
</tr>
<tr>
<td>Assistance with distraction or avoidance</td>
<td>3 (5)</td>
</tr>
</tbody>
</table>

**App Quality**

Multimedia Appendix 2 shows the subscale and overall MARS scores of all evaluated apps for reducing alcohol use. Interrater reliability was high (2-way mixed ICC=0.96; 95% CI 0.91-0.98). The average overall MARS score of all reviewed apps was 2.74 (SD 0.52). MARS item 19 “evidence base” reached the lowest mean score (0.00)—no information about scientific trialing or testing of the identified app could be obtained. Other low mean scores included quantity of information (0.75, SD 1.24), quality of information (0.94, SD 1.47), credibility (1.17, SD 0.58), visual information (1.40, SD 1.75), interest (1.97, SD 0.97), and entertainment (1.89, SD 0.99). The highest scores were obtained for gestural design (4.05, SD 0.63), ease of use (3.98, SD 0.66), accuracy of app description (3.94, SD 0.56), and target group (3.89, SD 0.84). The average score of all MARS items is shown in Figure 2. A total of 30% (19/63) of the evaluated apps reached an overall MARS score of ≥3.0. The categories of apps with strong support to reduce alcohol use and apps for SBIs reached the highest overall scores (3.74 [SD 0.18] and 3.42 [SD 0.32], respectively).
Out of the 63 identified mobile apps for reducing alcohol use, 4 apps contained the AUDIT, which is widely used in Russian PHC SBI. However, only 2 apps contained additional SBI elements and thus fulfilled both selection criteria. The 2 apps providing only the AUDIT (“I do not drink!” and “Sober One”) contained obvious errors. The I do not drink! app has a “urban” translation into Russian and contains only 9 out of the 10 AUDIT questions. The Sober One app is potentially more attractive for SBI as it provides a brief risk assessment. However, 1 standard dose is determined as 13.7 g of pure alcohol, which does not correspond to the value officially used in Russia [17,45]. Out of the 2 apps fulfilling both selection criteria (AUDIT and Alcoholism test), the AUDIT app applies the SBI algorithms from the official Russian guidelines [17]. The second app, Alcoholism test, evaluates the individual health risk and provides additional information about the harm associated with alcohol consumption, which may motivate a conversation with the patient. However, we did not find any information on the scientific testing or trialing of the identified apps.

Discussion

Principal Findings

This study is the first systematic search and evaluation of Russian-language mobile apps for reducing alcohol use in Russia. We identified and assessed 63 eligible apps, 2 of which could potentially be used in SBIs in Russian PHC facilities after improvements in content and scientific testing. The MARS app quality scores of the evaluated apps showed good functionality, aesthetics, and ease of use. However, there is ample room for improvement, especially in the area of scientific support and evidence base; no information on scientific trialing or testing of any of the apps could be obtained. Further, ABACUS scores indicated that most apps provide only few features to facilitate human behavior change, casting doubt on their effectiveness to change alcohol consumption habits. These weaknesses seem to be common not only in Russian-language apps but in comparable apps worldwide. In a recent Australian study using both MARS and ABACUS, English-language apps for reducing alcohol use obtained similar ratings as the apps evaluated in this study [46].

The analysis of the apps’ main features revealed some specific weaknesses and strong points of different app categories. Apps for calculating blood alcohol concentration and sobering time mostly used the Widmark formula developed in 1932 [42]. Out of the 2 apps fulfilling both selection criteria (AUDIT and Alcoholism test), the AUDIT app applies the SBI algorithms from the official Russian guidelines [17]. The second app, Alcoholism test, evaluates the individual health risk and provides additional information about the harm associated with alcohol consumption, which may motivate a conversation with the patient. However, we did not find any information on the scientific testing or trialing of the identified apps.

App Potential for SBI in the Russian PHC

This study is the first systematic search and evaluation of Russian-language mobile apps for reducing alcohol use in Russia. We identified and assessed 63 eligible apps, 2 of which could potentially be used in SBIs in Russian PHC facilities after improvements in content and scientific testing. The MARS app quality scores of the evaluated apps showed good functionality, aesthetics, and ease of use. However, there is ample room for improvement, especially in the area of scientific support and evidence base; no information on scientific trialing or testing of any of the apps could be obtained. Further, ABACUS scores indicated that most apps provide only few features to facilitate human behavior change, casting doubt on their effectiveness to change alcohol consumption habits. These weaknesses seem to be common not only in Russian-language apps but in comparable apps worldwide. In a recent Australian study using both MARS and ABACUS, English-language apps for reducing alcohol use obtained similar ratings as the apps evaluated in this study [46].

The analysis of the apps’ main features revealed some specific weaknesses and strong points of different app categories. Apps for calculating blood alcohol concentration and sobering time mostly used the Widmark formula developed in 1932 [42]. Out of the 2 apps fulfilling both selection criteria (AUDIT and Alcoholism test), the AUDIT app applies the SBI algorithms from the official Russian guidelines [17]. The second app, Alcoholism test, evaluates the individual health risk and provides additional information about the harm associated with alcohol consumption, which may motivate a conversation with the patient. However, we did not find any information on the scientific testing or trialing of the identified apps.

Discussion

Principal Findings

This study is the first systematic search and evaluation of Russian-language mobile apps for reducing alcohol use in Russia. We identified and assessed 63 eligible apps, 2 of which could potentially be used in SBIs in Russian PHC facilities after improvements in content and scientific testing. The MARS app quality scores of the evaluated apps showed good functionality, aesthetics, and ease of use. However, there is ample room for improvement, especially in the area of scientific support and evidence base; no information on scientific trialing or testing of any of the apps could be obtained. Further, ABACUS scores indicated that most apps provide only few features to facilitate human behavior change, casting doubt on their effectiveness to change alcohol consumption habits. These weaknesses seem to be common not only in Russian-language apps but in comparable apps worldwide. In a recent Australian study using both MARS and ABACUS, English-language apps for reducing alcohol use obtained similar ratings as the apps evaluated in this study [46].

The analysis of the apps’ main features revealed some specific weaknesses and strong points of different app categories. Apps for calculating blood alcohol concentration and sobering time mostly used the Widmark formula developed in 1932 [42]. Out of the 2 apps fulfilling both selection criteria (AUDIT and Alcoholism test), the AUDIT app applies the SBI algorithms from the official Russian guidelines [17]. The second app, Alcoholism test, evaluates the individual health risk and provides additional information about the harm associated with alcohol consumption, which may motivate a conversation with the patient. However, we did not find any information on the scientific testing or trialing of the identified apps.

Discussion

Principal Findings

This study is the first systematic search and evaluation of Russian-language mobile apps for reducing alcohol use in Russia. We identified and assessed 63 eligible apps, 2 of which could potentially be used in SBIs in Russian PHC facilities after improvements in content and scientific testing. The MARS app quality scores of the evaluated apps showed good functionality, aesthetics, and ease of use. However, there is ample room for improvement, especially in the area of scientific support and evidence base; no information on scientific trialing or testing of any of the apps could be obtained. Further, ABACUS scores indicated that most apps provide only few features to facilitate human behavior change, casting doubt on their effectiveness to change alcohol consumption habits. These weaknesses seem to be common not only in Russian-language apps but in comparable apps worldwide. In a recent Australian study using both MARS and ABACUS, English-language apps for reducing alcohol use obtained similar ratings as the apps evaluated in this study [46].

The analysis of the apps’ main features revealed some specific weaknesses and strong points of different app categories. Apps for calculating blood alcohol concentration and sobering time mostly used the Widmark formula developed in 1932 [42]. Out of the 2 apps fulfilling both selection criteria (AUDIT and Alcoholism test), the AUDIT app applies the SBI algorithms from the official Russian guidelines [17]. The second app, Alcoholism test, evaluates the individual health risk and provides additional information about the harm associated with alcohol consumption, which may motivate a conversation with the patient. However, we did not find any information on the scientific testing or trialing of the identified apps.

Discussion

Principal Findings

This study is the first systematic search and evaluation of Russian-language mobile apps for reducing alcohol use in Russia. We identified and assessed 63 eligible apps, 2 of which could potentially be used in SBIs in Russian PHC facilities after improvements in content and scientific testing. The MARS app quality scores of the evaluated apps showed good functionality, aesthetics, and ease of use. However, there is ample room for improvement, especially in the area of scientific support and evidence base; no information on scientific trialing or testing of any of the apps could be obtained. Further, ABACUS scores indicated that most apps provide only few features to facilitate human behavior change, casting doubt on their effectiveness to change alcohol consumption habits. These weaknesses seem to be common not only in Russian-language apps but in comparable apps worldwide. In a recent Australian study using both MARS and ABACUS, English-language apps for reducing alcohol use obtained similar ratings as the apps evaluated in this study [46].

The analysis of the apps’ main features revealed some specific weaknesses and strong points of different app categories. Apps for calculating blood alcohol concentration and sobering time mostly used the Widmark formula developed in 1932 [42]. Out of the 2 apps fulfilling both selection criteria (AUDIT and Alcoholism test), the AUDIT app applies the SBI algorithms from the official Russian guidelines [17]. The second app, Alcoholism test, evaluates the individual health risk and provides additional information about the harm associated with alcohol consumption, which may motivate a conversation with the patient. However, we did not find any information on the scientific testing or trialing of the identified apps.

Discussion

Principal Findings

This study is the first systematic search and evaluation of Russian-language mobile apps for reducing alcohol use in Russia. We identified and assessed 63 eligible apps, 2 of which could potentially be used in SBIs in Russian PHC facilities after improvements in content and scientific testing. The MARS app quality scores of the evaluated apps showed good functionality, aesthetics, and ease of use. However, there is ample room for improvement, especially in the area of scientific support and evidence base; no information on scientific trialing or testing of any of the apps could be obtained. Further, ABACUS scores indicated that most apps provide only few features to facilitate human behavior change, casting doubt on their effectiveness to change alcohol consumption habits. These weaknesses seem to be common not only in Russian-language apps but in comparable apps worldwide. In a recent Australian study using both MARS and ABACUS, English-language apps for reducing alcohol use obtained similar ratings as the apps evaluated in this study [46].

The analysis of the apps’ main features revealed some specific weaknesses and strong points of different app categories. Apps for calculating blood alcohol concentration and sobering time mostly used the Widmark formula developed in 1932 [42]. Out of the 2 apps fulfilling both selection criteria (AUDIT and Alcoholism test), the AUDIT app applies the SBI algorithms from the official Russian guidelines [17]. The second app, Alcoholism test, evaluates the individual health risk and provides additional information about the harm associated with alcohol consumption, which may motivate a conversation with the patient. However, we did not find any information on the scientific testing or trialing of the identified apps.

Discussion

Principal Findings

This study is the first systematic search and evaluation of Russian-language mobile apps for reducing alcohol use in Russia. We identified and assessed 63 eligible apps, 2 of which could potentially be used in SBIs in Russian PHC facilities after improvements in content and scientific testing. The MARS app quality scores of the evaluated apps showed good functionality, aesthetics, and ease of use. However, there is ample room for improvement, especially in the area of scientific support and evidence base; no information on scientific trialing or testing of any of the apps could be obtained. Further, ABACUS scores indicated that most apps provide only few features to facilitate human behavior change, casting doubt on their effectiveness to change alcohol consumption habits. These weaknesses seem to be common not only in Russian-language apps but in comparable apps worldwide. In a recent Australian study using both MARS and ABACUS, English-language apps for reducing alcohol use obtained similar ratings as the apps evaluated in this study [46].

The analysis of the apps’ main features revealed some specific weaknesses and strong points of different app categories. Apps for calculating blood alcohol concentration and sobering time mostly used the Widmark formula developed in 1932 [42]. Out of the 2 apps fulfilling both selection criteria (AUDIT and Alcoholism test), the AUDIT app applies the SBI algorithms from the official Russian guidelines [17]. The second app, Alcoholism test, evaluates the individual health risk and provides additional information about the harm associated with alcohol consumption, which may motivate a conversation with the patient. However, we did not find any information on the scientific testing or trialing of the identified apps.
background available in the App Store and the Google Play Store [50].

Some international websites such as iMedicalApps offer expert comments and reviews of medical apps to patients and health care professionals [51]. Unfortunately, none of these sites are available in Russian language. SBI for alcohol consumption is not yet broadly established and implemented at the level of PHC in Russia, although decisive action was taken to change this in the past 8 years [52]. Electronic SBIs might offer greater flexibility, and depending on their mode of implementation, potentially more anonymity to avoid stigma for PHC patients [26]. Their wide availability in the App Store and, more importantly, Google Play Store as the most popular marketplace for mobile apps in Russia, may offer new opportunities to expand personalized medical care for people with alcohol-related problems [53]. The use of mobile apps to facilitate the assessment of alcohol intake as well as the level of according risk is a promising approach and requires further study, especially in a country like the Russian Federation that is committed to implement SBIs as a routine procedure.

Limitations
Searches were carried out in the iOS App Store, Google Play Store, and the 4PDA forum. These stores regularly update their content, meaning that mobile apps may become unavailable over time. Furthermore, search options such as language and region settings affect the selection and order of results, thereby reducing the reproducibility of the searches. App contents were not analyzed in detail nor did we assess the apps’ potential to change human behavior in the long term. This assessment may represent an area for future research.

Conclusions
This study provides a structured overview of the main features, quality, and potential to change the alcohol-related health behavior of Russian-language apps for reducing alcohol use currently available in Russia. This overview can be used as a reference by alcohol consumers and health care professionals alike when choosing an app to facilitate the reduction of alcohol use. Although Russian-language apps for reducing alcohol use were found to be aesthetically pleasing, functional, and easy to use, most apps contained a low number of features that facilitate changes in lifestyle behavior and lacked information about scientific trialing or testing. Only 2 identified apps contained the AUDIT and additional brief intervention elements and could thus potentially be used for SBI in the Russian PHC after rigorous scientific evaluation of their effectiveness. Overall, our findings underline the need to develop evidence-based apps to mitigate alcohol consumption in Russia and elsewhere.

Authors’ Contributions
AB, MN, and BG conducted the app search and evaluation. AB and VW wrote the manuscript, and BG, CFB, and MN revised the manuscript.

Conflicts of Interest
None declared.

Multimedia Appendix 1
PROSPERO (Prospective Register of Systematic Reviews) protocol.
[DOCX File, 26 KB - mhealth_v10i1e31058_app1.docx ]

Multimedia Appendix 2
Russian-language mobile apps for reducing alcohol use.
[DOCX File, 1298 KB - mhealth_v10i1e31058_app2.docx ]

References


27. DIGITAL 2020 Your ultimate guide to the evolving digital world. We Are Social. URL: https://wearesocial.com/digital-2020 [accessed 2021-12-13]


37. 4PDA forum. 4PDA. URL: https://4pda.ru/forum/index.php?act=idx [accessed 2021-12-13]


41. Bunova A, Gornyi B. Razrabotka mobil'nogo prilozheniya dlya provedeniya skrininga i profilakticheskogo konsul'tirovaniya patsiyentov s riskom pagubnogo potrebleniya alkogolya (Development of a mobile application for screening and consultation of patients with harmful alcohol consumption). In: Profilakticheskaya Meditsina. 2019 Presented at: Non-communicable diseases and public health in Russia; September 16; Moscow p. 38.


Abbreviations

ABACUS: App Behavior Change Scale
AUDIT: Alcohol Use Disorders Identification Test
AUDIT-C: Alcohol Use Disorders Identification Test-Consumption
ICC: intraclass correlation coefficient
iOS: iPhone operating system
MARS: Mobile App Rating Scale
PHC: primary health care
PROSPERO: Prospective Register of Systematic Reviews
SBI: screening and brief intervention
WHO: World Health Organization

©Anna Bunova, Veronika Wiemker, Boris Gornyi, Carina Ferreira-Borges, Maria Neufeld. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 10.01.2022. 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.
Prioritization of Quality Principles for Health Apps Using the Kano Model: Survey Study

Abstract

Background: Health apps are often used without adequately taking aspects related to their quality under consideration. This may partially be due to inadequate awareness about necessary criteria and how to prioritize them when evaluating an app.

Objective: The aim of this study was to introduce a method for prioritizing quality attributes in the mobile health context. To this end, physicians were asked about their assessment of nine app quality principles relevant in health contexts and their responses were used as a basis for designing a method for app prioritization. Ultimately, the goal was to aid in making better use of limited resources (eg, time) by assisting with the decision as to the specific quality principles that deserve priority in everyday medical practice and those that can be given lower priority, even in cases where the overall principles are rated similarly.

Methods: A total of 9503 members of two German professional societies in the field of orthopedics were invited by email to participate in an anonymous online survey over a 1-month period. Participants were asked to rate a set of nine app quality principles using a Kano survey with functional and dysfunctional (ie, positively and negatively worded) questions. The evaluation was based on the work of Kano (baseline), supplemented by a self-designed approach.

Results: Among the 9503 invited members, 382 completed relevant parts of the survey (return rate of 4.02%). These participants were equally and randomly assigned to two groups (test group and validation group, n=191 each). Demographic characteristics did not significantly differ between groups (all P>.05). Participants were predominantly male (328/382, 85.9%) and older than 40 years (290/382, 75.9%). Given similar ratings, common evaluation strategies for Kano surveys did not allow for conclusive prioritization of the principles, and the same was true when using the more elaborate approach of satisfaction and dissatisfaction indices following the work of Timko. Therefore, an extended, so-called “in-line-of-sight” method was developed and applied for this evaluation. Modified from the Timko method, this approach is based on a “point of view” (POV) metric, which generates a ranking coefficient. Although the principles were previously almost exclusively rated as must-be (with the exception of resource efficiency), which was not conducive to their prioritization, the new method applied from the must-be POV resulted in identical rankings for the test and validation groups: (1) legal conformity, (2) content validity, (3) risk adequacy, (4) practicality, (5) ethical soundness, (6) usability, (7) transparency, (8) technical adequacy, and (9) resource efficiency.

Conclusions: Established survey methodologies based on the work of Kano predominantly seek to categorize the attributes to be evaluated. The methodology presented here is an interesting option for prioritization, and enables focusing on the most important criteria, thus saving valuable time when reviewing apps for use in the medical field, even with otherwise largely similar categorization results. The extent to which this approach is applicable beyond the scenario presented herein requires further investigation.

(JMIR Mhealth Uhealth 2022;10(1):e26563) doi:10.2196/26563
KEYWORDS
Kano; quality principles; mobile apps; physicians; surveys and questionnaires; evaluation studies; mHealth; health apps

Introduction

Background

Independent of their proficiency with apps and the respective usage contexts, users are often unfamiliar with the intricacies of the specific aspects that are essential for recognizing an app’s quality. Even apps covering health contexts are often marketed without having been evaluated by experts, and with only minimally relevant and reliant information being provided (eg, regarding scientific studies [1,2]). Thus, for end users, making an informed decision about whether or not to use an app is not an easy task, independent of whether they are health care professionals, patients with chronic conditions, or even laypeople with a more generic interest in health apps.

There are numerous, more or less elaborate, tools, norms, and lists of quality criteria that either target developers or aim at aiding those interested in an app in their decision process (eg, [3-8]), and many of the aspects they cover overlap. However, even if interested parties are aware of these approaches, if a quick assessment is desired, these approaches may sometimes be seen as going too far or being too complex. Both the paucity of readily available information or expert assessments [1,9] in identifying apps that can be recognized as trustworthy, as well as the difficulty in identifying suitable criteria for an initial and independent assessment, can mean that apps often fail to realize the potential attributed to them for medical care and prevention [10-13]. Checklists that interested users may apply to apps (eg, [4,14,15]) often target careful curation of a list of apps for later use, but may be too extensive for practical application and quick assessments in everyday medical practice. It may therefore be helpful to develop and apply a process for identifying a subset of criteria or quality principles listed in such tools considered to be particularly relevant for a specific target group, which may be achieved by means of prioritization.

As a foundation for this study, we used nine basic quality principles for health apps that were previously compiled [16,17] and evaluated [18,19] in a multistep process: (1) practicality, (2) risk adequacy, (3) ethical soundness, (4) legal conformity, (5) content validity, (6) technical adequacy, (7) usability, (8) resource efficiency, and (9) transparency. Participants in both of the aforementioned evaluation studies were first requested to provide initial assessments regarding the perceived relevance of these principles. They were then provided with applied app store descriptions and asked to determine whether they deemed the textual information sufficient to satisfy the above principles. Subsequently, they were asked to apply 25 questions operationalizing the nine principles to the same store descriptions. Between each of the steps, they were asked whether or not they would consider using the respective app based on the available information. During the course of these studies [18,19], as participants familiarized themselves with the quality criteria, they were able to make a more confident, but increasingly critical, assessments of the apps based on the available information.

These previous studies with medical students [18] and members of the German Society for Internal Medicine [19] showed that the participants predominantly perceived all nine of the above quality principles as important. For both studies, the data were evaluated using two (randomly assigned and equally sized) test and validation groups [18,19]. Although there were no significant differences in the answers obtained for the nine principles between the two groups, solely based on assigned relevance, rankings (and thus any prioritizations based on them) would have differed [18] between the groups as well as between the two studies. Apart from slightly lower relevance ratings for resource efficiency, all other quality principles were seen as either “important” or “very important”; however, owing to their closeness with respect to the ratings, any order of the principles based on these ratings seemed to have been influenced by statistical noise rather than sound calculations. Nevertheless, in both of the aforementioned studies [18,19], some participants expressed fear that the application of even these few principles would be too time-consuming for use in an everyday care context. As even the relevance-related questions for the criteria that were asked in these studies did not allow for their ranking, we therefore aimed to establish a method that would meet the demand for a better focus on quality aspects for mobile health (mHealth) apps that would be perceived as particularly relevant in the community.

Hypotheses

We hypothesized that methods established to assess product attributes in marketing-related research might also be suitable for categorizing quality attributes for mHealth apps. We tested this hypothesis based on an exemplary Kano survey related to the nine aforementioned quality principles. In this type of survey, questions are implemented based on a model developed by Noriaki Kano in the 1970s and 1980s. The “Kano model” is often used in the context of marketing or for refining products, specifically with regard to customer satisfaction with a product’s features in mind [20]. As Kano noted, there need not be a linear relationship between satisfaction or dissatisfaction and the fulfillment of a need [21]; thus, to be able to nevertheless assess a product, he proposed using so-called “functional” and “dysfunctional” questions that not only assess a participant’s opinion about a feature being available but also about it not being provided.

On its own, if successful at all, such a Kano survey–based categorization can only provide a rough prioritization at best, based on ranking the categories according to their fitness for the question at hand. As this approach may fail in cases where the attributes under consideration are rated similarly, we established our second hypothesis that it should be possible to nevertheless prioritize the product attributes studied (in our case, the nine quality principles) by developing and applying an extended method on the basis of the data collected.

Objectives

This study builds upon the foundation laid by previous studies in the health app quality context. This work was motivated by
interest to find and apply a method that helps to more finely differentiate between a chosen set of quality attributes to be used in such a setting. As indicated above, although there are a variety of tools for this task or lists of quality principles for different app types in the mHealth domain, there are voices lamenting that despite these tools being academically sound, applying them in a real-world setting or for a large number of apps may be too tedious [22].

In our evaluation, the proposed method was applied to the nine predefined health app quality principles to determine whether it is feasible to determine an adequate and stable ranking of such criteria to be used for prioritization in facilitating app assessments should the need arise.

Basic Design of the Study

Our approach is based on a group of popular techniques for classifying quality attributes that are often used in decision-making processes in the areas of marketing, management, or even a product’s design phase [23] if a decision is to be made about which (planned or existing) attributes of a product elicit customer satisfaction (and should thus be used or further investigated for a product) or dissatisfaction (making them superficial or even counterproductive for the product’s success). Following this line of thought, we used a survey design based on Kano’s model of attractive quality for classifying quality attributes (originally published in Japanese [20] and subsequently in English [24]), and applied various more elaborate evaluation techniques as specified in the literature (eg, those proposed by Timko as cited in Berger et al [25]) to the acquired data. 

Using the Kano survey data and available evaluation methods, it may be conceivable to find sufficiently differing categorizations of the quality principles that allow for selecting a particularly relevant subset of principles based on their assigned (Kano or derived) category, whereas principles in lower-ranking or less-desirable categories are treated as deferred or are even removed from further consideration. As applied to the nine quality principles, we suspected that even if the principles are largely seen as similarly important, some might be viewed as more attractive, essential, or indifferent than others. Based on a per-category ranking (depending on the perceived relevance of the categories for the use case), we deemed it possible to determine at least a partial prioritization.

As the first idea was unfortunately quickly disproved due to the largely similar categorizations of the nine principles based on the acquired survey data, as a second approach, we tried to better take into account to what degree a product’s attributes, or in our case the app quality principles, contribute to (customer) satisfaction or dissatisfaction, specifically based on the work proposed by Timko in Berger et al [25]. Our assumption was that by appropriately taking both the numeric values for satisfaction as well as dissatisfaction into account, it should be possible to determine a numeric representation in the form of a ranking coefficient (eg, using a ratio of the two values or similar approaches) that could lay the foundation for finding a relatively stable means for prioritization of app quality principles based on this value.

Methods

Data Acquisition

Implementation

Data collection for the study took place in the form of an anonymous and data protection–compliant online survey, implemented using the SoSci Survey [26] installation provided at Hannover Medical School. The survey was open for 1 month (between December 2, 2019, and January 2, 2020), and using the mailing lists of both the German Society for Orthopedics and Trauma Surgery (DGOU) and the Orthopedics and Trauma Surgery Professional Association (BVOU); a total of 9503 members of these societies were invited to participate.

Prior to sending the survey invitation, the study was reviewed by the Ethics Committee of Hannover Medical School (application number 8746_BO_K_2019). In the vote dated November 4, 2019, no ethical or legal objections were raised.

Structure of the Survey

The actual survey itself was conducted in two parts. The first part contained questions about the German Digital Healthcare Act (DVG [27]) that, at the time of the survey, had recently been ratified. Participants were presented with questions about their familiarity with this act, their opinions about its coverage, and whether they were at all considering making use of the possibility to prescribe health apps based on the processes specified in the DVG. The data corresponding to this part of the survey were previously evaluated and published [28].

To acquire demographic data, those responding to the survey were asked questions related to age and gender, as well as about their work history and environment (how long they had been working; their current function; and whether they were working in private practice, at a clinic, or another institution). To allow a basic assessment about their familiarity with mHealth, they were also asked about their private and work-related usage of mHealth apps, and whether any patients asked them either about specific health apps or about a recommendation for a health app. However, the demographic data are only presented to describe the participating physicians. Apart from exemplary calculations given in the Discussion, these data were not part of the analyses presented in this paper.

The work presented herein specifically deals with the second part of the survey. As mentioned in the Introduction, a predefined set of nine quality principles (practicality, risk adequacy, ethical soundness, legal conformity, content validity, technical adequacy, usability, resource efficiency, and transparency) was employed as a basis for the evaluation. The set of quality principles has previously been published [16, 17] along with their evaluations [18,19].

In the context of the work presented here, following Kano’s method, for each of the nine quality principles, the participants were presented with a set of so-called functional and dysfunctional questions (see Table 1). Answer options for both types of questions were “I would be very pleased,” “I’d expect this,” “I don’t care,” “I could accept that,” and “That would really bother me.”
Using the Kano model, based on the answers given for both functional and dysfunctional questions (see Table 3), a product’s features can be categorized as attractive (A), if its presence leads to satisfaction but there is no (additional) dissatisfaction if it is missing [25]; must-be (M), if the respective feature is deemed essential (ie, if it does not improve satisfaction if available, but leads to extreme dissatisfaction if missing) [29]; one-dimensional (O), also referred to as the performance (P) category in the literature, if both availability and lack of the feature cause satisfaction and dissatisfaction, respectively [25], thus representing a feature that customers explicitly demand; indifferent (I), if the feature (or the lack thereof) influences neither satisfaction nor dissatisfaction, thus being ideal for elimination if a reduction in overhead is desired [30]; reverse (R), if dissatisfaction is caused if the feature is available and satisfaction if it is missing; and questionable (Q) if the answers given to the functional and dysfunctional questions are in contradiction [25] (eg, if both answers are specified as “I would be very pleased”).

Table 1. Quality principles with the corresponding questions (translated from the original German-language version) for functional and dysfunctional aspects, as required by the Kano model.

<table>
<thead>
<tr>
<th>Principle</th>
<th>Functional question</th>
<th>Dysfunctional question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practicality</td>
<td>What would you say if apps could be used for the intended purpose?</td>
<td>What would you say if apps could not be used for the intended purpose?</td>
</tr>
<tr>
<td>Risk adequacy</td>
<td>What would you say if apps did not pose a disproportionate health, social, or economic risk to users?</td>
<td>What would you say if apps posed disproportionate health, social, or economic risks to users?</td>
</tr>
<tr>
<td>Ethical soundness</td>
<td>What would you say if discrimination and stigmatization were avoided when developing, offering, and using apps?</td>
<td>What would you say if discrimination or stigmatization were not avoided when developing, offering, operating, and using apps?</td>
</tr>
<tr>
<td>Legal conformity</td>
<td>What would you say if apps were compliant with data protection regulations as well as professional and health regulations?</td>
<td>What would you say if apps failed to comply with data protection, professional, or health regulations?</td>
</tr>
<tr>
<td>Content validity</td>
<td>What would you say if the content used in apps was valid and trustworthy?</td>
<td>What would you say if the content used in apps was not valid or not trustworthy?</td>
</tr>
<tr>
<td>Technical adequacy</td>
<td>What would you say if apps were easy to maintain and could be used independent of a specific platform?</td>
<td>What would you say if apps were hard to maintain or could not be used independent of a specific platform?</td>
</tr>
<tr>
<td>Usability</td>
<td>What would you say if apps were designed and implemented according to the requirements of the target group(s)?</td>
<td>What would you say if apps were not designed and implemented to meet the needs of the target group(s)?</td>
</tr>
<tr>
<td>Resource efficiency</td>
<td>What would you say if apps were to use resources such as battery and computing power efficiently?</td>
<td>What would you say if apps made only inefficient use of resources such as battery or computing power?</td>
</tr>
<tr>
<td>Transparency</td>
<td>What would you say if apps provided transparent information about inherent quality features?</td>
<td>What would you say if apps did not provide transparent information about inherent quality characteristics?</td>
</tr>
</tbody>
</table>

In addition to the functional and dysfunctional questions, the participants were also asked to rate the perceived relevance for each of the nine principles (Table 2). In this case, answers could be given using a 5-point scale: “very important,” “important,” “neutral,” “less important,” and “unimportant.”

Table 2. Questions regarding the relevance for each of the nine quality principles (translated from the original German version).

<table>
<thead>
<tr>
<th>Principle</th>
<th>Perceived relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practicality</td>
<td>How important is it to you that apps can be used for the intended purpose?</td>
</tr>
<tr>
<td>Risk adequacy</td>
<td>How important is it to you that apps are low risk in terms of health, social, or economic risks?</td>
</tr>
<tr>
<td>Ethical soundness</td>
<td>How important is it to you to avoid discrimination and stigmatization when developing, offering, operating, and using apps?</td>
</tr>
<tr>
<td>Legal conformity</td>
<td>How important is it to you that data protection, professional, and health regulations are respected in apps?</td>
</tr>
<tr>
<td>Content validity</td>
<td>How important is the validity and trustworthiness of the health-related content presented and used in an app to you?</td>
</tr>
<tr>
<td>Technical adequacy</td>
<td>How important are easy maintainability and platform-independent or cross-platform usability of apps to you?</td>
</tr>
<tr>
<td>Usability</td>
<td>How important is the target group–oriented design and operation of apps to you?</td>
</tr>
<tr>
<td>Resource efficiency</td>
<td>How important to you is the efficient use of resources through apps, for example in terms of battery and computing power?</td>
</tr>
<tr>
<td>Transparency</td>
<td>How important is it to you that apps provide transparent information about inherent quality features?</td>
</tr>
</tbody>
</table>

Categorization of Answers According to Kano

For each quality principle, the “functional” question was always presented first, followed by the “dysfunctional” question, and that for relevance. However, for each participant, the order in which the questions were shown was randomly assigned to alleviate bias based on an attribute’s position in the list.
Both the reverse and questionable categories may, for example, be due to inadequate wording of the questions employed in the survey or side effects from other (not necessarily easily explainable) factors that impact the answers. Especially for the questionable category, the answers given may also indicate that a participant was (for whatever reason) unwilling to answer in a sensible manner.

### Evaluation Strategies

For each of the nine quality principles, the answers provided by the participants for the functional and dysfunctional question pairs were then categorized based on Table 3, and the frequency that each category was assigned to each attribute was calculated. These counts were then used for further evaluation. As described previously [25,31], there are several strategies that can be applied for this task.

One approach is to determine the category for a feature based on its greatest frequency. Alternatively, an if-then-based approach can be adopted: if \((P+A+M)>(I+R+O)\), the category that corresponds to the maximum count for performance, attractive, or must-be is used; however, if \((P+A+M)>(I+R+O)\), the category corresponding to the maximum of indifferent, reverse, or questionable as the category assigned to the feature under consideration is used.

Both of these approaches work best if those surveyed are somewhat consistent in their answers for a specific feature, or at least show a clear tendency toward a specific category for that feature. However, these approaches do not work quite as well if the responses are distributed more evenly across several categories such as attractive, performance, must-be, and indifferent. Moreover, if different features elicit similar responses, it may be difficult to discriminate between them. This may hamper the usefulness of the approach in the context of categorization.

Timko ([25]) proposed an additional method, as he noted that based on the aforementioned mode statistic, the results may seem somewhat skewed. For example, for two features with only attractive and indifferent ratings, albeit one with a 90-to-10 attractive-to-indifferent ratio and the other with only a 60-to-40 attractive-to-indifferent ratio, the assigned category will be attractive for both. Thus, a third method tries to alleviate these disadvantages.

This method uses the previously obtained counts to calculate two distinct values: one representing the relative value of meeting a customer requirement (namely, “what if we’re better” in contrast to a competitor) and the other representing the relative cost of not meeting the customer requirement (ie, worse than the competition). The two values, as defined in Berger et al [25], are calculated as follows:

- **Better** = \((A+P)/(A+O+M+I)\), with 0 ≤ Better ≤ 1
- **Worse** = \(-(O+M)/(A+O+M+I)\), with -1 ≤ Worse ≤ 0

On average, satisfaction will increase for attractive and one-dimensional (performance) attributes, which is why, in the literature, “Better” is also often denoted as the satisfaction index [32,33], and satisfaction decreases if one-dimensional and must-be elements are not adequately represented. For this reason, “Worse” is often called the dissatisfaction index [32,33]. Both questionable as well as reverse answers are ignored in Timko’s approach, but nevertheless, the calculations do respect a possible spread of the attributes under consideration over the different categories.

The Worse-Better pairing for calculated attributes can be plotted on a two-dimensional and easy-to-interpret graph. Commonly, the values for each attribute are additionally multiplied by the average relevance the participants assign to each attribute to improve discrimination between value pairs for features located in direct vicinity to each other. According to Timko, when deciding which attributes to keep or to omit, one should choose...
those for which satisfaction (ie, the Better score) is higher, since they add more to customer satisfaction, whereas on the Worse axis, one should aim for more negative values, as they prevent dissatisfaction [25] (Figure 1).

**Figure 1.** Two-dimensional representation of Worse-Better pairings for the Kano quality categories [25]. For easier interpretation, Worse is shown with its absolute value.

Designing an Improved Methodology for Prioritization

Discussions among the authors led to the conclusion that established methods such as those described above were suffering from only being able to assign broadly defined categories to the attributes under consideration, without allowing for a more granular consideration that actually respects the relative location of the attributes under consideration. This is particularly relevant when the attributes to be compared (represented by their Worse and Better coordinates) are (predominantly) located in one of the four quadrants and are therefore assigned to the same category (ie, indifferent, must-be, attractive, or one-dimensional). With this in mind, we designed an “in-line-of-sight” method that allows for rankings depending on different points of view on the coordinate system.

This new approach makes it possible to establish a reference to the proximity of an attribute’s (or quality principle’s) coordinate points to the respective outermost corner (corresponding to the point most clearly representing the quadrant), and further respects their relative positions for obtaining the ranking.

This approach will now be explained in more detail by way of an example, using the must-be quadrant as a point of reference. Starting from the outermost point of this category, denoted by the coordinates \((Worse = -1, Better = 0)\), for each attribute (or quality principle), the Euclidean distance between this point and the respective coordinate is first calculated. An increasing distance to the must-be corner represents a greater proximity to one of the three other categories (and is, as such, less desirable).

For further improved differentiation between quality principles, even in the case of (almost) identical Euclidean distances, an angle is then determined based on the chosen secondary ranking strategy. In our example (and all further calculations shown in this paper), we decided to prefer points with less pronounced Worse values (ie, those that have less potential for causing dissatisfaction according to Timko). For this purpose, we chose to calculate an offset based on the angle (denoted by \(\alpha\)) between the x-axis of the coordinate system and the line defined by the corner point’s coordinate \(p=(-1,0)\) as well as the respective quality principle’s \(q=(-Worse, Better)\) coordinate (see Figure 2). As \(\alpha\) is only supposed to aid with differentiation between points with similar distance values, it needs to be rescaled to an appropriate value range. First, \(\alpha\) is divided by the maximum possible angle (ie, 90°) and then multiplied with \(0.05 \times 2 \approx 0\) (representing 5% of the maximum possible distance of the square root of 2 in the coordinate system). The distance and adapted angle value are then summarized (hereinafter referred to as the ranking coefficient \(f\)), and the resulting value for \(f\) is then used for ranking the quality principles according to ascending order as follows:

For simplification, as the plots use an inverted x-axis for representing the Worse value, all statements (as well as the angle calculations) concerning the left- or right-hand location of any point or axis mentioned in relation to the coordinate system refer to this inverted plot. For the other three quadrants, if necessary, rankings may be performed in a similar manner.
Statistics Tools
The R language and environment for statistical computing, version 4.0, was used for all evaluations, along with accompanying packages such as dplyr, ggplot2, arsenal, and others [34-36].

Results
Data
Of those who answered our survey, only 382 actually completed all of its parts, and were thus included in the evaluation presented here. This corresponds to a return rate of 4.02% of the 9503 potential participants.

Using the sample_frac function provided by the dplyr package [34], the available participants were randomly assigned to the test (group A, n=191) and validation (group B, n=191) groups.

Baseline Demographics of the Participants
To rule out differences between the two groups due to demographic factors, these were first compared. There were no statistically significant differences between the groups with respect to baseline demographics (P>.05 for all factors, see Table 4). Overall, the participants were predominantly male and older than 40 years (290/382, 75.9%). In line with the age structure, over three-quarters of the participants had a work experience of more than 10 years (288/328, 75.4%; excluding retirees, 19/328, 5.8%) and were working in higher-level functions (attendings, chiefs, or specialists in private practice; 284/328, 74.3%). The majority of participants worked in a hospital setting (acute care or university hospital; 232/382, 60.7%). As we had only surveyed members of two German orthopedic societies, the proportion of those who were not active in Germany was low, as expected (10/382, 2.6%).

Although the participants overwhelmingly stated that they were highly interested or interested in digital technology (316/382, 82.7%), this was not mirrored by the proportion of those admitting to app use in private or work settings. Only slightly over one-fifth of those participating had already been asked by patients about a specific app or about recommending an app (see Table 4 for full data).
Table 4. Base demographics for all participants and for those assigned to the test group (A) and validation group (B).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Group A (n=191), n (%)</th>
<th>Group B (n=191), n (%)</th>
<th>Total (N=382), n (%)</th>
<th>P value&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age group (years)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21-30</td>
<td>9 (4.7)</td>
<td>7 (3.7)</td>
<td>16 (4.2)</td>
<td>.87</td>
</tr>
<tr>
<td>31-40</td>
<td>34 (17.8)</td>
<td>42 (22.0)</td>
<td>76 (19.9)</td>
<td></td>
</tr>
<tr>
<td>41-50</td>
<td>46 (24.1)</td>
<td>44 (23.0)</td>
<td>90 (23.6)</td>
<td></td>
</tr>
<tr>
<td>51-60</td>
<td>62 (32.5)</td>
<td>59 (30.9)</td>
<td>121 (31.7)</td>
<td></td>
</tr>
<tr>
<td>&gt;60</td>
<td>40 (20.9)</td>
<td>39 (20.4)</td>
<td>79 (20.7)</td>
<td></td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
<td>.38</td>
</tr>
<tr>
<td>Female</td>
<td>24 (12.6)</td>
<td>30 (15.7)</td>
<td>54 (14.1)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>167 (87.4)</td>
<td>161 (84.3)</td>
<td>328 (85.9)</td>
<td></td>
</tr>
<tr>
<td><strong>Work experience</strong></td>
<td></td>
<td></td>
<td></td>
<td>.93</td>
</tr>
<tr>
<td>Not yet working</td>
<td>2 (1.0)</td>
<td>1 (0.5)</td>
<td>3 (0.8)</td>
<td></td>
</tr>
<tr>
<td>&lt;1 year</td>
<td>2 (1.0)</td>
<td>2 (1.0)</td>
<td>4 (1.0)</td>
<td></td>
</tr>
<tr>
<td>1-5 years</td>
<td>10 (5.2)</td>
<td>14 (7.3)</td>
<td>24 (6.3)</td>
<td></td>
</tr>
<tr>
<td>6-10 years</td>
<td>19 (9.9)</td>
<td>25 (13.1)</td>
<td>44 (11.5)</td>
<td></td>
</tr>
<tr>
<td>11-20 years</td>
<td>50 (26.2)</td>
<td>44 (23.0)</td>
<td>94 (24.6)</td>
<td></td>
</tr>
<tr>
<td>21-30 years</td>
<td>54 (28.3)</td>
<td>50 (26.2)</td>
<td>104 (27.2)</td>
<td></td>
</tr>
<tr>
<td>&gt;30 years</td>
<td>44 (23.0)</td>
<td>46 (24.1)</td>
<td>90 (23.6)</td>
<td></td>
</tr>
<tr>
<td>Retired</td>
<td>10 (5.2)</td>
<td>9 (4.7)</td>
<td>19 (5.0)</td>
<td></td>
</tr>
<tr>
<td><strong>Professional level</strong></td>
<td></td>
<td></td>
<td></td>
<td>.75</td>
</tr>
<tr>
<td>Student</td>
<td>1 (0.5)</td>
<td>0 (0.0)</td>
<td>1 (0.3)</td>
<td></td>
</tr>
<tr>
<td>In training/resident</td>
<td>23 (12.0)</td>
<td>25 (13.1)</td>
<td>48 (12.6)</td>
<td></td>
</tr>
<tr>
<td>Attending</td>
<td>60 (31.4)</td>
<td>52 (27.2)</td>
<td>112 (29.3)</td>
<td></td>
</tr>
<tr>
<td>Chief</td>
<td>38 (19.9)</td>
<td>39 (20.4)</td>
<td>77 (20.2)</td>
<td></td>
</tr>
<tr>
<td>Specialist (private practice)</td>
<td>47 (24.6)</td>
<td>48 (25.1)</td>
<td>95 (24.9)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>21 (11.0)</td>
<td>27 (14.1)</td>
<td>48 (12.6)</td>
<td></td>
</tr>
<tr>
<td>Not answered</td>
<td>1 (0.5)</td>
<td>0 (0.0)</td>
<td>1 (0.3)</td>
<td></td>
</tr>
<tr>
<td><strong>Work setting</strong></td>
<td></td>
<td></td>
<td></td>
<td>.49</td>
</tr>
<tr>
<td>Acute care: standard care level</td>
<td>63 (33.0)</td>
<td>50 (26.2)</td>
<td>113 (29.6)</td>
<td></td>
</tr>
<tr>
<td>Acute care: maximum care level</td>
<td>32 (16.8)</td>
<td>37 (19.4)</td>
<td>69 (18.1)</td>
<td></td>
</tr>
<tr>
<td>University hospital</td>
<td>21 (11.0)</td>
<td>29 (15.2)</td>
<td>50 (13.1)</td>
<td></td>
</tr>
<tr>
<td>Rehabilitation center</td>
<td>8 (4.2)</td>
<td>7 (3.7)</td>
<td>15 (3.9)</td>
<td></td>
</tr>
<tr>
<td>Medical care center</td>
<td>6 (3.1)</td>
<td>9 (4.7)</td>
<td>15 (3.9)</td>
<td></td>
</tr>
<tr>
<td>Private practice</td>
<td>40 (20.9)</td>
<td>44 (23.0)</td>
<td>84 (22.0)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>21 (11.0)</td>
<td>14 (7.3)</td>
<td>35 (9.2)</td>
<td></td>
</tr>
<tr>
<td>Not answered</td>
<td>0 (0.0)</td>
<td>1 (0.5)</td>
<td>1 (0.3)</td>
<td></td>
</tr>
<tr>
<td><strong>Geographic location</strong>&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td>.26</td>
</tr>
<tr>
<td>Germany</td>
<td>187 (98.9)</td>
<td>183 (95.8)</td>
<td>370 (97.4)</td>
<td></td>
</tr>
<tr>
<td>Austria</td>
<td>0 (0.0)</td>
<td>2 (1.0)</td>
<td>2 (0.5)</td>
<td></td>
</tr>
<tr>
<td>Switzerland</td>
<td>2 (1.1)</td>
<td>3 (1.6)</td>
<td>5 (1.3)</td>
<td></td>
</tr>
<tr>
<td>Other: European Union</td>
<td>0 (0.0)</td>
<td>2 (1.0)</td>
<td>2 (0.5)</td>
<td></td>
</tr>
<tr>
<td>Other: not yet listed</td>
<td>0 (0.0)</td>
<td>1 (0.5)</td>
<td>1 (0.3)</td>
<td></td>
</tr>
</tbody>
</table>
Interest in digital technology

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Group A (n=191), n (%)</th>
<th>Group B (n=191), n (%)</th>
<th>Total (N=382), n (%)</th>
<th>P valuea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highlty interested</td>
<td>76 (39.8)</td>
<td>81 (42.4)</td>
<td>157 (41.1)</td>
<td>.71</td>
</tr>
<tr>
<td>Interested</td>
<td>84 (44.0)</td>
<td>75 (39.3)</td>
<td>159 (41.6)</td>
<td></td>
</tr>
<tr>
<td>Neutral</td>
<td>19 (9.9)</td>
<td>25 (13.1)</td>
<td>44 (11.5)</td>
<td></td>
</tr>
<tr>
<td>Less interested</td>
<td>8 (4.2)</td>
<td>8 (4.2)</td>
<td>16 (4.2)</td>
<td></td>
</tr>
<tr>
<td>Not interested</td>
<td>4 (2.1)</td>
<td>2 (1.0)</td>
<td>6 (1.6)</td>
<td></td>
</tr>
</tbody>
</table>

Uses apps in private settings

<table>
<thead>
<tr>
<th>Answer</th>
<th>Group A (n=191)</th>
<th>Group B (n=191)</th>
<th>Total (N=382)</th>
<th>P valuea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>69 (36.1)</td>
<td>70 (36.6)</td>
<td>139 (36.4)</td>
<td>.92</td>
</tr>
<tr>
<td>No</td>
<td>122 (63.9)</td>
<td>121 (63.4)</td>
<td>243 (63.6)</td>
<td></td>
</tr>
</tbody>
</table>

Uses apps for work

<table>
<thead>
<tr>
<th>Answer</th>
<th>Group A (n=191)</th>
<th>Group B (n=191)</th>
<th>Total (N=382)</th>
<th>P valuea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>63 (33.0)</td>
<td>73 (38.2)</td>
<td>136 (35.6)</td>
<td>.29</td>
</tr>
<tr>
<td>No</td>
<td>128 (67.0)</td>
<td>118 (61.8)</td>
<td>246 (64.4)</td>
<td></td>
</tr>
</tbody>
</table>

Been asked about an app/recommendation

<table>
<thead>
<tr>
<th>Answer</th>
<th>Group A (n=191)</th>
<th>Group B (n=191)</th>
<th>Total (N=382)</th>
<th>P valuea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>43 (22.5)</td>
<td>43 (22.5)</td>
<td>86 (22.5)</td>
<td>&gt; .99</td>
</tr>
<tr>
<td>No</td>
<td>148 (77.5)</td>
<td>148 (77.5)</td>
<td>296 (77.5)</td>
<td></td>
</tr>
</tbody>
</table>

aPearson $\chi^2$ test.
bNot answered: group A, n=2.

Data Evaluation

**Descriptive Evaluation of the Survey Results**

Similar to the participants’ demographics, in the Kano-based questionnaire, there were no statistically significant differences between the training and validation groups with respect to answers given for the functional and dysfunctional questions, as well as the perceived relevance for the nine app quality criteria (see Figures 3, 4, and 5; for more detailed counts, proportions, and $P$ values for the available answers, see Multimedia Appendix 1, Tables S1-S3).

Figure 3. Distribution of answers for the functional questions. For legibility reasons, smaller values are not printed (see Multimedia Appendix 1 for the complete list of values).
Figure 4. Distribution of answers for the dysfunctional questions. For legibility reasons, smaller values are not printed (see Multimedia Appendix 1 for the complete list of values).

Figure 5. Ratings for relevance of the nine quality principles, as perceived by the participants. For legibility reasons, smaller values are not printed (see Multimedia Appendix 1 for the complete list of values).

Categorization According to Kano
Using Kano’s basic evaluation described in the “Evaluation Strategies Applied” subsection within the Methods, namely choosing the category with the largest number of counts as that to assign to each quality principle, the nine evaluated quality principles were exclusively categorized as must-be (see Table 5). This gives all attributes equal impact, which made it impossible to prioritize certain quality principles as desired, despite differences in ratings.
Table 5. Categorization of the answers for the functional and dysfunctional questions related to the nine quality principles, based on the category with the maximum count.

<table>
<thead>
<tr>
<th>Quality principle</th>
<th>Test group, A (n=191)</th>
<th>Validation group, B (n=191)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M&lt;sup&gt;a&lt;/sup&gt;</td>
<td>P&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Practicality</td>
<td>127</td>
<td>42</td>
</tr>
<tr>
<td>Risk adequacy</td>
<td>127</td>
<td>48</td>
</tr>
<tr>
<td>Ethical soundness</td>
<td>120</td>
<td>40</td>
</tr>
<tr>
<td>Legal conformity</td>
<td>148</td>
<td>27</td>
</tr>
<tr>
<td>Content validity</td>
<td>139</td>
<td>42</td>
</tr>
<tr>
<td>Technical adequacy</td>
<td>83</td>
<td>68</td>
</tr>
<tr>
<td>Usability</td>
<td>103</td>
<td>49</td>
</tr>
<tr>
<td>Resource efficiency</td>
<td>63</td>
<td>40</td>
</tr>
<tr>
<td>Transparency</td>
<td>103</td>
<td>43</td>
</tr>
</tbody>
</table>

<sup>a</sup>M: must-be.
<sup>b</sup>P: performance.
<sup>c</sup>A: attractive.
<sup>d</sup>I: indifferent.
<sup>e</sup>R: reverse.
<sup>f</sup>Q: questionable.

For example, for resource efficiency, less than half as many answer pairs were categorized under must-be compared with those for content validity (Group A: 63 vs 139 or 45.3%; Group B: 69 vs 140 or 49.3%); nevertheless, both principles were still equally categorized as must-be.

**If-Then-Based Approach**

The situation did not improve when employing the if-then approach; the results were equivalent to those shown in Table 5.

**Timko Approach**

Even using the method proposed by Timko [25], with or without using the average values for perceived importance, the situation only changed marginally, as shown in Table 6 and Figure 6. Visually, the value pairs were still in close vicinity to each other. Without factoring in perceived relevance, all values firmly remained categorized as must-be; only when accounting for relevance, one quality principle, specifically resource efficiency, showed a categorization change from must-be to indifferent. Apart from this principle (which, now being rated indifferent is deemed to be of less importance), prioritization of the remaining attributes was elusive, despite apparent (visual and numeric) differences.

Table 6. Better and Worse values without (denoted by a subscripted N) and with factoring in the average value of perceived relevance (or importance, denoted by a subscripted I) for each principle.

<table>
<thead>
<tr>
<th>Quality principle</th>
<th>Group A</th>
<th>Group B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Better&lt;sub&gt;N&lt;/sub&gt;</td>
<td>Worse&lt;sub&gt;N&lt;/sub&gt;</td>
</tr>
<tr>
<td>Practicality</td>
<td>0.28</td>
<td>-0.91</td>
</tr>
<tr>
<td>Risk adequacy</td>
<td>0.27</td>
<td>-0.94</td>
</tr>
<tr>
<td>Ethical soundness</td>
<td>0.26</td>
<td>-0.86</td>
</tr>
<tr>
<td>Legal conformity</td>
<td>0.15</td>
<td>-0.92</td>
</tr>
<tr>
<td>Content validity</td>
<td>0.23</td>
<td>-0.96</td>
</tr>
<tr>
<td>Technical adequacy</td>
<td>0.47</td>
<td>-0.80</td>
</tr>
<tr>
<td>Usability</td>
<td>0.37</td>
<td>-0.80</td>
</tr>
<tr>
<td>Resource efficiency</td>
<td>0.45</td>
<td>-0.55</td>
</tr>
<tr>
<td>Transparency</td>
<td>0.33</td>
<td>-0.78</td>
</tr>
</tbody>
</table>
**Figure 6.** Better and Worse pairings for the training (Group A) and validation (Group B) groups, plotted with and without the average value for perceived importance. The arrows represent the corresponding coordinate shift from the original values to those factoring in the perceived importance for each quality principle.

**Table 7.** Ranking the quality principles based on distance to the must-be corner and angle toward the right-most boundary.

<table>
<thead>
<tr>
<th>Quality principle</th>
<th>Coordinate distance between groups</th>
<th>Group A (test group)</th>
<th>Group B (validation group)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Distance, d</td>
<td>Angle, α</td>
<td>Ranking coefficient, f</td>
</tr>
<tr>
<td>Practicality</td>
<td>0.00</td>
<td>0.32</td>
<td>51</td>
</tr>
<tr>
<td>Risk adequacy</td>
<td>0.03</td>
<td>0.29</td>
<td>53</td>
</tr>
<tr>
<td>Ethical soundness</td>
<td>0.05</td>
<td>0.35</td>
<td>38</td>
</tr>
<tr>
<td>Legal conformity</td>
<td>0.05</td>
<td>0.23</td>
<td>37</td>
</tr>
<tr>
<td>Content validity</td>
<td>0.01</td>
<td>0.24</td>
<td>59</td>
</tr>
<tr>
<td>Technical adequacy</td>
<td>0.02</td>
<td>0.51</td>
<td>48</td>
</tr>
<tr>
<td>Usability</td>
<td>0.03</td>
<td>0.45</td>
<td>43</td>
</tr>
<tr>
<td>Resource efficiency</td>
<td>0.05</td>
<td>0.70</td>
<td>26</td>
</tr>
<tr>
<td>Transparency</td>
<td>0.05</td>
<td>0.46</td>
<td>34</td>
</tr>
</tbody>
</table>

**In-Line-of-Sight Method**

Table 7 shows the rankings for both groups based on the *must-be* quadrant, as this is where the attributes predominantly clustered. Angles were calculated in the direction of the one-dimensional (performance) category.

The distances between Better-Worse pairings for both groups (ie, the distance between the two groups) only differed insignificantly: they always remained below 5% the maximum possible distance within the coordinate square (ie, $0.05 \times (0,0),(-1,1)] = 0.05 \times \sqrt{2} = 0.05 \times 1.14142 = 0.0707$).

Based on the described method, the ranking for the quality principles was identical for both groups, with legal conformity ranked first, followed by content validity, risk adequacy, practicality, ethical soundness, usability, transparency, technical adequacy, and finally, resource efficiency.

**Gender Influence**

There was only a slight difference in the quality principle–related assessments between male and female participants. As there were too few female participants to prevent outliers from unduly influencing the results to continue evaluating groups A and B separately in this regard, the overall group of all participants was stratified by gender. There were only small differences in prioritization, despite (significant) disparities between both strata regarding the actual placement of the principles in the coordinate system (**Figure 7** and **Table 8**).
Figure 7. Plot of the Better and Worse coordinates per principle stratified by gender.

Table 8. Ranking of the quality principles based on the distance of the Better and Worse coordinates to the outermost corner of the must-be quadrant, using the in-line-of-sight method for all participants, stratified by gender.

<table>
<thead>
<tr>
<th>Quality principle</th>
<th>Female participants</th>
<th>Male participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coordinate distance between strata</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Distance, d</td>
<td>Angle, α</td>
</tr>
<tr>
<td>Practicality</td>
<td>0.085</td>
<td>0.37</td>
</tr>
<tr>
<td>Risk adequacy</td>
<td>0.086</td>
<td>0.28</td>
</tr>
<tr>
<td>Ethical soundness</td>
<td>0.130</td>
<td>0.31</td>
</tr>
<tr>
<td>Legal conformity</td>
<td>0.115</td>
<td>0.23</td>
</tr>
<tr>
<td>Content validity</td>
<td>0.077</td>
<td>0.19</td>
</tr>
<tr>
<td>Technical adequacy</td>
<td>0.070</td>
<td>0.57</td>
</tr>
<tr>
<td>Usability</td>
<td>0.062</td>
<td>0.48</td>
</tr>
<tr>
<td>Resource efficiency</td>
<td>0.160</td>
<td>0.68</td>
</tr>
<tr>
<td>Transparency</td>
<td>0.094</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Stratification by Interest in Digitization

There were notable differences in ratings between those with a stated interest in digitization and those who lacked interest in this topic, again considering only the overall group and discarding groups A and B due to the low number of participants in the “little to no interest” stratum (Figure 8). For the latter group, the principles were almost exclusively located in the indifferent quadrant, or, in the case of legal conformity, content validity, and risk adequacy, near the border between the indifferent and must-be quadrants.

Nevertheless, the prioritization remained largely similar with that of the interest-based stratification, with only minor differences (see Table 9).
Figure 8. Plot of the Better and Worse coordinates per principle stratified by interest in the topic.

Table 9. Ranking of the quality principles based on the distance of the Better and Worse coordinates to the outermost corner of the must-be quadrant, using the in-line-of-sight method for all participants, stratified by their interest in digitization.

<table>
<thead>
<tr>
<th>Quality principle</th>
<th>Interested participants</th>
<th>Uninterested participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coordinate distance between strata</td>
<td>Ranking coefficient, $f$</td>
</tr>
<tr>
<td>Practicality</td>
<td>0.44</td>
<td>0.31</td>
</tr>
<tr>
<td>Risk adequacy</td>
<td>0.42</td>
<td>0.28</td>
</tr>
<tr>
<td>Ethical soundness</td>
<td>0.36</td>
<td>0.34</td>
</tr>
<tr>
<td>Legal conformity</td>
<td>0.36</td>
<td>0.23</td>
</tr>
<tr>
<td>Content validity</td>
<td>0.44</td>
<td>0.24</td>
</tr>
<tr>
<td>Technical adequacy</td>
<td>0.34</td>
<td>0.51</td>
</tr>
<tr>
<td>Usability</td>
<td>0.50</td>
<td>0.42</td>
</tr>
<tr>
<td>Resource efficiency</td>
<td>0.17</td>
<td>0.66</td>
</tr>
<tr>
<td>Transparency</td>
<td>0.34</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Discussion

Principal Results

As shown in the literature (eg, [23,25,37,38]) as well as our own results, established methods for working with the results of Kano surveys are well-suited to determining generic user perceptions of product attributes of a health app, such as the quality principles that the participants of our survey were confronted with.

Nevertheless, when using Kano’s original approach, or even the more promising approach proposed by Timko [25] (with or without inclusion of the perceived relevance of the principles), in our case, the nine attributes remained firmly tethered to the must-be category (see Figure 6), with only resource efficiency crossing into the indifferent realm once perceived importance was included in the calculation. However, there were no one-dimensional or even attractive attributes. Solely based on established evaluation methods for Kano surveys, we therefore
fell short of obtaining the desired ranking to be used for potentially prioritizing the health app quality principles. Simply applying the Kano method and its categorizations to the quality principles initially did not allow for prioritization, which confirmed the previously noted similarity of the ratings [18,19], with again only resource efficiency standing out. As reported previously [18,19], the discrepancy between this quality principle and the other eight principles supports the assumption that resource efficiency likely only plays a minor role in today’s mostly very powerful devices, since health-related apps in particular presumably place little demand on the devices.

To counteract this lack of differentiation between the principles, we then developed the so-called “in-line-of-sight” method, which, based on the numeric values representing satisfaction as well as dissatisfaction with the respective attribute or quality principle, determines a ranking coefficient while also accounting for different points of view (depending on the purpose of the desired prioritization). This method should also be flexible enough to be adapted to different circumstances depending on the use case and user ratings provided.

In our exemplary evaluation for the ranking from the must-be perspective, we chose a rather conservative approach, factoring in an angle that leads to lower Worse values being preferred, while accepting that by choosing this approach, the values for Better will also decrease.

This corresponds to the definition of the must-be category: a lack of the respective quality principle would be perceived more strongly than the positive effect that would be achieved if the characteristics consistent with the quality principle were present. When changing the perspective to another quadrant, similar considerations need to be applied, with calculations being adapted accordingly. For example, when changing the perspective to the attractive quadrant, it would be more useful to aim at a higher priority of Better values, as this better represents the definition of this category.

**Kano Survey Interpretation: Potential for Linguistic Inconsistencies?**

Although the Kano model is popular and is often used in a wide variety of contexts, linguistic inaccuracies in its application have arisen over the years, which in some publications have led to difficulties in its correct application or to supposed inconsistencies ([29], citing [25]). The problem originates from an inaccurate translation of Kano’s key concept transliterated as “atarimae,” which has been translated as must-be in many English-language publications. Must-be seems to have first been used in the early 1990s by Shoji Shiba when presenting the Kano model to English-speaking audiences [29]. However, apparently, the meaning of “atarimae” would be better represented by the terms “natural,” “obvious,” “expected,” “ordinary,” or “normal.” This change should be applied to the category name must-be as well as the corresponding customer response, which is often given as “It must be that way,” but, as noted by Horton and Goers [29], should rather be represented by translations along the lines of the aforementioned suggestions.

When Kano surveys are translated into other languages, this inaccuracy may be passed on to a varying degree, potentially further complicating the situation. In our (German language) questionnaire, however, we already included the wording representing “I take this for granted” (German: “Setze ich voraus”) as an answer option for the participants instead of must-be, thus more closely following Kano’s original idea. To stay in line with most of the literature, we nevertheless decided to stick to the must-be term, although this aspect needs to be kept in mind. This change in interpretation may also provide an explanation for the results we obtained for the nine quality principles, with all of them being located in the must-be category.

In contrast to common usage scenarios for Kano surveys that aim at selecting attributes one should further investigate, we applied the model to a set of attributes, namely our quality principles, that had already been painstakingly compiled [16,17] (among others based on various norms (eg, [3,39-42]), as well as the literature (eg, [5,6,43,44]). This may provide an additional explanation for why, in the survey presented here, all quality principles were rated as must-be, or following the adapted interpretation, as “obvious” or “something to be taken for granted.” That is, the quality principles simply followed obvious requirements that were mentioned as essential in the aforementioned sources, and that one would expect users to be able to rate objectively (at least to a certain degree); they were, however, not selected in order to trigger enthusiasm. Their placement in the must-be quadrant is therefore easily explained, and the sole exception for resource efficiency being placed in the indifferent category may possibly be due to the fact that today’s mobile devices are commonly equipped with sufficient computing power—at least for physicians, who often probably have access to rather high-end devices—so that resources are not a factor that warrants considerable attention.

**Selection of the Evaluation Method Used as a Basis of this Work**

In addition to the linguistic aspects, there is no clear verdict about the methodology one should apply foremost when evaluating Kano model–based surveys. While there is a large variety of methods to choose from, based on various theoretical concepts, the discussion is still open as to which of them is most appropriate (in general or for a specific use case) and has the greatest validity. Although there are various empirical evaluations of different approaches in the context of Kano surveys that are described in the literature (eg, [37,38,45-47]), determining which of these particular approaches is best seems to be near impossible.

As stated by Mikulić and Prebezac [23], the validity and reliability of the various approaches cannot be determined with certainty: there is simply no known comparison that can be taken as the ground truth. Which method is chosen is therefore rather often a matter of whether (1) the theoretical justification of the respective approach appears valid, (2) the increase in information when applying the respective approach actually contributes to the solution of the problem, and (3) which (recognizable) technical strengths and weaknesses the approach has.
For the purposes of this paper, Timko’s approach (first introduced in [25]) was therefore chosen as a foundation, as it is easy to understand and also easily allows for integration of the self-stated relevance of the attributes to be evaluated. Additionally, compared to Kano’s initial idea, where, essentially, all 25 possible answer combinations are directly mapped to only 6 possible categories, one may feel the need for a more differentiated, continuous method of analyzing the data to better assess how different attributes are similar or dissimilar, and our enhanced approach follows this line of thought.

Limitations

Selection of the Quality Principles Employed in This Study

New information technologies, including online information or specific (mobile) apps, place additional demands on those employing them, especially in professional health care contexts. Professionals employing such technologies need to ensure that they are safe and pose no harm to those in their care. Regulatory oversight as well as evidence-based literature are often found lacking [48]. Economic questions such as the paucity of information related to cost-effectiveness or cost-utility [49,50], or even aspects related to reimbursement [51,52] may also play a role in whether or not the technologies are actually adopted in everyday practice.

Without at least a basic understanding of the relevant quality aspects (and how to apply them), or uncertainties regarding their safety and security, acceptance may suffer, which may also limit the potential of these technologies [48,53]. However, there is no general consensus, even among experts, as to what exactly constitutes “quality” in this context and how it can be assessed for specific scenarios (eg, to rate health-related apps) [54].

To identify items of relevance, such as for inclusion in various tools [4-8] meant to aid in assessing such technologies that are to be provided to the respective target groups (eg, physicians or other health care personnel), it is important to identify certain key aspects in the hope that these fulfill the information needs and information-seeking behaviors of users [55]. Many authors use rather detailed approaches and criteria to enable this information-seeking and more easily assess the quality of health-related apps, and they often target specific (professional) user groups [54].

For this purpose, in close collaboration with various stakeholders (eg, experts convened on behalf of eHealth Suisse), the nine quality principles used here were compiled [16,17] and evaluated [18,19]. In this context, we were able to show that, despite its broad scope and lack of details, and being almost unanimously regarded as (highly) relevant by the participants of both previous studies, the predefined set of quality principles was still well-suited to provide the respective participants with pointers to aspects relevant for determining an app’s quality and fine-tuning their usage decisions. After having been sensitized to the topic of quality principles, and having applied these principles to exemplary app descriptions, the participants of both previous studies were able to make a much more differentiated assessment of the app descriptions that were provided, and were much more critical in their decision on whether or not to potentially use the corresponding app.

Survey Design

Although we had initially considered an additional qualitative approach, specifically to ask the participants to directly rank the principles as they saw fit, a major reason that made us abandon this course of action was that the data presented here were part of a larger project (as mentioned above, the first part of the analysis of the acquired data is already published [28]), and it was decided by the team that an additional (sorting) questionnaire would be too much of a burden for those participating in the survey. Of the two alternatives for designing the part of the survey presented here (ie, continuing to rely on the Kano model or using the qualitative sorting approach), the choice ultimately fell on Kano. This was based on our hope to be able to use the data obtained for implicit assessment instead of running the risk that the previously established, highly similar assessments of the principles would make it difficult for the participants to determine a specific order. Because we did not initially know how many people would participate, we were concerned that it would be difficult to determine an overall ranking for the nine principles if too few people participated and we only relied on the explicitly stated rankings. It was hoped that based on Kano’s methods, using the provided answers and ensuing categorizations, we would be able to at least determine a rough prioritization for the overall group of participants, in our case, by giving principles in the must-be or one-dimensional categories precedence over those in the attractive or indifferent categories.

Study Participants

Despite having contacted a relatively large number of potential participants, with only 4.02% (382/9503) of those who were initially invited actually completing the survey, the response rate was low. Based on this response rate and demographic factors, the results, specifically those related to any rankings of attributes presented here, may not be fully representative of physicians overall or even those specializing in orthopedic or trauma surgery.

One of the possibly most relevant demographic factors for which one might potentially expect an impact on the assessments is the gender of the participants. Overall, the gender distribution of the participants roughly corresponded to the ratio expected in orthopedics. In our survey, 85.9% (328/382) of the participants were male and 14.1% (54/382) were female. Thus, there were only slightly fewer women than would have been expected in the field of orthopedics and trauma surgery, according to data provided by the Bundesärztekammer, with 17.63% (3611/20,477), as of December 31, 2020, of those in the fields of orthopedics or orthopedics and trauma surgery being women [56].

However, gender seems to only have exerted a limited influence on prioritization, which is in line with our previous work [18,19], where there were also only minor differences in the quality principle–related assessments between male and female participants. Differences were particularly pronounced for resource efficiency and ethical soundness (see the column
describing the coordinate distance between both strata in Table 8, as well as Figure 7 for the actual coordinates). The former was placed near the (neutral) center for female participants, whereas for male participants, it was clearly placed in the indifferent quadrant. Content validity and usability (along with transparency for female participants) were somewhat closer to the one-dimensional quadrant than the other principles in both strata. In case of the female participants, the point cloud was also shifted more toward the one-dimensional quadrant compared with that of their male peers, and the coordinates were less scattered overall (Figure 7).

Regarding the ranking of the principles, for the female participants (n=54), content validity ranked first and legal conformity ranked second (Table 8). For the male (n=328) participants, this order was reversed. The same was true for ethical soundness and practicality. Apart from resource efficiency, all quality principles were found in the must-be quadrant (Figure 7).

Nevertheless, the prioritization was roughly similar for the two demographic groups: for the female participants (n=54), content validity ranked first and legal conformity was placed second (Table 8), whereas this order was reversed for the male (n=328) participants. The same was true for ethical soundness and practicality. Apart from resource efficiency, all quality principles were found in the must-be quadrant (Figure 7).

Considering interest in digitization (Figure 8 and Table 9), digitally affine participants (aggregated data for “neutral,” “interested,” or “highly interested”; n=360) were considerably overrepresented due to the chosen survey method. Participants with little interest in the topic, or those lacking access to the techniques used, responded much less frequently than those showing more enthusiasm toward digitization, thus potentially biasing the results as well. However, for the limited number of participants (n=22) who cared only little about digitization (values aggregated for being “less interested” or “not interested” in the topic), but nevertheless participated, it was primarily the placement of the points representing the quality principles in the coordinate system that differed strikingly from the other participants (Figure 8). There was also a striking difference in the placement of the principles within the coordinate system, which is probably not solely attributable to the imbalance between the sizes of the two groups. Disinterested participants rated the principles as indifferent, or, in the case of legal conformity, content validity, and risk adequacy, near the border between the indifferent and must-be quadrants (Figure 8). Nevertheless, rankings remained largely similar independent of digital affinity. For those stating a more or less pronounced interest into digitalization, the order of practicality and risk adequacy was reversed compared with that of the participants with little to no interest. Among disinterested participants, there were also small deviations in the rank for legal conformity and content validity (reverse rank 1 and 2, respectively) as well as technical adequacy and usability (rank 6 and 8, respectively; see Table 9). Legal conformity, content validity, and risk adequacy occupied the top ranks among participants with or without interest, but the order for content validity and legal conformity differed. The lower ranks were occupied by usability, transparency, technical adequacy, and resource efficiency, albeit with a somewhat differing order.

The difference in locations of the principles in the coordinate system (Figure 8), but not in the prioritizations obtained for the two groups (Table 9), lends support to the feasibility of applying our method to quality principles in the mHealth app domain, and supports the need for better education of (potential) users of mHealth apps. Although medical professionals such as our participants are—or at any rate should be—aware of the need for quality (as demanded by professional ethics) for all tools they apply in care contexts, it seems as though for those lacking interest in digitization, this mental transfer apparently does not work for the uninterested participants, as shown by their indifferent ratings. Educational campaigns such as those by professional societies that emphasize the need for quality not only in conventional care but also in the digital domain, including mHealth apps, may help to raise awareness in this regard even for those who are not (yet) familiar or comfortable with the use of such technologies in their daily work.

Altogether, an additional, hopefully larger-scale, study should be implemented to obtain more conclusive data for these as well as other demographic strata, such as by recruiting additional participants with the aid of other professional organizations or by including additional target groups such as patient organizations, universities providing medical education, and others.

Implementation

We believe to have found a methodology that is well-adapted to the demands of finding a prioritization of app quality principles in the case of very similar categorizations, clustered either in one of the four categories of must-be, one-dimensional, attractive, or indifferent obtained using a Kano questionnaire.

Of course, our method needs further validation, and, depending on the scenario in which it is applied, it might be helpful to adapt the strategy of how the angles (or their direction) are calculated. This may depend on multiple factors. For example, when considering ratings based on must-be, it seems sensible to always perform sorting based on the distance to the one-dimensional rather than to the indifferent quadrant, as an indifferent opinion, per se, does not elicit identification with the product (or its attributes).

However, if one switches perspective to the one-dimensional category, it may well depend on the type of product, its application areas, as well as its target user group, along with the attributes actually being evaluated if it makes more sense to calculate the angles used in determining the sorting against must-be or attractive. For products targeting professionals, it might, for example, make more sense to sort the quality principles depending on their closeness to must-be, whereas for marketing purposes, attractive qualities may be more promising. Again, for the attractive corner, similar arguments as for the approach taken for must-be apply, with the angle toward one-dimensional rather than indifferent, which likely makes more sense in most scenarios.

If attributes were clustered in the indifferent corner, the question of the direction to base any attribute sorting on is again more
open (ie, toward either attractive or must-be). The decision may also depend somewhat on the purpose, design, and area of use of the product under consideration; in the case of a professional product, it may potentially make more sense to build the ranking based on must-be as a reference, since attractiveness does not necessarily reflect professional quality.

**Outlook and Comparison With Previous Work**

Further proof of the validity of the method and its transferability to other interest groups, quality attributes, or application scenarios is still pending. Future work will particularly have to address further validation of the method with regard to the evaluation involving other user groups (eg, patients, caregivers) or to the application for prioritization of other attributes, whether for use in medical or general apps, or for the evaluation of other attribute lists outside the app domain.

However, especially with regard to the determined ranking of the quality criteria we chose for this evaluation, we believe that a comparison of the perception of relevance between the results of the previous studies (eg, [19], where participants were working in a different medical field) and those shown here is a strong indicator that the results are likely transferable. Similar to the current work, participants of previous studies had also been asked to provide their opinion regarding the relevance of the nine quality principles, and the participating physicians rated the relevance of the quality principles similar to the current group of participants (see Figure 9), with only minor (and statistically negligible) differences between the previous study [19] and the data obtained from the participants of this study. Table S4 in Multimedia Appendix 1 shows the overall relevance ratings and P values for the comparison between the two studies. However, for the sake of streamlining the comparison between studies, the respective test and validation samples, as they were used in both studies, were aggregated. Similarly, to stay in line with Albrecht et al [19], the answer options for “very important” and “important” were summarized using the term “important,” while those for “less important” and “unimportant” were aggregated as “not important.”

As shown in Figure 9 and Table S4 in Multimedia Appendix 1, there are notable similarities between both studies: the proportion of participants that rated resource efficiency as important was decidedly lower (current study: 260/382, 68.1% participants; previous study: 270/441, 61.6%) than it was for all other quality principles, where the perceived importance was in the range of 84%-98%, again for both studies.

**Figure 9.** Relevance ratings for the nine quality principles: comparison between this survey and previously published work [19]. See Table S4 in Multimedia Appendix 1 for the corresponding P values of this comparison. DGIM: German Association for Internal Medicine, German: “Deutsche Gesellschaft für Innere Medizin e.V.”.

<table>
<thead>
<tr>
<th>Group A (Test)</th>
<th>Group B (Validation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practicability</td>
<td>66.5%</td>
</tr>
<tr>
<td>Risk adequacy</td>
<td>62.4%</td>
</tr>
<tr>
<td>Ethical soundness</td>
<td>67.0%</td>
</tr>
<tr>
<td>Legal conformity</td>
<td>81.7%</td>
</tr>
<tr>
<td>Content validity</td>
<td>75.8%</td>
</tr>
<tr>
<td>Technical adequacy</td>
<td>67.1%</td>
</tr>
<tr>
<td>Usability</td>
<td>57.1%</td>
</tr>
<tr>
<td>Resource efficiency</td>
<td>42.4%</td>
</tr>
<tr>
<td>Transparency</td>
<td>59.1%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>%</th>
<th>That would really bother me</th>
<th>I could accept that</th>
<th>I don't care</th>
<th>I'd expect this</th>
<th>I would be very pleased</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Conclusions**

The agreement with respect to perceived relevance between both studies, as shown above, leads to the following conclusions.

For both previous studies [18,19], there was no clear pathway for prioritization of the principles should the need arise, apart from resource efficiency consistently being the least popular quality principle, with ensuing lesser relevance. However, in today’s medical world, time is a valuable commodity, and in fact, a lack of time or too much effort being required to adequately assess all relevant aspects is often mentioned as a barrier both to accessing information [55] as well as to employing apps in specific situations (eg, for consultations [57]). Although health apps may initially give the impression of being able to save time and reduce effort, professional ethics (eg, [58,59]) demand that those working in medical professions must ensure that any (digital) tools they use are up to the expected professional standards. In the digital world, even aided by various tools meant to aid in the process, health care professionals often remain unsure of which factors they need to consider in this context, especially if the tools require extensive effort. This may possibly contribute to the many—real or perceived—barriers toward successfully using apps in care settings or for health-related purposes in general.

Of course, an all-encompassing, unaided, and professionally conducted evaluation of apps will neither be possible nor practical in most scenarios, largely due to a lack of technical expertise. However, physicians and other health care professionals should at least be enabled to assess available
information in the context of their work, such as based on a set of questions [19] that address basic quality principles. Even for such limited lists, being able to determine a ranking of the questions or quality principles seems sensible for assessing highly available information with priority; if the initially evaluated factors already lead to a rejection, the remaining factors can justifiably be disregarded, thus saving the time that a full, structured assessment based on such questions covering all available information sources (eg, from the app store, on manufacturer websites, and other sources) would take. For longer lists of quality principles or rating criteria applied to mHealth apps, the benefits of being able to determine a sensible and context-adapted prioritization, based on feedback obtained from the respective peer group, may be even greater, counteracting or at least somewhat alleviating arguments that many of the available rating tools or quality principles are—due to the large number of details they cover—too cumbersome for real-world applications outside of academic evaluations [22].

In contrast to other approaches based on the Kano method (eg, [23,25]) that predominantly strive for categorization of the attributes being evaluated, the methodology presented here may provide an interesting option that additionally allows for the prioritization of quality principles in cases of largely similar categorization results or initial user perceptions. This may aid in giving precedence to the most relevant (prioritized) principles, deferring those with lesser priority. To what extent the method will be applicable beyond the usage scenario described here will require more extensive investigations.

However, it also remains an open question as to how one could deal with cases where for a larger number of attributes, there are multiple close clusters of attributes found in different quadrants. One possible solution to this might be to sort attributes in each cluster as described above, and to then perform a prioritization of the clusters themselves (with attributes in the attractive quadrant probably being the most relevant) in order to arrive at a full ranking of all attributes to be considered.

Nevertheless, the proposed prioritization may provide a means for professional organizations that want to give their members a recommendation as to which quality principles should be applied with priority in digital domains, independent of whether this is done for the generic set of app-related quality principles or principles that are more subject-specific (eg, for use in a particular medical specialty or for a specific user group).

Acknowledgments
The authors would like to thank the DGOU and the BVOU for the logistical support of the survey. Special thanks also go to Prof Bernhard Breil for the valuable discourse.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Additional tables.
[PDF File (Adobe PDF File), 1735 KB - mhealth_v10i1e26563_app1.pdf ]

References


Abbreviations

BVOU: Berufsverband für Orthopädie und Unfallchirurgie (Professional Association for Orthopedics and Trauma Surgery)
DGOU: Deutsche Gesellschaft für Orthopädie und Unfallchirurgie (German Society for Orthopedics and Trauma Surgery)
DVG: German Digital Healthcare Act
mHealth: mobile health

Edited by L Buis; submitted 18.12.20; peer-reviewed by P van der Keylen, S Badawy, E Omaki; comments to author 15.04.21; revised version received 10.06.21; accepted 05.11.21; published 11.01.22.

Please cite as:
Malinka C, von Jan U, Albrecht UV. Prioritization of Quality Principles for Health Apps Using the Kano Model: Survey Study. JMIR Mhealth Uhealth 2022;10(1):e26563
URL: https://mhealth.jmir.org/2022/1/e26563
doi: 10.2196/26563
PMID:35014965

©Christin Malinka, Ute von Jan, Urs-Vito Albrecht. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 11.01.2022. 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 App to Improve Oral Anticoagulation Adherence in Patients With Atrial Fibrillation: Prospective Observational Study

Keitaro Senoo, MD; Tomonori Miki, MD, PhD; Takashi Ohkura, MD; Hibiki Iwakoshi, MD; Tetsuro Nishimura, MD; Hirokazu Shiraishi, MD, PhD; Satoshi Teramukai, PhD; Satoaki Matoba, MD, PhD

1Department of Cardiac Arrhythmia Research and Innovation, Graduate School of Medical Science, Kyoto Prefectural University of Medicine, Kyoto, Japan
2Department of Cardiovascular Medicine, Graduate School of Medical Science, Kyoto Prefectural University of Medicine, Kyoto, Japan
3Departments of Biostatistics, Graduate School of Medical Science, Kyoto Prefectural University of Medicine, Kyoto, Japan

Corresponding Author:
Keitaro Senoo, MD
Department of Cardiac Arrhythmia Research and Innovation
Graduate School of Medical Science, Kyoto Prefectural University of Medicine
Kyoto-shi, Kamigyo-ku Kajii-cho 465, Kawaramachi-Hirokoji
Kyoto, 602-8566
Japan
Phone: 81 8031117168
Email: swcqg251@yahoo.co.jp

Abstract

Background: Poor adherence to oral anticoagulation in elderly patients with atrial fibrillation (AF) has been shown to negatively impact health care costs, morbidity, and mortality. Although various methods such as automated reminders, counseling, telephone support, and patient education have been effective in improving medication adherence, the burden on health care providers has been considerable. Recently, an attempt has been made to improve medication adherence without burdening health care providers by using smartphone apps; however, the use of the app for elderly patients with AF is still limited.

Objective: The purpose of this study was to determine whether the newly developed smartphone app for patients with AF (the Smart AF), which integrates education, automatic reminder, and patient engagement strategies with a simple user interface, can improve medication adherence in elderly patients with AF.

Methods: Patient enrollment was carried out by obtaining informed consent from patients with AF attending Kyoto Prefectural University of Medicine hospital between May 2019 and September 2020. Follow-up was planned at 1, 3, and 6 months after enrollment, and questionnaire reminders were automatically sent to patient apps at designated follow-up time points. A questionnaire-based survey of medication adherence was performed electronically using the self-reported 8-item Morisky Medication Adherence Scale (MMAS-8) as the survey tool.

Results: A total of 136 patients with AF were enrolled in this study. During the follow-up period, 112 (82%) patients underwent follow-up at 1 month, 107 (79%) at 3 months, and 96 (71%) at 6 months. The mean age of the enrolled patients was 64.3 years (SD 9.6), and male participants accounted for 79.4% (108/136) of the study population. The mean CHADS2 (congestive heart failure, hypertension, age, diabetes, previous stroke, or transient ischemic attack) score was 1.2, with hypertension being the most common comorbidity. At the time of enrollment, 126 (93%) and 10 (7%) patients were taking direct oral anticoagulants and warfarin, respectively. For medication adherence as measured according to the MMAS-8, MMAS scores at 1 month, 3 months, and 6 months were significantly improved compared with baseline MMAS scores (all P values less than .01). The overall improvement in medication adherence achieved by the 6-month intervention was as follows: 77.8% (14/18) of the patients in the high adherence group (score=8) at baseline remained in the same state, 45.3% (24/53) of the patients in the medium adherence group (score=6 to <8) at baseline moved to the high adherence group, and 72% (18/25) of the patients in the low adherence group (score <6) moved to either the medium or high adherence group.

Conclusions: The Smart AF app improved medication adherence among elderly patients with AF. In the realm of medication management, an approach using a mobile health technology that emphasizes education, automatic reminder, and patient engagement may be helpful.

(JMIR Mhealth Uhealth 2022;10(1):e30807) doi:10.2196/30807

https://mhealth.jmir.org/2022/1/e30807
KEYWORDS
atrial fibrillation; smartphone app; anticoagulants; drug adherence; education; patient involvement

Introduction

Atrial fibrillation (AF), the most common chronic arrhythmia, affected 33.5 million people worldwide, in 2010, and the number is projected to double by 2030 [1,2]. AF is associated with a 5-fold increased risk of stroke, and oral anticoagulation (OAC) is required for people at moderate-to-high risk of stroke [3]. Poor adherence to medication in the real world can alter efficacy and safety estimates from randomized controlled trials, leading to poorer health outcomes and greater health care costs [4,5]. A recent meta-analysis found that suboptimal adherence to and persistence with direct oral anticoagulants (DOACs) is common [6]. For example, patients with AF do not take a DOAC every 4 days; one-third of the patients have less than 80% adherence; real-world persistence with DOACs is lower than in randomized controlled trials; and patients with poor adherence have a higher risk of stroke.

To date, various greater efforts in monitoring and interventions have been used to improve OAC adherence. For instance, the AEGEAN (Assessment of an Education and Guidance Programme for Eliquis Adherence in Non-Valvular Atrial Fibrillation) trial explored the impact of education (ie, using booklets and reminder tools) and telephone follow-up using a virtual clinic on adherence to apixaban. However, in the AEGEAN trial, electronically measured adherence did not differ between the usual care and intervention groups, with adherence rates of 88.5% and 88.3%, respectively [7]. A study by Shore et al [8] reported that enhanced pharmacist engagement and longer patient monitoring and follow-up were associated with greater adherence to dabigatran. In a study by Desteghe et al [9], remote monitoring of daily medication uses and daily remote monitoring and individualized feedback by telephone resulted in very high adherence to DOACs, with 99.0% demonstrating adherence and 96.8% demonstrating regimen adherence. FACILITA (strategies for improving dabigatran adherence for stroke prevention in patients with non-valvular atrial fibrillation) study revealed that a mixed intervention, consisting of patient education and a simple calendar reminder for drug intake, was an effective strategy to improve adherence to dabigatran (to 91% and 89% at 6 and 12 months, respectively, compared with 65% and 63% for the control group) [10]. To improve medication adherence, various methods such as reminders, counseling, telephone support, and patient education are effective; however, these long-term interventions impose a considerable burden on health care providers [11,12].

Recently, attempts have been made to increase medication adherence without burdening health care providers by using smartphone apps in various fields [13-15]. The Health Buddies app (DAE Studios) was developed as a tool to improve adherence by providing a virtual contract with the patients’ grandchildren [16]. The mAIF app was developed to integrate clinical decision support, education, and patient-involvement strategies [17]. Although some success was achieved in these studies, there were problems such as the complexity of the app user interface.

The purpose of this study was to determine whether the newly developed smartphone app for patients with AF (the Smart AF app), which integrates education, automatic reminder, and patient engagement strategies with a simple user interface, can improve medication adherence in elderly patients with AF.

Methods

Features of the Smart AF App

The Smart AF app was developed by Health Tech Innovation Center, in association with Kyoto Prefectural University of Medicine, funded by the BMS/Pfizer Japan Thrombosis Investigator Initiated Research Program (JRISTA). There are 3 features of the Smart AF app (Figure 1).

Figure 1. Features of the smartphone app for atrial fibrillation (the Smart AF app).

Educational Program

There are 7 components to the patient educational program. With videos that are approximately 1-2 minutes in length, patients can learn about AF and learn self-management methods, including how to detect and treat AF, the importance of anticoagulation, and the treatment of comorbidities.

Patient Engagement

After inputting information about the characteristics of their AF history, other medical history, medication information (eg,
type of antithrombotic medication or other concomitant medications), and lifestyle behaviors, each patient’s CHADS2 score is automatically calculated, enabling physicians to easily understand their risk for stroke. Furthermore, the app can take inputs on step counts, sleep time, and presence or absence of symptoms and, by sharing this information with health care providers, the proactive involvement of the patient in their own care can be promoted and supported.

**Reminder Alarm**
A reminder is automatically transmitted through the app in the morning and evening daily to prevent forgetting to take medication.

**Outcome Measures**
Follow-up was planned at 1, 3, and 6 months after enrollment, and a reminder email for the survey was automatically sent to the patient’s app at the time of follow-up. The self-reported 8-item Morisky Medication Adherence Scale (MMAS-8) was used as the survey instrument [18-20]. The MMAS-8 score assesses patients’ self-reported adherence to their anticoagulant medication. According to the MMAS-8 score (range 0-8), adherence was defined as high (score 8), medium (score 6 to <8) or low (score <6).

**Eligibility Criteria for Participants**
Patient enrollment was carried out by obtaining informed consent from patients with AF attending Kyoto Prefectural University of Medicine hospitals between May 2019 and September 2020. All participants provided electronic informed consent and were assigned a password. Inclusion criteria were as follows: documented diagnosis of AF; current prescription for OACs (ie, dabigatran, rivaroxaban, apixaban, edoxaban, and warfarin) for at least 3 weeks; and ownership of a mobile phone. Individuals less than 20 years of age, those with valvular AF, and those who had been taking OACs for less than 3 weeks were excluded. This single-center prospective observational study was approved by the Institutional Review Board of the Kyoto Prefectural University of Medicine (ERB-C-1429).

**Data Analysis**
Statistical analyses were performed using R version 3.6.1 (R Foundation for Statistical Computing). Continuous variables are expressed as mean and SD, and categorical variables as number and percentage. MMAS-8 scores at 4 time points (baseline, 1, 3, and 6 months later) were plotted to illustrate changes in these variables over time. The Wilcoxon signed-rank test was used to evaluate differences in MMAS-8 scores between baseline and 1, 3, and 6 months. For the 96 patients who were able to complete the follow-up up to 6 months, the change in adherence from the time of enrollment to 6 months is reflected graphically by a low, middle, and high MMAS score. The percentage of patients who activated the app at least once per day is expressed as the mobile app retention rate, and the mean retention rate with SD per week at 1, 3, and 6 months was calculated and graphed. Associated factors were also identified using simple regression analysis. “Low retention,” defined as the percentage of days in which the app was activated at least once, was ≤10% during the observation period. Differences with P<.05 were considered to be statistically significant.

This study was funded by JRISTA. Permission for use of the MMAS-8 scale and its coding has been acquired, and a license agreement is available from MMAR, LLC, Donald E Morisky, ScD, ScM, MSPH, 294 Lindura Ct., United States.

**Results**
Between May 2019 and September 2020, a total of 136 patients with AF were enrolled in this study. During the follow-up period, 112 (82%) patients underwent the follow-up survey at 1 month, 107 (79%) at 3 months, and 96 (71%) at 6 months. The mean age of the enrolled patients was 64.3 years (SD 9.6), and males accounted for 79.4% (108/136) of the study population; 89.7% (122/136) of the patients were married and had family members currently residing with them; 63.9% (87/136) of the patients had attended college as their highest level of education, and 57.3% (78/136) were currently working; 75% (102/136) were currently practicing regular dietary habits, and 19.8% (27/136) engaged in daily exercise; 74.2% (101/136) of the patients had a history of AF treatment for ≥1 year, and hypertension was the most common comorbidity, with a mean CHADS2 score of 1.2. Moreover, 38.9% (53/136) of the patients experienced palpitation symptoms, 130 (96%) patients with a Hospital Anxiety and Depression Scale (HADS)-A score of ≥8 (major anxiety), 30 (22%) patients with a HAD-D score of ≥11 (major depression), and 95 (70%) patients with a HADS-T score of ≥20 (major anxiety and depression). At enrollment, 126 (93%) patients were taking DOAC, and 10 (7%) were taking oral warfarin (Table 1).

For medication adherence, as measured according to the MMAS-8, the low adherence group was 27.9% (n=38) at baseline, 16.0% (n=18) at 1 month, 15.9% (n=17) at 3 months, and 12.5% (n=12) at 6 months. Medium adherence was 55.1% (n=75) at baseline, 43.8% (n=49) at 1 month, 44.9% (n=48) at 3 months, and 44.8% (n=43) at 6 months. High adherence group was 16.9% (n=23) at baseline, 40.2% (n=45) at 1 month, 39.3% (n=42) at 3 months, and 42.7% (n=41) at 6 months, respectively. Compared with baseline MMAS scores, MMAS scores at 1 month, 3 months, and 6 months were significantly improved (all P<0.01) (Figure 2).

Furthermore, the overall improvement in medication adherence achieved by the intervention was as follows: 77.8% (14/18) of the patients in the high adherence group at baseline remained there; 45.3% (24/53) of the patients in the medium adherence group at baseline moved to the high adherence group; and 72% (18/25) of the patients in the low adherence group moved to either the median or high adherence groups (Figure 3).

The mobile app retention rate is a plot of the percentage of patients who activate the app at least once per day (Figure 4).
Table 1. Demographics of the participants (N=136).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>64.3 (9.6)</td>
</tr>
<tr>
<td>Male, n (%)</td>
<td>108 (79.4)</td>
</tr>
<tr>
<td>Married, n (%)</td>
<td>122 (89.7)</td>
</tr>
<tr>
<td>Living together, n (%)</td>
<td>123 (90.4)</td>
</tr>
<tr>
<td>University education level, n (%)</td>
<td>87 (64.0)</td>
</tr>
<tr>
<td>Full-time or part-time employed</td>
<td>78 (57.4)</td>
</tr>
<tr>
<td>Regular meal, n (%)</td>
<td>102 (75.0)</td>
</tr>
<tr>
<td><strong>Habit of exercise, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Every day</td>
<td>27 (19.9)</td>
</tr>
<tr>
<td>Sometimes</td>
<td>54 (39.7)</td>
</tr>
<tr>
<td>None</td>
<td>55 (40.4)</td>
</tr>
<tr>
<td>Duration of atrial fibrillation (&gt;1year)</td>
<td>101 (74.3)</td>
</tr>
<tr>
<td><strong>Type of oral anticoagulants, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>DOAC(^a), OD(^b) (edoxaban, rivaroxaban)</td>
<td>86 (63.2)</td>
</tr>
<tr>
<td>DOAC, BID(^c) (apixaban, dabigatran)</td>
<td>40 (29.4)</td>
</tr>
<tr>
<td>Warfarin</td>
<td>10 (7.4)</td>
</tr>
<tr>
<td>CHADS(^d) score (SD)</td>
<td>1.2 (1.1)</td>
</tr>
<tr>
<td>Congestive heart failure, n (%)</td>
<td>18 (16.5)</td>
</tr>
<tr>
<td>Hypertension, n (%)</td>
<td>65 (59.6)</td>
</tr>
<tr>
<td>Age ≥75 (years), n (%)</td>
<td>13 (9.6)</td>
</tr>
<tr>
<td>Diabetes mellitus, n (%)</td>
<td>19 (17.4)</td>
</tr>
<tr>
<td>Prior stroke or TIA(^e), n (%)</td>
<td>9 (8.3)</td>
</tr>
<tr>
<td>With symptom, n (%)</td>
<td>53 (39.0)</td>
</tr>
<tr>
<td><strong>HADS(^f) scale, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>HADS-A (≥8—major anxiety; total 21 points)</td>
<td>130 (95.6)</td>
</tr>
<tr>
<td>HADS-D (≥11—major depression; total 21 points)</td>
<td>30 (22.1)</td>
</tr>
<tr>
<td>HADS-T (≥20—major anxiety and depression; total 42 points)</td>
<td>95 (69.9)</td>
</tr>
</tbody>
</table>

\(^a\)DOAC: direct oral anticoagulants.
\(^b\)OD: once a day.
\(^c\)BID: twice a day.
\(^d\)CHADS\(^2\): congestive heart failure, hypertension, age, diabetes, previous stroke, or transient ischemic attack.
\(^e\)TIA: transient ischemic attack.
\(^f\)HADS: Hospital Anxiety and Depression Scale.
Retention rates for 1, 3, and 6 months were 0.48 (SD 3), 0.36 (SD 3), and 0.27 (SD 1), respectively. In the univariate regression analysis, predictors for a higher retention rate were older age ($P=.047$), regular meals ($P=.04$), and HADS-A ≥8 (major anxiety, $P=.04$), respectively (Table 2).
Table 2. Predictors for retention rate.

<table>
<thead>
<tr>
<th>Characteristics and variables</th>
<th>Values</th>
<th>( P ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.605</td>
<td>.047</td>
</tr>
<tr>
<td>Male</td>
<td>-0.442</td>
<td>.95</td>
</tr>
<tr>
<td>Married</td>
<td>-4.134</td>
<td>.67</td>
</tr>
<tr>
<td>Living together</td>
<td>-0.349</td>
<td>.97</td>
</tr>
<tr>
<td>Education level</td>
<td>-1.756</td>
<td>.77</td>
</tr>
<tr>
<td>Employed</td>
<td>0.096</td>
<td>.99</td>
</tr>
<tr>
<td>Regular meal</td>
<td>13.56</td>
<td>.04</td>
</tr>
</tbody>
</table>

**Habit of exercise**

- Every day: 3.023, \( P = .68 \)
- Sometimes: 3.871, \( P = .52 \)
- None: -5.845, \( P = .33 \)
- Duration of atrial fibrillation (>1 year): -2.087, \( P = .76 \)
- Oral anticoagulants, BID\(^b\) (reference as OD\(^c\)): 1.538, \( P = .81 \)
- CHADS\(^d\)2 score: -0.257, \( P = .90 \)
- Polypharmacy: -0.376, \( P = .98 \)
- With symptom: -1.533, \( P = .80 \)

**HADS\(^e\) scale**

- HADS-A (≥8—major anxiety): 28.875, \( P = .04 \)
- HADS-D (≥11—major depression): -10.946, \( P = .12 \)
- HADS-T (≥20—major anxiety and depression): -2.533, \( P = .69 \)

\(^a\)B: univariate regression analysis.
\(^b\)BID: twice a day.
\(^c\)OD: once a day.
\(^d\)CHADS2: congestive heart failure, hypertension, age, diabetes, previous stroke, or transient ischemic attack.
\(^e\)HADS: Hospital Anxiety and Depression Scale.

**Discussion**

**Main Findings**

The Smart AF app has been developed, which integrates patient education, reminder alarm, and patient engagement strategies without increasing the burden on health care providers. Elderly patients with AF who used the Smart AF app demonstrated significantly improved MMAS-8 at 6 months compared to baseline. The Smart AF app demonstrated a significant decrease in the low adherence group and an increase in the high adherence group at 6 months compared with baseline.

**Improvement of Adherence**

Current clinical guidelines for AF advocate incorporating patient preferences regarding treatment and support and involving patients in management decisions [21,22]. The participation of patients with AF in the process of developing and refining the app through patient involvement resulted in the Smart AF app. For example, to facilitate the use of the app, age-related aspects considered to be useful in the design of mHealth (mobile health) tools, including large screens, large fonts, and ease of navigation, were incorporated into it. As a result, we adopted a simple layout and large navigation buttons that are easy to use, even by elderly individuals. Among the few studies that evaluated the effectiveness of smartphone apps in improving medication adherence in those with AF, research investigating the mAF app revealed that it had a complex design with a great deal of content, such as education programs, clinical decision support material, patient involvement in self-care, structured follow-up, and many areas that users needed to manipulate manually [17]. Regarding the Health Buddies App, there are some problems, such as a lack of interest in the characteristic games of the app, to the point that grandchildren who become “Buddies” must also participate; moreover, although achieving tasks and goals inside the app was fun, long-term behavioral change was difficult because these achievements do not correlate directly with patient health conditions [16].

The Smart AF app was developed with a focus on patient education, reminder alarms, and patient engagement in self-care, and we tried to keep the app simple without any other features.
In fact, the FACILITA study also revealed that a mixed intervention, consisting of patient education and a simple calendar reminder of drug intake, was an effective strategy to improve adherence to dabigatran (to 91% and 89% at 6 and 12 months, respectively, compared with 65% and 63% for the control group) [10]. Therefore, despite the simplicity of the app’s contents, we believe that not only did the reminder alarms encourage patients to take their daily medication, but by having them record their medication in the app’s calendar, we could encourage their participation and make them more aware of their engagement in self-care by managing their daily health.

Previous studies have also reported that tailor-made educational interventions can significantly improve anticoagulation management of warfarin [23]. In addition, the IMPACT-AF (Integrated Management Program Advancing Community Treatment of Atrial Fibrillation) study, which investigated the use of oral anticoagulation in patients with AF, found that it was improved by a multifaceted, multilevel (including at medication initiation as well as at follow-up) educational program implemented by physicians [24]. The Smart AF app has educational contents to help patients quickly solve their questions about AF. Thus, the smart AF app can enhance patient education and medication reminders at any time and place. In addition, we believe that we have succeeded in encouraging long-term patient engagement in self-care by avoiding complex operations and time-consuming input as much as possible.

Retention Rate of the Smart App

The decrease in the retention rate of the app over time is an important issue to be considered. The mean retention rates at 1, 3, and 6 months were 0.48 (SD 3), 0.36 (SD 3), and 0.27 (SD 1), respectively. A decline in app use over time is also a concern that has been highlighted in previous reports [25].

Although the app offers the advantage of completely remote recruitment and enrollment, lack of human communication may mean less motivation for the participants to continue compared with studies conducted face-to-face.

The Smart AF app was designed to pop up notifications (reminder) every morning and evening, even when the app had not been activated or opened. Such pop-up feature may have succeeded in improving medication adherence regardless of whether the app was activated or not. However, long-term use is essential for the app to affect users. Clearly, additional efforts to improve retention rates are necessary. The characteristics associated with the retention time of apps have not been well studied. In a study of patients with asthma, being female and older was related to longer retention [26]. In our study, 95 patients enrolled in the study had an HADS-T score of 20 or higher (major anxiety and depression), and most of them had anxiety and depression. As shown in Table 2, older patients, patients with a regular diet, and patients with anxiety had higher retention rates. Conversely, younger patients, those with irregular diets, and those without anxiety had lower retention rates. To deliver mHealth effectively, it is important to identify patient domain factors, such as psychological factors, dietary regularity, and age, that mark suitable candidates for the app.

In the future, to further increase continued app usage rates, we believe that we must work to incorporate an interactive design; more specifically, design elements that respond immediately to patient operations and behaviors.

Limitations

There were limitations to our study and findings. First, this was a single-center study, and the patients were not prospectively randomized into intervention (the app group) and usual care groups. Second, while important in this type of study, the reliance on measurement of self-reported results is a challenge often encountered. The data obtained from these indices may be supported by more objective observations. For example, we used a self-reported medication adherence tool to examine medication adherence. Older adults taking multiple medications may not be able to accurately report their medication use status due to poor memory or confusion. Therefore, it may be useful to complement self-reported measures using more objective measures of medication adherence (eg, medical record review and pharmacy documentation). Third, other new interventions, strategies, and technologies designed to enhance long-term adherence to DOACs need to be developed and investigated because patients with AF are a large and diverse patient population, and not all will have access to newer mHealth tools. Nonadherence is often caused by a multitude of factors, indicating the necessity of providing patients with tailored and more personalized tools. Lastly, this app only showed an improvement in adherence for 6 months; therefore, future studies are needed for long-term improvement.

Conclusion

The Smart AF app improved medication adherence among elderly patients with AF. In the realm of medication management, an approach using an mHealth technology that emphasizes education, automatic reminder, and patient engagement may be helpful. The challenge that emerged, however, was the decline in the rate of persistent use of the app over time; therefore, continuous doctor-patient interaction via the app will be necessary in the future.

Conflicts of Interest

KS is an executive director of the General Incorporated Association of Health Tech Innovation Center, Osaka, Japan.

References


Abbreviations

- AEGEAN: Assessment of an Education and Guidance Programme for Eliquis Adherence in Non-Valvular Atrial Fibrillation
- AF: atrial fibrillation
- CHADS2: congestive heart failure, hypertension, age, diabetes, previous stroke, or transient ischemic attack
- DOAC: direct oral anticoagulant
- FACILITA: strategies for improving dabigatran adherence for stroke prevention in patients with non-valvular atrial fibrillation
- HADS: Hospital Anxiety and Depression Scale
- IMPACT-AF: Integrated Management Program Advancing Community Treatment of Atrial Fibrillation
- JRISTA: the BMS/Pfizer Japan Thrombosis Investigator Initiated Research Program
- mHealth: mobile health
- MMAS-8: 8-item Morisky Medication Adherence Scale
- OAC: oral anticoagulation

©Keitaro Senoo, Tomonori Miki, Takashi Ohkura, Hibiki Iwakoshi, Tetsuro Nishimura, Hirokazu Shiraishi, Satoshi Teramukai, Satoaki Matoba. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 07.01.2022. 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.
App-Based Relaxation Exercises for Patients With Chronic Neck Pain: Pragmatic Randomized Trial

Daniel Pach1,2, MD; Susanne Blödt1, PhD; Jiani Wang1, MSc; Theresa Keller3, PhD; Beatrice Bergmann1, MD; Alizé A Rogge1, MSc; Jürgen Barth2, PhD; Katja Icke1; Stephanie Roll1, PhD; Claudia M Witt1,2,4, MD, MBA

1Institute of Social Medicine, Epidemiology and Health Economics, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany
2Institute for Complementary and Integrative Medicine, University Hospital Zurich and University of Zurich, Zurich, Switzerland
3Institute of Biometry and Clinical Epidemiology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany
4Center for Integrative Medicine, University of Maryland School of Medicine, Baltimore, MD, United States

Corresponding Author:
Claudia M Witt, MD, MBA
Institute for Complementary and Integrative Medicine
University Hospital Zurich and University of Zurich
Sonneggstrasse 6
Zurich, 8091
Switzerland
Phone: 41 44 2552396
Email: claudia.witt@uzh.ch

Abstract

Background: Chronic neck pain is a highly prevalent condition. Learning a relaxation technique is recommended by numerous guidelines for chronic neck pain. Smartphone apps can provide relaxation exercises; however, their effectiveness, especially in a self-care setting, is unclear.

Objective: The aim of this pragmatic randomized trial is to evaluate whether app-based relaxation exercises, including audio-based autogenic training, mindfulness meditation, or guided imagery, are more effective in reducing chronic neck pain than usual care alone.

Methods: Smartphone owners aged 18 to 65 years with chronic (>12 weeks) neck pain and the previous week’s average neck pain intensity ≥4 on the Numeric Rating Scale (0=no pain to 10=worst possible pain) were randomized into either an intervention group to practice app-based relaxation exercises or a control group (usual care and app for data entry only). For both groups, the follow-up data were collected using app-based diaries and questionnaires. The primary outcome was the mean neck pain intensity during the first 3 months based on daily measurements. Secondary outcomes included neck pain based on weekly measurements, pain acceptance, neck pain–related stress, sick-leave days, pain medication intake, and adherence, which were all measured until the 6-month follow-up. For the primary analysis, analysis of covariance adjusted for baseline neck pain intensity was used.

Results: We screened 748 participants and enrolled 220 participants (mean age 38.9, SD 11.3 years; mean baseline neck pain 5.7, SD 1.3 points). The mean neck pain intensity in both groups decreased over 3 months; however, no statistically significant difference between the groups was found (intervention: 4.1 points, 95% CI 3.8-4.4; control: 3.8 points, 95% CI 3.5-4.1; group difference: 0.3 points, 95% CI −0.2 to 0.7; P=.23). In addition, no statistically significant between-group differences regarding neck pain intensity after 6 months, responder rate, pain acceptance, pain medication intake, or sick-leave days were observed. There were no serious adverse events that were considered related to the trial intervention. In week 12, only 40% (44/110) of the participants in the intervention group continued to practice the exercises with the app.

Conclusions: The study app did not effectively reduce chronic neck pain or keep the participants engaged in exercising in a self-care setting. Future studies on app-based relaxation interventions should take into account the most recent scientific findings for behavior change techniques.

Trial Registration: ClinicalTrials.gov NCT02019134; https://clinicaltrials.gov/ct2/show/NCT02019134

International Registered Report Identifier (IRRID): RR2-10.1186/1745-6215-15-490
Introduction

Neck pain is a global public health issue entailing a high socioeconomic burden [1,2]; moreover, it is one of the top 5 global chronic pain conditions in terms of prevalence and cause of disability [3,4]. According to the data from the European Social Survey 2014 [5], approximately 40% of all respondents reported back or neck pain. These results indicated the highest prevalence of back or neck pain in Germany (54.05%).

In most cases, neck pain is nonspecific [1]. Hence, the treatment is complex and costly. Pharmacological approaches are often used to alleviate chronic pain; however, these approaches include possible risks of tolerance, dependence, and addiction when using opioids [6,7]. Moreover, previous research showed that exercise treatment might also be beneficial in patients with neck pain [3].

Mind–body therapies focus on the interactions among the brain, mind, body, and behavior and their effects on health and disease [8]. As components of mind–body medicine, relaxation techniques have gained wide acceptance within conventional medicine [9]. The relaxation response leads to a variety of physiological benefits that may enhance pain relief through reduced sympathetic activity, decreased muscular tension, modulated pain awareness, and increased release of endogenous opioids [10,11]. Studies directly comparing the effects of self-administered versus therapist-administered interventions found similar effects on pain reduction [12]. Moreover, according to the recent Neck Pain Guideline of the German Society of General Practice and Family Medicine [13], learning a relaxation technique is recommended for patients with nonspecific chronic neck pain that lasts for >12 weeks. Thus, relaxation techniques alone or in addition to conventional medical care can influence the treatment and rehabilitation of chronic neck pain. However, the accessibility of cognitive and mind–body therapies for chronic low back pain and neck pain remains a major challenge [14].

Medical smartphone apps or other mobile digital health solutions can allow easy access to self-care activities [15] and support behavior changes by incorporating features such as the provision of information, tracking of activity, or providing feedback. A review [16] identified 606 mindfulness apps; however, only 3.8% (23/606) of those apps actually provided mindfulness training, and only 1 app [17] was evaluated in a randomized controlled trial (RCT). Another review [8] on apps with self-management support functions for people with persistent pain identified only 2 evidence-based apps; however, none of them were for chronic pain.

In this study, we aim to conduct a pragmatic app-based RCT to evaluate whether app-based audio relaxation exercises are more effective in reducing chronic neck pain than usual care.

Methods

Study Design

The trial design and methods have been published elsewhere [18] and have not been changed afterward. The study app remained frozen without any updates during the trial. We conducted a 2-armed, randomized, parallel-group, single-center pragmatic trial to investigate the effectiveness of additional relaxation exercises delivered by a smartphone app compared with usual care alone. Participants were randomized in a 1:1 ratio to either the app-based relaxation intervention group or the control group. The trial flow is presented in Figure 1.

The intervention duration was 6 months, with the primary outcome summarizing the effect of the first 3 months.
Participants and Setting
The first participant was randomized on March 31, 2014, and the final data recording was on January 11, 2017, in Berlin, Germany. Information on the study was posted with brochures and posters in universities, gyms, and general practitioners’ offices. Moreover, the study was advertised in local subways from December 2014 to July 2015. Eligibility was checked by a study nurse at the study site. Eligible participants completed the paper-and-pencil baseline questionnaires. Then, the study nurse helped the participants install the app on their own smartphones and provided a randomly allocated code to activate the study app and the respective app features according to the group allocation. Participants received compensation of €20 (US $22.60) after participating in the study.

The inclusion criteria were as follows: aged 18-65 years, chronic neck pain within at least the past 12 weeks, average neck pain intensity ≥4 on the Numeric Rating Scale (NRS; 0=no pain to 10=worst possible pain) in the previous week, possession of a smartphone (iOS or Android), willingness to be randomized and follow the app-delivered interventions, and willingness to enter data through the study app.

Participants were excluded if their neck pain was caused by a known malignant disease, trauma, the presence of a known rheumatic disorder, a history or planned surgery of the spinal column of the lower neck in the next 6 months, known neurological symptoms (e.g., radicular symptoms because of a prolapsed disk), regular intake of analgesics (more than once per week) because of additional disease, intake of centrally acting analgesics, or a history of severe acute or chronic disorders that did not allow participation in the study.

Further exclusion criteria were known alcohol or substance abuse, insufficient German language skills, current application for a pension claim, participation in another clinical trial during the 6 months before the study and parallel to the study, applying regular relaxation techniques, mindfulness meditation, or any other mindfulness-based therapy 6 weeks before the study or planned in the next 6 months.

Participants in both groups were allowed to continue with their usual care (medical and nonmedical); however, the regular application of any other relaxation techniques, including mindfulness meditation or mindfulness-based training, was not permitted.

The follow-up data (daily, weekly, and at the third and sixth month) were collected through the app-based questionnaires and by in-app tracking of the length of the practiced exercises. Serious adverse events were documented during the study period to evaluate safety.
The Relaxneck App

Overview

The study app Relaxneck was developed by the Institute of Social Medicine, Epidemiology and Health Economics, Charité–Universitätsmedizin Berlin, Germany, together with Smart Mobile Factory, Berlin, Germany, which is an agency focused on mobile solutions [18]. The app supported iOS and Android systems and was available in the German Apple Appstore and the Google Play Store free of charge. However, the app could only be activated by entering an individual code assigned to each study participant by the study nurse.

The app supported notification features, a diary, and questionnaire options for all participants, whereas it provided audio relaxation exercises only for those in the intervention group. The app’s user interface and content were available in the German language (Figure 2). The app concept was approved by the data protection officer of the Charité–Universitätsmedizin Berlin.

Figure 2. Screenshots of the study app (dashboard, relaxation exercises, and questionnaires).

App-Based Relaxation Interventions

Overview

The duration of the audios for the relaxation interventions, as well as their intensity and dosage; the use of push notifications; the diary content: and the German translation of guided imagery instructions resulted from stakeholder engagement during the planning phase of the study [18].

There were 3 types of exercises (autogenic training, mindfulness meditation, and guided imagery), with a length of 15 minutes each, that were available in 2 versions (female and male voices) in the study app for the intervention group. They were accompanied by a short instructional text (Figure 2). Relaxation exercises could be applied in different positions (sitting, walking, and lying) according to the participants’ needs. It was recommended to apply a relaxation exercise daily or at least 5 days per week for 6 months.

Autogenic Training

Autogenic training is a form of self-relaxation technique that is commonly used to treat stress disorders, pain, and anxiety [19-21]. Autogenic training was developed by the German psychiatrist Johannes Schultz in 1932. It focuses on the physical sensation of the breath or heartbeat and visualizes the body as warm, heavy, or relaxed [21]. Participants learn to react to 6 verbal commands, such as “my arms are very heavy,” “my heart beats regularly and calm,” and “my belly is warm,” to make the body feel relaxed [18].

Mindfulness Meditation

Mindfulness is a practice based on Vipassana (ie, insight) meditation, which has Buddhist roots. It is defined as “paying attention in a particular way: on purpose, in the present moment and in a nonjudgmental way” [22]. It focuses on the breath and uses it as an anchor when the mind starts to wander [18]. This concept is also used in mindfulness-based stress reduction developed by Kabat-Zinn [22-24].

Guided Imagery

In guided imagery, the mind is directed to intentionally create images to produce positive changes [25]. The audio guides the participants to visualize or conjure a place that is associated with positive feelings such as safety, security, and well-being. The guided imagery audio is accompanied by soft background music and directs visualization and imagination to a pleasant and peaceful place that has meaning for the participant to replace negative or stressful feelings [26].
Behavior Change Techniques in the App

To enhance changes in participants’ behavior, behavior change techniques (BCTs) can be implemented in intervention settings [27]. As this was not a common feature in app development in 2013, we retrospectively analyzed the Relaxneck app using the BCT taxonomy (version 1) by Michie et al [27] to identify BCTs that were represented in the app, although not formally preplanned.

App for the Control Group

Participants in the control group downloaded the same app as the intervention group. All study data after baseline measurements were collected by means of app-based diaries and questionnaires. The participants were able to activate reminders for the questionnaire notifications. However, no intervention features, that is, relaxation exercises, were accessible in their version of the app. The relaxation exercises were activated after 6 months after all the survey data were collected. In addition, participants could continue using usual care, defined as all medical and nonmedical treatments, while using the app; however, relaxation techniques, mindfulness meditation, or any other mindfulness-based trainings were not permitted to be practiced during the study.

Outcome Measurements

The primary outcome measure was the mean neck pain intensity during the first 3 months of intervention based on daily measurements of pain intensity on the NRS (0=no pain to 10=worst possible pain) [18].

The secondary outcome parameters included the mean pain intensity during the first 6 months after randomization based on daily measurements, the mean intensity measured weekly (using NRS) as the average pain intensity of the previous 7 days over 3 and 6 months, pain acceptance (German version of Chronic Pain Acceptance Questionnaire [28]), neck pain–related stress, sick-leave days, and pain medication intake. Data on adherence, self-reported general changes in neck pain, suspected adverse reactions, and serious adverse events were additionally collected [18].

If a weekly survey had not been completed, the patient received an SMS text message as a reminder; if 2 consecutive weekly surveys had not been completed, the patient was contacted by telephone call; if there was no response after 2 calls, the patient received a reminder letter.

The number of participants who practiced the exercises was recorded to reflect exercise adherence over time. Practice of the exercise was defined by (1) tracking the number (and duration) of applied types of intervention with the app and (2) asking the participants weekly about the number of applied types of intervention without using the app. The complete stop of filling in any data with the study app was defined as participant dropout. Adverse events and suspected adverse reactions (only in the intervention group) were assessed after 3 and 6 months.

Sample Size

According to previous literature [29], an effect size of 0.62 has been described for mind–body therapies compared with no intervention in a group setting. We assumed a smaller effect size of 0.4 (Cohen d, baseline adjusted) for individual self-care relaxation exercise compared with usual care alone, as individuals might be less focused and consequently less adherent in a self-care setting [18]. To obtain a power of 80% using a 2-sided t test with a significance level of .05, 100 participants for each treatment group were needed (a total of 200 participants). Thus, a final sample size of 110 participants per group (220 in total), allowing a dropout rate of 9.1%, was required.

Randomization, Allocation, and Implementation

Eligible participants were randomized to either the intervention (app-based relaxation and usual care) or the control (usual care only) group using blocked randomization with variable block lengths and an allocation ratio of 1:1, that is, 110:110 participants. The randomization sequence was generated by a data manager who was not involved in the analysis of the data or the enrollment of the patients; SAS (version 9.3, SAS Inc) was used for this process. The randomization list was included in a safe Microsoft Access database to ensure that it was not accessible during the randomization process of individual participants and that the screened patients were strictly consecutively enrolled. The randomization process was conducted by the study office at the Institute of Social Medicine, Epidemiology and Health Economics. To ensure allocation concealment, first, the study team added the participants’ information into the database, and then, random allocation of the participants into the intervention or control group was performed.

Statistical Analysis

For the primary analysis of the primary outcome (mean pain intensity over 3 months measured as the daily pain intensity), an analysis of covariance with a fixed factor of treatment group, adjusted for the baseline NRS value (fixed covariate), was performed. The analysis was based on the full analysis set (all available data without imputation of missing values, as only a small number of missing values was expected based on experiences with a previous app-based study conducted by our study team in a similar study setting [30]) based on the intention-to-treat principle with a 2-sided significance level of .05.

All the secondary analyses were explorative, and P values were interpreted as such. The secondary outcomes were analyzed for the full analysis set, similar to the primary analysis, depending on the scale and distribution of the outcome, that is, analysis of covariance or logistic regression, adjusted for the respective baseline value. For sensitivity analysis, the primary analysis of the primary outcome was repeated based on the per-protocol population.

Subgroup analyses were performed on the primary outcome by including an interaction term (subgroup variable by treatment) in the main model and performing separate analyses for each subgroup. Subgroups were specified with covariates in age, education (>10 years of school education or ≤10 years of school education), sex (male or female), and duration of disease. Kaplan–Meier survival analysis was conducted to investigate
whether the app features (with or without app-based intervention content) predicted the dropout of app use.

SAS version 9.4 (SAS Inc) was used for data analysis, except for the Kaplan–Meier survival analysis for adherence, which was conducted using SPSS version 22.0 (SPSS Inc).

**Ethics**

The study was approved by the local ethics review board at the Charité–Universitätsmedizin, Berlin (approval number Relaxneck EA 1/259/13). The study was conducted according to the common standard guidelines for clinical trials (Declaration of Helsinki and, where applicable, the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use and Good Clinical Practice revised version, Somerset West, Republic of South Africa, 1996).

All study participants provided oral and written informed consent. The trial was registered at ClinicalTrials.gov (NCT02019134), and the study protocol has been published elsewhere [18].

**Results**

**Baseline Characteristics**

Of the 748 screened participants, 220 (29.4%) were eligible for the study and gave informed consent. They were randomized either to the app-based intervention group (110/220, 50%) or to the usual care group (110/220, 50%).

The sociodemographic and clinical characteristics of the participants at baseline are presented in Table 1. The participants had a mean age of 38.9 (SD 11.3) years and an average education, with 70% (154/220) having ≥ 10 years of school education. Of the 220 participants, 35 (15.9%) participants had a migration background. In the previous 7 days, the average neck pain on the NRS was 5.7 (SD 1.3) points, and 26.8% (59/220) of participants had taken medication for neck pain.

Although both groups were comparable at baseline, we observed small differences regarding gender (intervention vs control: female 74/110, 67.3% vs 79/110, 71.8%), partnership status (56/110, 50.9% vs 66/110, 60%), migration background (14/110, 12.7% vs 21/110, 19.1%), duration of neck pain (mean 79.2, SD 74.8 months vs mean 86.4, SD 97.7 months), and number of sick-leave days (mean 1.7, SD 3.6 days vs mean 2.1, SD 4.5 days) after randomization.
Table 1. Baseline demographic and clinical characteristics of the trial groups (N=220).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>App-based intervention (n=110)</th>
<th>Control (n=110)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>37.9 (11)</td>
<td>39.8 (11.6)</td>
</tr>
<tr>
<td><strong>Gender, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>74 (67.3)</td>
<td>79 (71.8)</td>
</tr>
<tr>
<td>Male</td>
<td>36 (32.7)</td>
<td>31 (28.2)</td>
</tr>
<tr>
<td>BMI (kg/m²), mean (SD)</td>
<td>24.5 (4.6)</td>
<td>23.9 (4.1)</td>
</tr>
<tr>
<td>Graduation after ≥10 years of school, n (%)</td>
<td>79 (71.8)</td>
<td>75 (68.2)</td>
</tr>
<tr>
<td><strong>Size of household, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single-person</td>
<td>32 (29.1)</td>
<td>34 (30.9)</td>
</tr>
<tr>
<td>2-person</td>
<td>44 (40)</td>
<td>42 (38.2)</td>
</tr>
<tr>
<td>Multiperson</td>
<td>34 (30.9)</td>
<td>34 (30.9)</td>
</tr>
<tr>
<td>Partnership, n (%)</td>
<td>56 (50.9)</td>
<td>66 (60)</td>
</tr>
<tr>
<td>Migration background&lt;sup&gt;a&lt;/sup&gt;, n (%)</td>
<td>14 (12.7)</td>
<td>21 (19.1)</td>
</tr>
<tr>
<td>Neck pain intensity in the previous 7 days (NRS&lt;sup&gt;b,c&lt;/sup&gt;), mean (SD)</td>
<td>5.7 (1.4)</td>
<td>5.8 (1.3)</td>
</tr>
<tr>
<td>Neck pain–related stress intensity in the previous 7 days (NRS&lt;sup&gt;c&lt;/sup&gt;), mean (SD)</td>
<td>5.4 (1.9)</td>
<td>5.3 (2.1)</td>
</tr>
<tr>
<td>Duration of neck pain (months), mean (SD)</td>
<td>79.2 (74.8)</td>
<td>86.4 (97.7)</td>
</tr>
<tr>
<td>Sick-leave days, mean (SD)</td>
<td>1.7 (3.6)</td>
<td>2.1 (4.5)</td>
</tr>
<tr>
<td>Medication intake against neck pain, n (%)</td>
<td>28 (25.5)</td>
<td>31 (28.2)</td>
</tr>
<tr>
<td>Pain acceptance, mean (SD)</td>
<td>73.3 (16.7)</td>
<td>73.6 (15.9)</td>
</tr>
<tr>
<td>Subscale pain willingness, mean (SD)</td>
<td>30.1 (10.1)</td>
<td>31.1 (8.2)</td>
</tr>
<tr>
<td>Subscale activity engagement, mean (SD)</td>
<td>43.2 (8.8)</td>
<td>42.4 (9)</td>
</tr>
<tr>
<td><strong>Expected effectiveness of relaxation exercise, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recovery</td>
<td>1 (0.9)</td>
<td>5 (4.5)</td>
</tr>
<tr>
<td>Distinct improvement</td>
<td>54 (49.1)</td>
<td>61 (55.5)</td>
</tr>
<tr>
<td>Light improvement</td>
<td>55 (50)</td>
<td>44 (40)</td>
</tr>
<tr>
<td>No improvement</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Ineffective</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>Expected effectiveness of no relaxation exercise, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recovery</td>
<td>0 (0)</td>
<td>1 (0.9)</td>
</tr>
<tr>
<td>Distinct improvement</td>
<td>3 (2.7)</td>
<td>6 (5.5)</td>
</tr>
<tr>
<td>Light improvement</td>
<td>15 (13.6)</td>
<td>18 (16.4)</td>
</tr>
<tr>
<td>No improvement</td>
<td>89 (80.9)</td>
<td>81 (73.6)</td>
</tr>
<tr>
<td>Ineffective</td>
<td>3 (2.7)</td>
<td>4 (3.6)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Based on a study by Schenk et al [31].

<sup>b</sup>NRS: Numeric Rating Scale.

<sup>c</sup>Lower values indicate better status.

**Outcomes**

Less intense mean neck pain was observed in both groups during the first 3 months compared with the baseline (Table 2). However, there was no significant difference in the primary outcome of the mean neck pain intensity during the first 3 months between the intervention and control groups (group difference 0.3, 95% CI –0.2 to 0.7; \( P=.23 \)). In addition, no significant differences in the mean neck pain intensity between the 2 groups during the second 3 months (group difference –0.1, 95% CI –0.7 to 0.4; \( P=.62 \)) or during the entire 6 months (group difference 0.1, 95% CI –0.3 to 0.6; \( P=.62 \)) were found.

The subgroup analysis also yielded comparable primary outcomes between participants of different genders, ages, education levels, and disease durations.
Table 2. Primary and secondary outcomes (adjusted for sex and baseline value; N=220).

<table>
<thead>
<tr>
<th>Outcome</th>
<th>App-based intervention, mean (95% CI)</th>
<th>Control, mean (95% CI)</th>
<th>Differences intervention versus control, mean (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neck pain intensity during first 3 months (NRS&lt;sup&gt;a,b&lt;/sup&gt;)</td>
<td>4.1 (3.8 to 4.4)</td>
<td>3.8 (3.5 to 4.1)</td>
<td>0.3 (–0.2 to 0.7)</td>
<td>.23</td>
</tr>
<tr>
<td>Neck pain intensity (NRS&lt;sup&gt;b&lt;/sup&gt;)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second 3 months</td>
<td>3.6 (3.2 to 4)</td>
<td>3.7 (3.4 to 4.1)</td>
<td>–0.1 (–0.7 to 0.4)</td>
<td>.62</td>
</tr>
<tr>
<td>First 6 months</td>
<td>3.9 (3.6 to 4.2)</td>
<td>3.8 (3.5 to 4.1)</td>
<td>0.1 (–0.3 to 0.6)</td>
<td>.62</td>
</tr>
<tr>
<td>Average neck pain during previous 7 days (NRS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First 3 months</td>
<td>4.3 (4 to 4.6)</td>
<td>4 (3.8 to 4.3)</td>
<td>0.2 (–0.2 to 0.7)</td>
<td>.24</td>
</tr>
<tr>
<td>Second 3 months</td>
<td>3.8 (3.4 to 4.1)</td>
<td>3.9 (3.6 to 4.3)</td>
<td>–0.2 (–0.7 to 0.3)</td>
<td>.52</td>
</tr>
<tr>
<td>First 6 months</td>
<td>4.1 (3.8 to 4.4)</td>
<td>4 (3.7 to 4.3)</td>
<td>0.2 (–0.3 to 0.6)</td>
<td>.49</td>
</tr>
<tr>
<td>Pain acceptance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>After 3rd month</td>
<td>75.4 (73 to 77.8)</td>
<td>75.8 (73.4 to 78.1)</td>
<td>–0.4 (–3.8 to 3)</td>
<td>.83</td>
</tr>
<tr>
<td>After 6th month</td>
<td>76.1 (73.7 to 78.4)</td>
<td>75.8 (73.6 to 78.1)</td>
<td>0.2 (–3 to 3.5)</td>
<td>.89</td>
</tr>
<tr>
<td>Participants with medication intake against neck pain, proportion (%)&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>During 6 months</td>
<td>49.5 (39.8 to 59.3)</td>
<td>52.4 (42.4 to 62.2)</td>
<td>0.97 (0.5 to 1.8)</td>
<td>.69</td>
</tr>
<tr>
<td>Numbers of weeks with pain medication</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First 3 months</td>
<td>2 (1.5 to 2.5)</td>
<td>2 (1.4 to 2.5)</td>
<td>0.01 (–0.7 to 0.8)</td>
<td>.98</td>
</tr>
<tr>
<td>Second 3 months</td>
<td>2 (1.4 to 2.6)</td>
<td>2 (1.5 to 2.6)</td>
<td>–0.03 (–0.8 to 0.8)</td>
<td>.93</td>
</tr>
<tr>
<td>First 6 months</td>
<td>3.7 (2.7 to 4.7)</td>
<td>3.9 (2.9 to 4.9)</td>
<td>–0.2 (–1.7 to 1.2)</td>
<td>.75</td>
</tr>
<tr>
<td>Neck pain–related stress</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First 3 months</td>
<td>4 (3.7 to 4.3)</td>
<td>3.8 (3.5 to 4.1)</td>
<td>0.2 (–0.2 to 0.7)</td>
<td>.32</td>
</tr>
<tr>
<td>Second 3 months</td>
<td>3.6 (3.2 to 3.9)</td>
<td>3.6 (3.2 to 4)</td>
<td>0 (–0.6 to 0.5)</td>
<td>.88</td>
</tr>
<tr>
<td>First 6 months</td>
<td>3.9 (3.6 to 4.2)</td>
<td>3.7 (3.4 to 4)</td>
<td>0.2 (–0.3 to 0.6)</td>
<td>.46</td>
</tr>
<tr>
<td>Responder rate, proportion (%)&lt;sup&gt;c,d,e&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>After third month</td>
<td>29.4 (21 to 38.9)</td>
<td>35.6 (26.4 to 45.6)</td>
<td>0.75 (0.4 to 1.4)</td>
<td>.33</td>
</tr>
<tr>
<td>After sixth month</td>
<td>35.9 (26.8 to 45.7)</td>
<td>37.5 (28.2 to 47.5)</td>
<td>0.93 (0.5 to 1.7)</td>
<td>.80</td>
</tr>
<tr>
<td>Sick-leave days</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>After third month</td>
<td>1.2 (0.4 to 2)</td>
<td>1.5 (0.7 to 2.3)</td>
<td>–0.3 (–1.4 to 0.9)</td>
<td>.66</td>
</tr>
<tr>
<td>After sixth month</td>
<td>1.1 (0.6 to 1.6)</td>
<td>1 (0.5 to 1.5)</td>
<td>0.1 (–0.6 to 0.8)</td>
<td>.81</td>
</tr>
<tr>
<td>Concomitant treatment, proportion (%)&lt;sup&gt;c,d&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>After third month</td>
<td>40 (30.8 to 49.8)</td>
<td>45.5 (35.9 to 55.2)</td>
<td>0.82 (0.5 to 1.4)</td>
<td>.50</td>
</tr>
<tr>
<td>After sixth month</td>
<td>47.3 (37.7 to 57)</td>
<td>43.6 (34.2 to 53.4)</td>
<td>1.20 (0.7 to 2.1)</td>
<td>.69</td>
</tr>
</tbody>
</table>

<sup>a</sup>NRS: Numeric Rating Scale.

<sup>b</sup>Lower values indicate better status.

<sup>c</sup>Between-group differences are presented as odds ratio (95% CI) instead of mean (95% CI).

<sup>d</sup>Proportions are not adjusted.

<sup>e</sup>Either at least 50% pain reduction or at least 2.5 points on the Numeric Rating Scale compared with baseline.

Furthermore, there were no significant differences between the mean average neck pain based on weekly measurements in either group during the first 3 months (group difference 0.2, 95% CI –0.2 to 0.7; P=.24), second 3 months (group difference –0.2, 95% CI –0.7 to 0.3; P=.52), or the entire 6 months (group difference 0.2, 95% CI –0.3 to 0.6; P=.49). The chance of being a responder was similar for both groups after 3 months (odds ratio 0.75, 95% CI 0.4-1.4) and after 6 months (odds ratio 0.93, 95% CI 0.5-1.7).

There were also no significant differences in pain acceptance between the groups after 3 months (group difference –0.4, 95% CI –3.8 to 3; P=.83) and 6 months (group difference 0.2, 95% CI –3 to 3.5; P=.89).
There was no significant difference between the proportions of participants who took pain medication among both groups during the whole follow-up period of 6 months (odds ratio 0.97, 95% CI 0.5-1.8; \(P=.68\)). The number of weeks with pain medication did not differ between the groups in the first 3 months, second 3 months, and 6 months. The number of sick-leave days and pain acceptance did not differ between the groups.

The sensitivity and subgroup analyses did not change the pattern of the results, and we found no significant difference between female and male participants in a subgroup analysis of the primary outcome.

**App-Based Exercise Time and Study Dropout**

The overall time spent exercising declined with time. In the first week, almost all participants (109/110, 99.1%) in the intervention group practiced the exercises with the app. However, only 40% (44/110) of the participants continued to practice the exercises (for any length) in week 12, and 30% (33/110) of the participants continued to practice the exercises (for any length) in week 26. The declining trend was similar over the study phase when comparing the number of participants who practiced relaxation exercises of any length with the number of participants who practiced relaxation exercises for at least 10 minutes per week (Figure 3).

The Kaplan–Meier survival curves in Figure 4 display the study dropouts. There was no significant difference in the curves between the 2 groups according to the log-rank test (\(P=.44\)). Approximately 74.5% (82/110) of participants in the intervention group and 79.1% (87/110) of participants in the control group used the study app to answer the survey questions until the end of the study (week 26).

**Figure 3.** Number of participants practicing the exercises over time.

**Figure 4.** Probability of dropout in using the study app by group.
Self-perceived Neck Pain Change

Overall, 60% (66/110) of participants in the intervention group reported that they felt the neck pain improved significantly or slightly after 3 and 6 months, in contrast to approximately 30% (33/110) of participants in the control group who said the same (Figure 5).

Figure 5. Self-perceived improvement of neck pain.

BCTs in the App

Most parts of the app’s user interface implementations can be characterized as prompt and cues BCT, such as the dashboard dialog showing the number of questionnaires remaining to be processed. Moreover, the prompt and cues BCT was combined with the action planning BCT to remind participants to fill out their weekly diaries. Participants could determine the time and date of the reminders (action planning and prompt and cues BCT).

To ensure proper performance of the relaxation exercises, all the exercises were explained by experienced clinicians in an audio recording (instruction on how to perform the behavior BCT). The Relaxneck app provided the full name, profession, professional title, and workplace of the audio recording instructors to ensure quality and safety for the participants (credible source BCT).

Safety Data

There were 5 serious adverse events recorded only in the control group, including cancer, sudden hearing loss, nerve injury and spinal tap, tonsillectomy, and an accident causing a fracture of the upper arm. None of them was considered related to the trial or the trial intervention.

Discussion

Principal Findings

In our trial, additional app-based self-relaxation techniques were not more effective than usual care alone for the reduction of chronic neck pain in a pragmatic setting. The results were consistent across all outcomes. The evaluated self-relaxation techniques were safe to use; however, they did not effectively relieve chronic neck pain during this app-based study.

There are a few possible reasons that helped to understand why the intervention did not improve pain. The study app’s design was not updated during the study (developed in 2014) and did not include more elaborate BCTs, such as feedback about the correct application of the intervention and monitoring [27]. As the retrospective BCT analysis showed, only prompt and cues BCT was mainly used, whereas modern digital interventions or consumer apps widely apply BCTs [32,33]. In mobile health settings, personalized feedback from the app would be a promising virtual communication tool to enhance patient engagement and adherence [34]. Biofeedback and self-monitoring of changes are very important in relaxation- and mindfulness-based therapies for pain. Moreover, it must be considered that our study mainly measured self-reported outcomes. The study may have benefited from parameters such as step count as a measure of physical activity or sleep duration as a proxy for sleep [35]. At the time when the study was planned, wearables were not widely implemented, and it was more difficult to link these measures with an app because of interoperability issues. However, the type and duration of the audio recordings used as interventions were measured and used as measures of adherence. Although tracked outcomes may have added a more objective point of view, the implementation would have added a much larger complexity during the development of the app. In addition, mindfulness-based therapies are very often designed with progressive lengths or difficulties [36]. In our trial, the participants were required to practice 3 relaxation exercises of almost the same length repeatedly across the whole intervention period. This could have limited the participants’ interest and the treatment effect. Finally, our app focused on audio relaxation alone instead of incorporating a whole theoretical framework such as mindfulness-based stress reduction or a comprehensive pain management strategy. Therefore, it is likely that the intervention of the study app was not powerful enough to improve chronic pain.

Adherence to the trial intervention was low compared with other app-based studies conducted by our research group [30,37]. The number of participants who performed the relaxation exercises slightly after 3 and 6 months, in contrast to approximately 30% (33/110) of participants in the control group who said the same (Figure 5).
diminished during the course of the study. Potential explanations may again be the lack of an elaborate BCT concept or that chronic pain decreases motivation [38], especially to perform prescribed physical activities and exercises [39]. However, the number of practiced exercises of any length or >10 minutes remained similar over time. This might indicate that users who feel attached to the app-based relaxation exercise at the beginning finish the whole exercise process in most cases.

Although our study intervention was asynchronous, that is, contact with a health care provider and app intervention occurred at different time points, future mobile health studies may also include synchronous interventions in which health care providers could offer real-time interventions to the users. This approach might be helpful to improve the app and study adherence. However, this approach might also increase the complexity of the intervention and increase the costs.

In our trial, stopping the app-based intervention did not necessarily predict stopping the answering of the app-based survey questions. Only 30% (33/110) of the participants continued to practice the app-based relaxation exercises until the end of the follow-up; however, 74.5% (82/110) of participants used the app to answer survey questions until the end of the trial. Meanwhile, adherence to app use for answering survey questions was not affected by whether the app contained intervention features. The proportion of participants who used the app regularly to answer surveys until the end of the study was rather similar in both groups. A possible explanation for the good response rate in both groups could be our reminder system for the questionnaires or the paid compensation for the efforts.

Although all other outcomes did not show statistically significant group differences, most participants in the intervention group reported self-perceived improvement of neck pain, whereas most participants in the control group reported no change or worsening of neck pain. This result might be attributed to a digital placebo effect. The concept of the digital placebo effect has already been discussed in mental health studies [40]. A good example could be seen in a study involving a smartphone app that was designed to help patients self-monitor and record their symptoms of depression. Even without any direct therapeutic intervention, smartphone-based self-monitoring significantly reduced the symptoms [41]. Future studies should investigate the perceived changes in pain and the placebo-like effects of smartphone interventions.

Strengths and Limitations

Our app-based RCT was performed in a pragmatic setting. In addition, stakeholder engagement was implemented in the design of the trial and intervention [18]. Hence, the selection of the relaxation exercises and the length of the exercises were defined during stakeholder meetings to facilitate patient-centered therapy. Moreover, the study included a sufficient number of participants to answer our research question. Thus, our findings were considered generalizable in a real-life setting.

Some limitations have to be considered for this trial. The trial recruitment took rather long (32 months), possibly because of our conventional on-site recruitment strategy with paper-and-pencil baseline questionnaires. During that time, smartphone technologies, designs, and perceptions experienced numerous changes. For example, it is unclear whether the app’s user interface was perceived as outdated by the participants. For future app-based studies, web-based recruitment and the incorporation of an app-based baseline survey could accelerate the overall trial process [15]. This acceleration of the trial process might also increase the relevance of the results.

Potential selection bias with an impact on the generalizability of the results might be another limitation of this study. The trial was conducted from 2014 to 2017. All study participants needed to own a smartphone. However, at that time, the number of smartphone owners in Germany (approximately 50%) was substantially lower than the current number (approximately 72%) [42]. It is unclear whether this affected the characteristics of our study population. To address a broader user base, we decided to build the study app for both the main platforms (iOS and Android).

Unfortunately, our sample size could not enable gender disaggregation. Gender might influence behavioral change, use patterns, and adherence to app use [43]. Some app-based studies have reported that gender is a strong predictor of the discontinuation of relaxation app use [37,44]. In this study, approximately 69.5% (153/220) of the participants were women. It would be interesting to discover the role of sex and gender in participants’ adherence in future studies.

During the development of the app, we did not follow a preplanned BCT concept, and only basic BCTs were implemented, as shown in the post hoc review of the BCT techniques used. However, regarding behavioral change and intervention effects, a meta-analysis [45] concluded that implementing more (than one) theory is unlikely to improve intervention effectiveness. Future studies should be conducted to better understand the impact of BCTs on intervention outcomes for interventions for chronic pain.

Finally, the trial was single-blinded, as we could not blind the participants. However, it is common that participants cannot be blinded in nonpharmacological complex intervention trials and eHealth trials.

Comparison With Previous Work

Mind–body therapies are considered to be relatively safe [46]. However, only a few studies have been conducted on chronic neck pain. There were not enough trials for the Institute for Clinical and Economic Review (ICER) to summarize the effectiveness of cognitive and mind–body therapies for chronic neck pain [14]. According to a systematic review that investigated the effects of mindfulness- and relaxation-based interventions in an eHealth setting [47], only a few studies reported positive effects on pain, and no study reported positive effects on stress or mindfulness.

However, some eHealth studies have been conducted for chronic lower back pain. Heapy et al [48] reported that the efficacy of cognitive behavioral therapies (CBTs) delivered remotely using telephone and the internet for chronic back pain is not inferior to that of in-person CBTs. Kristjánsdóttir et al [49] reported that smartphone app–based interventions with personalized
feedback can reduce catastrophizing in women with chronic widespread pain. Instead of relaxation exercises alone, CBT, including emotion recognition, mindfulness exercises, and empathic communication, was highlighted in these studies. It seems that the evidence for only relaxation is rather low compared with systematic mind–body therapy or CBT for chronic pain. Therefore, future studies are required to investigate the effect of mind–body therapy on chronic neck pain within a comprehensive pain management strategy.

Conclusions
In conclusion, the evaluated study smartphone app, which included self-relaxation techniques such as autogenic training, mindfulness meditation, and guided imagery but without elaborate BCTs, was not more effective than usual care for chronic neck pain in a pragmatic trial. Further studies are needed to understand the potential of relaxation for neck pain and whether app-based mechanisms for relaxation and behavior change might be useful within a comprehensive pain management strategy for neck pain.

Acknowledgments
The authors thank Beatrice Eden and Iris Bartsch for app testing, as well as all the participants of this study. The authors thank their stakeholder team [18].

Authors’ Contributions
DP, SB, and CMW conceived and designed the study. TK performed the data analysis, and JW and the other coauthors performed the data interpretation. JW and DP wrote the first draft of the paper. All the authors discussed the results, commented on the paper, and approved the final paper.

Conflicts of Interest
This was an investigator-initiated trial. The app was developed for research purposes and is not a commercial product. The authors do not have any financial stake in the success of the app. CMW received research grants from the university for digital health projects from Krebsliga Schweiz, German Cancer Aid, The German health care Innovation Fund, and Newsenselab GmbH. Board positions related to digital health for mind and body (nonpaid) are as follows: Codirector of the Digitals Society Initiative of the University Zurich and President Schweizer Fachverband Mind Body Medicine

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.2).

[PDF File (Adobe PDF File), 91 KB - mhealth_v10i1e31482_app1.pdf]

References


13. DEGAM S1-Handlungsempfehlung Nackenschmerzen. DEGAM. 2016. URL: https://tinyurl.com/2p8432uw [accessed 2021-12-08]


https://mhealth.jmir.org/2022/1/e31482

JMIR Mhealth Uhealth 2022 | vol. 10 | iss. 1 | e31482 | p.160

(page number not for citation purposes)


Abbreviations

BCT: behavior change technique
CBT: cognitive behavioral therapy
ICER: Institute for Clinical and Economic Review
NRS: Numeric Rating Scale
RCT: randomized controlled trial
Enabling Research and Clinical Use of Patient-Generated Health Data (the mindLAMP Platform): Digital Phenotyping Study

Aditya Vaidyam1, MS; John Halamka2, MD; John Torous1, MBI, MD

1Beth Israel Deaconess Medical Center, Boston, MA, United States
2Mayo Clinic, Rochester, MN, United States

Corresponding Author:
John Torous, MBI, MD
Beth Israel Deaconess Medical Center
330 Brokline Avenue
Boston, MA, 02215
United States
Phone: 1 6176676700
Email: jtorous@bidmc.harvard.edu

Abstract

Background: There is a growing need for the integration of patient-generated health data (PGHD) into research and clinical care to enable personalized, preventive, and interactive care, but technical and organizational challenges, such as the lack of standards and easy-to-use tools, preclude the effective use of PGHD generated from consumer devices, such as smartphones and wearables.

Objective: This study outlines how we used mobile apps and semantic web standards such as HTTP 2.0, Representational State Transfer, JSON (JavaScript Object Notation), JSON Schema, Transport Layer Security (version 1.3), Advanced Encryption Standard-256, OpenAPI, HTML5, and Vega, in conjunction with patient and provider feedback to completely update a previous version of mindLAMP.

Methods: The Learn, Assess, Manage, and Prevent (LAMP) platform addresses the abovementioned challenges in enhancing clinical insight by supporting research, data analysis, and implementation efforts around PGHD as an open-source solution with freely accessible and shared code.

Results: With a simplified programming interface and novel data representation that captures additional metadata, the LAMP platform enables interoperability with existing Fast Healthcare Interoperability Resources–based health care systems as well as consumer wearables and services such as Apple HealthKit and Google Fit. The companion Cortex data analysis and machine learning toolkit offer robust support for artificial intelligence, behavioral feature extraction, interactive visualizations, and high-performance data processing through parallelization and vectorization techniques.

Conclusions: The LAMP platform incorporates feedback from patients and clinicians alongside a standards-based approach to address these needs and functions across a wide range of use cases through its customizable and flexible components. These range from simple survey-based research to international consortiums capturing multimodal data to simple delivery of mindfulness exercises through personalized, just-in-time adaptive interventions.

(JMIR Mhealth Uhealth 2022;10(1):e30557) doi:10.2196/30557

KEYWORDS
digital phenotyping; mHealth; apps; FHIR; digital health; health data; patient-generated health data; mobile health; smartphones; wearables; mobile apps; mental health, mobile phone

Introduction

Background
The medical field today is transitioning toward integrating patient-generated health data (PGHD) into clinical care to increase shared decision-making, coordination of care, patient safety, and clinical outcomes [1]. PGHD are central to this mission and are defined as data recorded or created by the patient or caregivers used to address health concerns. Examples include a longitudinal view of symptoms of patient status between clinic visits captured via an app or information related to daily adherence to treatment plans [1]. This may include a...
daily step count captured from a smartphone, medication surveys administered on the smartphone, and sleep quality data measured via a wearable sensor. Given the ability of smartphones and wearables to collect a myriad of continuous multimodal data relevant to care, such as heart rate, sleep, steps, and more, tools and systems to harness and use this vast amount of automatically generated PGHD are a health care priority. One challenge that remains toward this goal is the lack of technical infrastructure and organizational capability to handle the intake of accurate and valid PGHD from patient-owned consumer devices. Such standards and tools supporting the effective and compatible integration of PGHD are needed to enhance clinical insight and support research and data analysis before PGHD can actually impact routine care.

Digital Phenotyping+

The need for apps that can not only capture but also use and integrate PGHD is clear. The use of commercially available wearable technology for the acquisition of PGHD has seen a recent uptick owing to the effects of the COVID-19 pandemic and growing demand for telehealth services [2,3]. Today, over 80% of Americans own a smartphone device [4] and over 20% own a wearable device [5], reinforcing the potential of PGHD to improve clinical outcomes. For example, the Apple Watch today retails at US $350 with onboard nonmedical grade electrocardiography and oxygen saturation sensors that continuously measure and record data. Modern smartphones are also equipped with numerous sensors that generate a high volume of potentially clinically significant PGHD that could enable a better real-time understanding of cognition, mobility, sociability, and more through techniques, such as digital phenotyping and ecologic momentary assessment.

Digital phenotyping [6] is the construction of an individual-level phenotype using data collected from smartphones or wearable devices actively via user interaction (eg, surveys), or passively without user input (eg, sensors, such as an accelerometer). Although there are many digital phenotyping tools and systems used in health care and research contexts, a recent review identified nearly 50 of them [7], and few offer an integrated and standardized approach to analyze and respond to clinically actionable patient-generated data. Existing tools are primarily closed systems or consist of only a single app with little flexibility or customizability [7]. In a recent review, 85% of existing solutions supported active and passive sensing but only 33% supported clinical assessment, 30% supported predictive modeling of patient data, and 24% supported app-delivered interventions [7]. Furthermore, only 35% of the existing solutions showed a patient- or clinician-facing user interface [7]. However, the combination of all these features is necessary to meet the diverse needs of research and care delivery. A search outside the research literature and instead, across mental health smartphone apps in commercial marketplaces reported that only 1.1% supported sensors [8], suggesting that many research tools do not translate into accessible tools for patient or clinician use. Although diverse functionality and innovation continue to exist across the entire app space, we have argued that there is a need for multiple uses of the same app, instead of using multiple apps in a fragmented manner, toward better supporting clinical research, integration, and implementation [9].

Challenges in Integration

Creating PGHD tools that use sensor and digital phenotyping tools in a more patient- and health system-centric manner is a common goal, but it remains challenging to achieve. Despite the prevalence of existing electronic medical record standards and tools, a 2019 review on the integration of PGHD into clinical practice, “integration [...] was extremely limited, and decision support capabilities were for the most part basic” [10]. The most widely adopted medical record standards initiative that can be used to link PGHD to medical records is Fast Healthcare Interoperability Resources (FHIR), led by the Health Level 7 organization [11], which is now adopted by many major health care systems and industry partners, including Apple, Google, Amazon, Microsoft, and others [12]. Its companion projects SMART (Specific, Measurable, Achievable, Realistic, and Timely) [13] and SMART Markers [14] build on FHIR and enable integration of third-party modules into medical record systems, including patient mobile devices and sensors.

However, the FHIR ecosystem alone does not address a number of concerns specific to the integration of PGHD into clinical systems. FHIR and the current data interoperability standard (United States Core Data for Interoperability) [15] were not developed for continuous high-velocity data, and its implementation in the health care ecosystem today is primarily read-only, although its data gathering and write-back ability continues to evolve. For example, the FHIR core does not allow for semantic equivalence of data that can be used to automate data matching or harmonization. As a result, it is not possible to work with both cognitive assessment scores and mobility or sociability metrics using the same analysis pipeline. This increases the effort required and time taken to work with PGHD, as clinicians or researchers must first preprocess data of different semantic types individually before being able to work with a data set as a whole. Although R4 extensions, such as the mCODE core cancer model [16] are becoming a new way to expand FHIR’s supported vocabulary, they are still early in evolution and adoption. Today, the inability to standardize terminologies across interconnected systems, such as through a data dictionary, impedes effective export and analysis of different types of data from different data sources using the FHIR ecosystem [17,18]. These challenges preclude the adoption of FHIR as a PGHD-first standard for clinical and research use cases.

Thus, there remains a need for a flexible, interoperable, and extensible platform that enables the effective use of PGHD through widely accepted standards for both clinical and research needs. In this paper, we present a potential solution for the robust and effective acquisition and integration of PGHD into research and clinical care with tangible examples and open-source code.

Methods

Overview

To integrate PGHD into research and clinical care needs, our team has designed and developed the Learn, Assess, Manage, and Prevent (LAMP) platform that encompasses a robust set of protocols, standards, tools, and apps. Our team initially developed the mindLAMP smartphone app [19] as part of the
initial version of the LAMP platform. In this paper, we review a rearchitected and redeveloped platform comprising new frontend, backend, and analysis components to support PGHD, patient-centric care, and actionable digital phenotyping. This new platform, distinct from its predecessor, includes features such as customizable and schedulable activities, sensor data collection and analysis, messaging support with the care team, and more, available across modern web browsers and smartphone operating systems. The design and development of the platform was approached from both a patient- and clinician-focused approach as well as a semantic standards–based approach.

**Patient- and Clinician-Led Design**

The LAMP platform was designed and developed with continuous feedback from patients with serious mental illnesses and clinicians. Through a patient advisory panel, focus groups [20-22], clinical use, and feedback from a global consortium of users, mindLAMP has been co-designed iteratively with updates reflecting expanding ideas for its role. User input informed the adaptability, flexibility, and customizability of the LAMP platform, which resulted in a new user interface compared with the previous version, as shown in Figure 1. We established a formal system to enable anyone to suggest improvements, report bugs, and assess new features to ensure that all could partake in the iterative design rounds. This process also influenced aspects of the user experience, such as making mindLAMP available in multiple languages (English, Spanish, and Hindi) and designing to ensure easy addition of more. Key examples of patient feedback and the design outcomes they influenced are provided in Table 1.

**Figure 1.** (A) The home screen interface of the original (version 1 with red border) mindLAMP app; (B) the new and improved home screen interface of the (version 2 with green border) mindLAMP app that incorporates multiple activities and schedules into a heads-up tab called the Feed; (C) the Learn, Assess, Manage, and Prevent tabs group embedded activities together with helpful tooltips and icons, with new activity types (eg, tips, meditation); (D) returning data and insight to users was considered a priority in the redesigned user interface and thus more advanced charting tools were integrated into the Prevent tab, accessible to both patients and clinicians.

**Table 1.** Selected examples of patient feedback driving significant changes in the user experience and overall architecture of the Learn, Assess, Manage, and Prevent platform. Semantic (technical) standards–based approach.

<table>
<thead>
<tr>
<th>Sample patient feedback</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Apps like Facebook or Amazon are clear where I am lost in a sea of people or items and that is generally accepted, but with apps in the health space, who is involved, which institution is involved, level of comfort with the individuals and what data is collected, all of these factors are carefully calculated when I make a decision to join a study like this—an establishment of trust is crucial. Apps that track people incur a level of suspicion that changes between people, from none at all to a lot, perhaps depending on level of illness.”</td>
<td>The Learn, Assess, Manage, and Prevent platform was rebuilt around an open-source collaborative environment supported by the consortium; all development and data handling processes are disclosed in the privacy policy, and significant backend changes were made to patient data equity and ownership. Additional types of activities were added to the mindLAMP app, including tips (Figure 2), meditation, and other informational and management tools; each of these activities captures metadata during patient use that can be interpreted and incorporated into a clinical encounter.</td>
</tr>
<tr>
<td>“Yeah, because I don’t see any apps out there these days that help people with psychosis and when they’re getting sicker. It just seems...they just don’t help with certain things. This gives you control to go get help if somebody needs it. It’s like, the good thing about this app is, it’s getting the right information and it’s sending you somewhere, it’s almost as if you could go to the therapist with this information! You don’t want an app that’s just one sided [and siloed off from the therapist or delivery of care].”</td>
<td>The smaller heads-up summaries originally found in the first version of the mindLAMP app (Figure 2) were updated and expanded into an entire tab (Figure 2), providing more insight and customization into patient data.</td>
</tr>
<tr>
<td>“mindLAMP is a tool for me to get better: I want to know if I’m making progress and when, what am I deficient in, how am I deficient, and how to improve on it; that is, as a metrics-driven person.”</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2. Flow of the data collection process from native app to backend: (1) an activity specification describes the types of interactive elements available in the mindLAMP app, along with their possible configuration parameters; (2) when participants interact with a configured and scheduled activity (such as a mood survey based on the survey specification), all metadata and data from the interaction session is integrated into a single unit of patient-generated health data called events; (3) events are then submitted to the backend in real time as part of a continuously generated stream of patient-generated health data; and (4) clinicians and researchers are able to perform continuously updating queries on the data with their desired parameters.

In addition to a patient- and clinician-centric design approach, the LAMP platform was also architected with a semantic standards–based approach, considering technical best practices for future proofing and security or compliance across health care systems. The open standards listed in Table 2 were chosen specifically to foster an open ecosystem around the platform. For example, the platform’s programming interface adopts a repository model to store and configure patient-facing instruments, each with its own embedded user interface. These embedded user interfaces were developed using common and widely adopted web standards indicated in Table 2 (HTML5.0, Cascading Style Sheets 3.0, and ECMAScript 6). As a patient begins an interaction session, this embedded code is securely sandboxed by the mindLAMP user interface both within the smartphone app and the patient-facing web dashboard. In addition to providing a standard schematic of all structured documents encountered and processed in the LAMP platform using OpenAPI [23], the JSON Schema [24] data markup standard is used to provide developers of these interactive patient-facing instrument configurability and extensibility. With little required skill or upfront effort, developers can use the platform’s software development kit to create instruments with completely customizable user experiences that are then tuned and customized by clinicians for individual patients or by research coordinators for studies spanning many patients.
Table 2. Adopted semantic web standards, their use rationale, and implementation details.

<table>
<thead>
<tr>
<th>Standard</th>
<th>Description</th>
<th>Use</th>
<th>Reason chosen</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTTP 2 [22]</td>
<td>Ubiquitous web standard that declares and defines the semantics of client-server communication with a rich and readily available debugging and implementation toolset and ecosystem not available for custom binary protocols</td>
<td>Implemented by core programming libraries and the backend</td>
<td>In contrast to TCP-based binary data protocols requiring specialized tooling to access and work with data, almost all systems and tools are able to interact with web standards through the HTTP protocol.</td>
</tr>
<tr>
<td>REST [25]</td>
<td>Ubiquitous lightweight HTTP-based web standard that defines systems logically through accessibility and manipulation of remote resources instead of invocation of remote functions</td>
<td>Implemented by core programming libraries and the backend</td>
<td>In contrast to a custom implementation of remote function invocation that would require custom programming libraries to interface with, most web systems are able to interact with REST-based resources in a logical manner. In the absence of developer knowledge or pre-existing tools, it remains possible to communicate with RESTful systems.</td>
</tr>
<tr>
<td>JSON [26]</td>
<td>Ubiquitous web standard that supports structured (as opposed to tabular, ie, CSV files) formatting and markup of data using strict data types</td>
<td>Implemented across all components in the platform</td>
<td>In contrast to encoded binary data formats requiring specialized tooling to interpret and work with data, most programming environments support the JSON standard.</td>
</tr>
<tr>
<td>TLS version 1.3 [27]</td>
<td>Ubiquitous web standard that enables encryption of data in transit between client and server</td>
<td>Implemented by core modules and programming libraries, used by all components in the platform</td>
<td>No alternative</td>
</tr>
<tr>
<td>AES-256 [28]</td>
<td>Ubiquitous cryptographic standard that enables encryption of data at REST (on disk) by a database</td>
<td>Implemented by the backend and mandated by the backend and deployment configuration for the database within which data shall be stored</td>
<td>No alternative</td>
</tr>
<tr>
<td>JSON Schema [24]</td>
<td>Web standard that describes JSON-encoded data and metadata through ahead-of-time specification of a universally agreed upon schema, as opposed to inline schema provided only at runtime</td>
<td>Implemented by the backend and used by the frontend</td>
<td>Although binary protocols require a predetermined strict schema to format the data, JSON does not. JSON Schema provides ahead-of-time resolution of the contents of a data payload and can be used to validate and harmonize data as well.</td>
</tr>
<tr>
<td>OpenAPI [23]</td>
<td>Web standard that describes REST-based web services and metadata through ahead-of-time specification of a universally agreed upon schema</td>
<td>Implemented by the backend and core programming libraries and used by the frontend</td>
<td>In contrast to writing programming libraries and testing or validation tools, the generation of these tools and packages by the OpenAPI ecosystem increases productivity.</td>
</tr>
<tr>
<td>HTML5 [29]</td>
<td>Ubiquitous web standard that makes it possible to securely embed custom user interfaces backing patient-facing activities</td>
<td>Implemented by the frontend and all patient-facing activities, with wide support for CSS [3] and JavaScript 2016 (ES6)</td>
<td>No alternative</td>
</tr>
<tr>
<td>Vega [30]</td>
<td>Visualization grammar standard that encodes charts and graphs as JSON documents that are then rendered and viewed interactively by apps</td>
<td>Implemented using HTML5. Implemented by the frontend and Cortex analysis code</td>
<td>In contrast to static images and handwritten analysis code, the ability to declaratively generate interactive real-time charts through an embedded query reduces data science and clinician effort and fatigue.</td>
</tr>
</tbody>
</table>

aTCP: transmission control protocol.
bREST: Representational State Transfer.
cJSON: JavaScript Object Notation.
eAES-256: Advanced Encryption Notation.
fCSS: Cascading Style Sheets.
fES: ECMA Script.

**LAMP Platform**

The LAMP platform is a customizable clinical care management and neuropsychiatric research platform designed around PGHD, as detailed in Figure 3. It comprises numerous essential features, such as customizable clinician-defined activities (eg, surveys, breathing exercises, journaling, and cognitive tests), collection and analysis of mobile and wearable sensor data, push notification scheduling, care team-centric conversations, just-in-time adaptive interventions, prebuilt featureization, visualization, or analysis pipelines, and a companion integrated development environment (IDE). The LAMP platform is available for use across any modern desktop web browser as well as recent versions of iPhone operating system and Android.
through the mindLAMP app available on the Apple and Google app stores, as shown in Figure 4. The backend is deployable using enterprise-standard orchestration tools (Docker and Kubernetes [31]) and has already been deployed across several health care systems and is used today by patients, clinicians, and researchers. The companion Cortex data analysis toolkit integrates tightly across the platform to provide a unified processing pipeline for secondary active and passive data features (measurable behavioral characteristics extracted from raw data), interactive visualizations, and the generation of targeted and automated adaptive interventions. The IDE is bundled with support for the widely adopted Python, R, and JavaScript programming languages and built atop Jupyter Notebooks and Visual Studio Code for collaborative data analysis.

The LAMP platform is designed to be customizable to fit a wide range of use cases and requirements, eliminating the need for multiple apps hosting only a set of fixed, immutable content as well as the research concern of proprietary data formats and closed-source commercial analysis software. It also securely enables data interoperability and extensibility, avoiding the issue of data silos without external access of collected data for clinicians or patients. It integrates into existing hospital organization structure and is not limited to the sandbox on an individual’s smartphone, allowing the caregivers and patients to coexist on the same platform. These features combined allow the LAMP platform to engage the care team through interactive clinical decision support with adaptive responses to incoming PGHD. Where a self-contained app must focus on solving individual problems for specific stakeholders, the LAMP platform focuses on broader challenges around linking people with data and data to teams of interconnected stakeholders, from patients and clinicians and family members and the care team to administrators and research coordinators. For these reasons, the LAMP platform supports digital phenotyping+, the plus symbol indicates the ability to return and share PGHD with the individuals from which it is collected in a secure and ethical manner and the ability to integrate that data into a machine learning pipeline or other clinical decision support algorithms.

Figure 3. The Learn, Assess, Manage, and Prevent platform consists of three major components: (1) mindLAMP, the patient- and clinician-facing app and web dashboard; (2) Data center, providing secure storage and access to data; and (3) Cortex, the data analysis toolkit that enables adaptive interventions and interactive visualizations. API: application programming interface; HIPAA: Health Insurance Portability and Accountability Act; iOS: iPhone operating system; SDK: software development kit.
Figure 4. Screenshots of the home screen as viewed by a patient in the mindLAMP app on a smartphone device; system administrators install applets such as Survey, Breathe, or Tips that clinicians and researchers are able to configure and schedule so that the participants interact with the app and produce data and metadata.

Results

Overview

The rearchitected LAMP platform addresses the integration of PGHD into existing systems through a simplified extensible programming interface (application programming interface [API]) and an internal data representation that interlinks raw data and metadata with descriptive schema. The Cortex data analysis toolkit obviates the need for custom preprocessing or harmonization of disparate sources of data and removes barriers between the real-time collection of PGHD and subsequent featurization, analysis, or visualization. We present the process results and examples below but do not offer a hypothesis in line with other papers exploring informatics systems created for use in clinical care and research.

Integration to Existing Systems

To enable robust data analysis, adaptive interventions, and interoperability with a broad range of health care systems and services, the platform’s data repository and programming interface are based upon a concise semantic FHIR-compatible API. The platform’s API provides both predefined and pluggable schematics for patient-facing instruments, such as surveys and cognitive tests as well as for mobile and wearable sensors. The platform’s backend validates and harmonizes patient data upon receipt, retaining lossless FHIR compatibility in the process.

The platform provides a facility to query and transform data into FHIR-compatible bundle and resource types, in addition to other domain-specific tabular or structured data formats. Table 2 lists the clinical, regulatory, and software standards implemented and supported by the LAMP platform.

The LAMP platform’s internal data representation provides a simplified abstraction around PGHD in comparison with FHIR. Fundamentally, FHIR adapts a message and document-based exchange programming interface atop the representational exchange state transfer web standard protocol. The FHIR data structures (Figure 5) consist of over 90 modules for clinical use, insurance, billing, and other use based on the concept of resources, with each resource containing some raw data, metadata, a schema identifier, and a human-readable representation of the raw data. The schema identifier is used to reference how the data should be interpreted by a compatible system or machine. As the raw data contained within 2 resources of the same schema type may differ (eg, the use of the observation data type to represent both blood pressure and depression assessments results), data processing cannot be standardized across different data types. By organizing and accessing PGHD separately from generalized repositories of data, such as electronic health record systems using the FHIR API, common and shared analysis methods and processing tools that are standardized across such various data types can be used by clinicians and researchers.
Figure 5. The structured list of all supported Fast Healthcare Interoperability Resources core resources; the structured list of all Learn, Assess, Manage, and Prevent platform resources.

The LAMP platform declares only 12 core PGHD-centric resources (Figure 5) that remain focused on clinical and research use. The **Researcher** and **Study** resource types group together sets of **participants** as well as the **activities** and **sensors** they are able to interact with or collect data from. Upon data collection, **ActivityEvent** and **SensorEvent** describe and link the recorded data to its metadata and any specific customized parameters. The **Credential** resource provides security access controls to any of the aforementioned resources, and the **Tag** resource provides support for integration, extensibility, and backward compatibility. The semantic context of any recorded data and metadata is described by **ActivitySpec**, **SensorSpec**, and **TagSpec**.
Internal Data Representation

Understanding the need for seamless integration into existing health care systems, software, and services, the platform exposes its internal data representation through the LAMP protocol, a programming interface that enables integration with third-party services. For example, integration with Google Fit wearable devices that also implement this same push-based model, is possible by signing up on the Fitbit developer portal and connecting the data output of the Fitbit programming interface to the data input in the programming interface provided by the LAMP platform. In another example of seamless integration, clinicians and researchers can use automated scripts to synchronize data between the mindLAMP app and their existing record-keeping systems. Data can be proactively fetched and stored securely, and users of the platform are notified of any or all data from a particular patient using subscriptions, regardless of whether the data were generated by the mindLAMP app or a third-party data source.

The extensibility and flexibility of patient-facing instruments in the LAMP platform rely on the unique data structure and functionality provided by the LAMP protocol as shown in Figure 2. Each activity with which patients are able to interact is defined and encapsulated in an activity specification that contains the program code written using web-compatible standards, along with descriptors of the required input configuration and output data. When a patient begins an interactive session with any activity, session-wide metadata regarding who, what, and when are recorded. Each tap of the screen within the activity is then automatically validated and converted into a standardized data format called a temporal slice. When the user completes the interactive session, all the temporal slices are packaged into chronologically ordered events indexed under the patient’s identifier as a stream of continuously generated data. The data analyst is then able to query these data at any desired temporal resolution (eg, 1 millisecond, 1 day, and 1 year) and filter by the type of activity (eg, mood survey, anxiety survey, trails-making test, and meditation). The query can be mutated using transformation logic executed by the backend and subscribed such that newly uploaded data matching the query is reported in real time to the data analyst. This query framework can be used to better understand how participants use and engage with the activities available to them as part of the study, for example, by extracting a real-time metric of duration spent meditating in the app per participant.

As depicted in Figure 4, the flow of activity specifications to configured activities to their generated PGHD facilitates patient interaction with any kind of interactive web media, from static text for tips, to video content for learning modules, or audio content for breathing exercises. Instruments and their data can be monitored and maintained organization-wide for compliance and conformance. Multimedia Appendix 1 lists the sources of active and passive data currently available within the mindLAMP app and their data sources and types. This novel data organization and structure supported by the platform enables unification and harmonization of these different data types, with both backward compatibility to data types from legacy systems and future compatibility for data types for systems that are not yet available.

Data Analysis With Cortex

The same pipeline operates on both active and passive data, unifying the conceptual model for PGHD processing and obviating the need for individual analyses tied to custom code for specific sensor types across various devices. Sensor data are therefore subject to additional harmonization to account for the various differences in functionality and recording between Apple and Android devices. For example, accelerometer measurements taken on Apple devices are measured in units of gravity (G) with a frame of reference experiencing $-1 \text{ G}$ in the downward-facing axis, whereas measurements on Android are measured in meters per second square ($\text{m/s}^2$) without a frame of reference provided. As the platform automatically applies this harmonization step, the data analysis code does not require an intrinsic understanding of the source of the data. Samples of sensor data after harmonization are shown in Figure 6. Furthermore, in addition to raw sensors on smartphones or wearable devices, processed Apple HealthKit and Google Fit sensor data, such as activity recognition or heart rate variability, are available to the LAMP platform.

The Cortex data analysis toolkit further simplifies the extraction of passive data features as listed in Multimedia Appendix 1, with an example shown in Figure 7. Cortex provides prebuilt, parallelized and vectorized workflows in Python for PGHD extraction and featurization that operate across large data sets to generate interactive visualizations for the mindLAMP dashboard using the Vega visualization grammar (as listed in Table 2). It obviates the need to work directly with the LAMP protocol, allowing data scientists to reason about live actionable structured data entirely as data frames within their programming environment of choice. Through the vectorization of array operations and parallelization of function calls, Cortex is able to target high performance and cost-effectiveness, while maintaining data security and policy compliance. A sample execution plan for a particular analysis involving the GPS data is shown in Figure 8. The modular nature of PGHD captured by mindLAMP allows for personalization and creation of new digital biomarkers and analysis without the need for additional coding. Furthermore, the companion IDE manager abstracts away log-in and security issues by securely injecting an authenticated connection to the server into Cortex and the resulting analysis notebooks.
Figure 6. Samples of data from selected sensors in the mindLAMP app for a sample patient. Total duration (in seconds) spent in calls per day; cumulative number of steps taken per hour during a 24-hour rolling window; number of times the device’s screen was turned on per day; number of unique nearby devices (Wi-Fi or Bluetooth) encountered per day.

Figure 7. A visual representation of the various categories of activity and sensor data type features using standardized functions as part of the Cortex data analysis toolkit; shown as part of Cortex is the distinction between the primary and secondary feature types, where secondary features are composed of primary features as opposed to raw patient-generated health data. Availability of wearable sensors depends on the device type used and supported application programming interface; Apple Watch (HealthKit) sensors are shown here. DBT: dialectical behavioral therapy.
**Discussion**

**Principal Findings**

Research and clinical needs in digital medicine are evolving to use PGHD approaches to understand patient behavior and symptomatology [3]. To this end, by optimizing the system architecture for data throughput and substantial database write-loads, the LAMP platform supports high-performance data collection and real-time data analysis to enable, for example, larger machine learning models or just-in-time adaptive interventions that can leverage PGHD into actionable insights for patients and clinicians alike.

**Efficient Collection and Configuration**

Among the various approaches to data collection adopted in digital medicine, the pull-based model [13,14] shown in Figure 9, requires patients to activate a request and upload data from their mobile devices. This request can be scheduled and authorized. An example of the pull-based model is that during a clinical encounter, the clinician would use a portal to request data collection from the patient’s device for a period of 1 day; the patient would then have to approve this request in their smartphone app before data collection can begin. During the next clinical encounter, the clinician would be able to interpret the data in potentially meaningful ways.

The LAMP platform, however, adopts a push-based model shown in Figure 9, where, in contrast, clinicians or research coordinators configure and schedule activities for patients to use and sensors from which measurements should be passively recorded ahead of time. The patients’ devices receive a configuration request that activates data collection in the background. As it is collected, the data are uploaded (pushed) to the back end periodically, available for processing and clinical insight ahead of time. This push-based approach reduces latency from collection of PGHD to the usability of that PGHD, for example, as real-time alerts in the context of research studies (Figures 10 and 11), or toward clinical decision-making with custom rules and alerts. It is important that clinicians or research coordinators communicate clearly and establish consent with the patient or subject about the various types of data being collected and the frequency of the data push.
Figure 9. A pull-based model, in which clinicians must schedule data to be pulled from the device periodically versus a push-based model, in which clinicians preconfigure various sensors on the device to collect and push data to the server in real time.

Figure 10. The detailed coordination required among the many components of the Learn, Assess, Manage, and Prevent platform involved in the submission of a push notification for a survey or gift card email for completion of a study; an example of the reporting of live intervention processing as made possible by a push-based model. Upon participant enrollment, survey delivery, gift card delivery, or intervention triggering, a message is pushed and logged to the Slack messaging service, a push-based model, to alert the research coordinator in real time. API: application programming interface; APNS: Apple Push Notification Service; FCM: Firebase Cloud Messaging; LAMP: Learn, Assess, Manage, and Prevent; REST: Representational State Transfer; SNS: Simple Notification Service.
Figure 11. An example of the reporting of live intervention processing as made possible by a push-based model. Upon subject enrollment, survey delivery, gift card delivery, or intervention triggering, a message is pushed and logged to the Slack messaging service, also a push-based model, to alert the research coordinator in real time.

The data collection processes are executed both actively during patient interactions with the mindLAMP app as well as passively while the app or mobile device is not in use. These data are securely uploaded to the organization’s backend systems and can be used immediately upon receipt for data analysis and logic to select interventions to display to the patient. Using push notifications sent to the mobile device, the platform promotes a high level of engagement with patients without explicitly requiring approval for data upload. Once a research study or clinic is configured, its documenting configuration can be exported and reimported by other LAMP-compatible systems or interfaces. This allows reproducibility in both clinics and research studies, for example, by attaching the configuration file to a research manuscript or clinical protocol.

Consortium and Clinical Research Efforts
The LAMP platform is built and maintained collaboratively as an open platform to address the needs of many and integrate tools and resources to streamline workflows. In contrast to commercially available apps and services, the mindLAMP app may be used by organizations independently of our team through the deployment of a secure self-hosted backend. It can be customized and adapted without requiring specialized coding and deployment efforts, although others have also taken advantage of the extensibility of the platform to design and develop unique cognitive tests for their organization’s needs. Common data processing and analysis needs across clinical and research workflows are encapsulated by the platform and the Cortex data analysis toolkit to minimize the time between patient onboarding and affecting or assessing patient outcomes.

The LAMP platform is highly configurable to suit many needs across a broad range of both clinical and research use cases and strategies. Consortium partners are encouraged to share their use case and LAMP configurations. As shown in Textbox 1, there were many different potential configurations and use patterns across consortium members.
Textbox 1. Selected examples of configurations and use cases for the Learn, Assess, Manage, and Prevent platform across various consortium members.

**Ecological momentary assessment tool**
- A total of 17 adults with substance use disorder who recently discharged from the hospital for this diagnosis completed daily assessments delivered via mindLAMP of mood, anxiety, sleep, social activity, and craving. No sensors were configured, and push notifications were enabled as reminders. Participants were able to view select survey responses in the Prevent tab. Results will be presented by the study team at the 34th Annual European College of Neuropsychopharmacology meeting in October 2021.

**Digital phenotyping**
- A research study to examine circadian rhythms in bipolar disorder was conducted in which no activities were enabled or scheduled. The accelerometer, gyroscope, magnetometer, gravity, device motion, GPS, screen state, call and text, Bluetooth, and Wi-Fi sensors were enabled and configured to collect data at the highest possible frequency. Participants were able to view select sensor data in the Prevent tab.

**Both ecological momentary assessment and digital phenotyping**
- A research study was conducted in which 100 college students were remotely enrolled to use the mindLAMP app for 1 month. Participants took 1 scheduled daily survey and 1 scheduled weekly survey, with provided optional tips and resources for managing stress, depression, and anxiety. The accelerometer, gyroscope, magnetometer, gravity, device motion, GPS, screen state, call and text, Bluetooth, and Wi-Fi sensors were enabled and configured to collect data at the highest possible frequency. Results are summarized in a paper published in 2021 [32].

**Individual patient study**
- A research study was conducted in which 50 participants with schizophrenia or bipolar disorder used the mindLAMP app for 1 year. Each participant was scheduled a daily standard battery of several surveys, the Jewels cognitive test, and the Spatial Span cognitive test. Optional tips and resources were provided, and the journaling, scratch card, and breathing exercise activities were made available for participants to use on their own volition. Participants were able to view their own data in the Prevent tab and worked with the research coordinator and psychiatrist to create custom surveys specific to individual participants’ needs. For example, 1 participant chose to create a water intake survey. Each participant’s notifications were scheduled individually by the research coordinator, instead of at a set time across all participants. The accelerometer, gyroscope, magnetometer, gravity, device motion, GPS, screen state, and call and text sensors were enabled and configured to collect data at the highest possible frequency.

**Clinical use**
- A clinical team in California offered dialectical behavioral therapy diary cards to all patients via mindLAMP. Before clinic visits and during their daily lives, patients would fill out these app-based dialectical behavioral therapy diary cards on their mobile device and were able to see their previous responses in the Prevent tab. No other activities were made available to the patients and no sensors were enabled for data collection. The diary cards were reviewed during each clinical session with the dialectical behavioral therapy therapist.

**Digital clinic**
- A clinic was established in which patients used the mindLAMP app with a new care team in addition to their ongoing care. Each patient was scheduled a daily standard battery of several surveys. Required tips and resources were provided along with required journaling, scratch card, and breathing exercise activities that were scheduled according to patient preferences. Patients were able to view their own data in the Prevent tab and worked with the digital navigator and clinician to create custom surveys or activities specific to individual patient’s needs. For example, 1 patient requested a set of self-management tips and resources. Each patient’s notifications were scheduled individually by the digital navigator. On the basis of each patient’s clinical goals each week, accelerometer, gyroscope, magnetometer, gravity, device motion, GPS, screen state, call and text, and other sensors were turned off or on with the goal of capturing relevant and actionable information to help manage care. A protocol for the clinic is published here [33].

**Intervention tool**
- A research team used the cognitive games in mindLAMP as a tool for cognitive remediation to offer attention and memory training to patients with clinical high risk for psychosis. A paper summarizing the results was published in 2021 [34]. The app was offered in the Mandarin Chinese language for this study.

To aid these joint research and clinical efforts, the LAMP consortium was founded to connect partners using the LAMP platform. The design and development of the platform occurs in an open-source, collaborative environment that many have taken advantage of to suggest features, report bugs, add documentation, and improve the overall quality and efficacy of the LAMP platform. Through its flexibility and interoperability, the platform encourages integration and cross talk between clinical and research contexts, and to this end, supports the implementation of a digital clinic [33] and the creation of a digital navigator role [35].

**Next Steps**

The consortium further integrates directly into the development and feedback cycle using a community forum and bug tracking system, both available publicly. The community forum serves as a centralized resource for multiple teams or organizations to work with one another to assist with data analysis or troubleshooting and provide feedback about the LAMP platform. In addition, collaborators actively engage in making modifications to the source code (hosted through the public source code repository hosting service GitHub), make any suggested modifications or bug fixes, and then request that these
changes be merged upstream into the distribution of the LAMP platform that is used by all. To learn more about the LAMP platform or help contribute, one can join the consortium or visit the open-source repository [36].

Conclusions
Through the incorporation of patient- and clinician-centric feedback as well as a standards-based approach, the LAMP platform is designed to address important needs around the effective and compatible integration of PGHD into existing clinical systems for research and clinical care. It offers a flexible and comprehensive set of tools and solutions that can be configured and stitched together to function in a wide range of use cases, as used by members of the LAMP consortium. Its simplified programming interfaces are designed to securely handle a high throughput of PGHD as well as its companion metadata. With the integration of the Cortex data analysis toolkit, machine learning feature extraction, data processing, interactive visualization, and other essential tasks are simplified and coordinated seamlessly at low cost and high efficiency. In addressing technical challenges, the LAMP platform enables research and clinical teams to rapidly convert PGHD from widely accessible consumer smartphones and wearable devices into actionable clinical insights.

Acknowledgments
This work was supported by a philanthropic gift from Jeremy Wertheimer.

Conflicts of Interest
None declared.

Multimedia Appendix 1
A full listing of Active, Passive, and Cortex data types currently supported by the Learn, Assess, Manage, and Prevent platform along with a description and expected components of the data.

[DOCX File, 18 KB - mhealth_v10i1e30557_app1.docx ]

References
12. ITI homepage. ITI. URL: https://www.itic.org/ [accessed 2021-12-15]


16. mCODE™: Minimal Common Oncology Data Elements. mCODE. URL: https://mcodeinitiative.org/ [accessed 2021-12-15]


29. HTML Living Standard. HTML. URL: https://html.spec.whatwg.org [accessed 2021-12-15]

30. Vega. URL: https://vega.github.io [accessed 2021-12-15]


IDE: integrated development environment  
LAMP: Learn, Assess, Manage, and Prevent  
PGHD: patient-generated health data  
SMART: Specific, Measurable, Achievable, Realistic, and Timely