JMIR mHealth and uHealth

Impact Factor (2024): 5.4
Volume 10 (2022), Issue 10 ISSN 2291-5222 Editor in Chief: Lorraine Buis, PhD, MSI

Contents

Reviews

Original Papers

XSL•FO
RenderX

Corrigenda and Addenda

XSL•FO
RenderX

Review

Measurement Properties of Smartphone Approaches to Assess Physical Activity in Healthy Young People: Systematic Review

Belinda Parmenter[1], BHtlhSc, PhD; Claire Burley[1,2], BSc (Hons), MSc, PhD; Courtney Stewart[3], BSc (Hons); Jesse Whife[3], BSc; Katrina Champion[4], BAPsych (Hons), B Health, PhD; Bridie Osman[4], BSc (Hons); Nicola Newton[4], BPysch (Hons), PhD; Olivia Green[4], BA; Annie B Wescott[5], BA, PhD; Lauren A Gardner[4], BPysch (Hons), PhD; Rachel Visontay[4], BA (Hons); Louise Birrell[4], BSocSc, BPsych (Hons), PhD; Zachary Bryant[4], BComm, MPH; Cath Chapman[4], BA (Hons), PhD; David R Lubans[6,7,8], BEd, MSc, PhD; Matthew Sunderland[4], BPysch (Hons), PhD; Tim Slade[4], BA, PhD; Louise Thornton[4], BPsych (Hons), PhD

[1]School of Health Sciences, University of New South Wales, Kensington, Australia

[2]Centre for Healthy Brain Ageing, Discipline of Psychiatry & Mental Health, University of New South Wales, Sydney, NSW, Australia

[3]National Drug Research Institute and enAble Institute, Faculty of Health Sciences, Curtin University, Perth, WA, Australia

[4]The Matilda Centre, The University of Sydney, Sydney, NSW, Australia

[5]Galter Health Sciences Library & Learning Center, Northwestern University Feinberg School of Medicine, Chicago, IL, United States

[6]Centre for Active Living and Learning, College of Human and Social Futures, University of Newcastle, Callaghan, NSW, Australia

[7]Hunter Medical Research Institute, New Lambton Heights, NSW, Australia

[8]Faculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä, Finland

Corresponding Author:
Claire Burley, BSc (Hons), MSc, PhD
School of Health Sciences
University of New South Wales
Room 234, Level 2 Wallace Wurth Building
Corner of High Street and Botany Street
Kensington, NSW 2052
Australia
Phone: 61 0290653510
Email: c.burley@unsw.edu.au

Abstract

Background: Physical inactivity is a preventable risk factor for several chronic diseases and one of the driving forces behind the growing global burden of disease. Recent evidence has shown that interventions using mobile smartphone apps can promote a significant increase in physical activity (PA) levels. However, the accuracy and reliability of using apps is unknown.

Objective: The aim of our review was to determine the accuracy and reliability of using mobile apps to measure PA levels in young people. We conducted a systematic review guided by PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses).

Methods: Studies published from 2007 to 2020 were sourced from 8 databases—Ovid MEDLINE, Embase (Elsevier), Cochrane Library (Wiley), PsychINFO (EBSCOhost), CINAHL (EBSCOhost), Web of Science (Clarivate), SPORTDiscus (EBSCOhost), and IEEE Xplore Digital Library database. Studies were conducted in young people aged 10-24 years and without chronic illnesses, who evaluated a mobile app's ability to measure PA. Primary outcomes included validity, reliability, and responsiveness of the measurement approach. Duplicate screening was conducted for eligibility, data extraction, and assessing the risk of bias. Results were reported as a systematic review. The main physical activity measures evaluated for each study were the following: total PA time (min/day or min/week), total moderate to vigorous PA per week, daily step count, intensity measure (heart rate), and frequency measure (days per week).

Results: Of the 149 identified studies, 5 met the inclusion criteria (322 participants, 176 female; mean age 14, SD 3 years). A total of 3 studies measured criterion validity and compared PA measured via apps against PA measured via an Actigraph accelerometer. The 2 studies that reported on construct validity identified a significant difference between self-reported PA and

XSL•FO
RenderX

the objective measure. Only 1 of the 5 apps examined was available to the public, and although this app was highly accepted by young people, the app recorded PA to be significantly different to participants' self-reported PA.

Conclusions: Overall, few studies assess the reliability, validity, and responsiveness of mobile apps to measure PA in healthy young people, with studies typically only reporting on one measurement property. Of the 3 studies that measured validity, all concluded that mobile phones were acceptable and valid tools. More research is needed into the validity and reliability of smartphone apps to measure PA levels in this population as well as in populations with other characteristics, including other age groups and those with chronic diseases.

Trial Registration: PROSPERO CRD42019122242; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=122242

(JMIR Mhealth Uhealth 2022;10(10):e39085) doi:10.2196/39085

KEYWORDS

smartphone; mobile phone; mHealth; prevention; risk; physical activity; sedentary behavior; young people

Introduction

Physical inactivity is a preventable risk factor for several chronic diseases and one of the driving forces behind the growing global burden of disease [1,2]. Physical inactivity and excessive sedentary behavior are increasing, especially in young people. A review of Australia's health published in 2018 showed that 92% of young people aged 13-17 years did not meet the physical activity (PA) guidelines of 60 minutes of moderate to vigorous intensity PA each day [1]. Similar global trends have also been reported [2-4]. In addition, independent of PA, total sitting time and TV viewing time are also associated with greater risk for several major chronic disease outcomes [5]. However, data from Australia's health 2018 indicate that only 20% of young people meet the sedentary screen-based behavior guideline [1]. Recent cross-sectional and large population-based cohort studies published by the authors of this paper [6-8] (sample sizes of 3826, 6640, and 231,048 participants) indicated that, on average, 85.9% and 77.7 % of Australian adolescents engage in too much recreational screen time and do not meet the PA guidelines, respectively [6-8].

Recent evidence has shown that interventions using mobile smartphone apps or activity trackers can promote a significant increase in PA levels [9]. Smartphone apps can track PA and enable continuous self-monitoring and feedback on PA through heart rate, step counts, and exercise type, duration, and intensity. Other benefits of using smartphones for PA tracking are high rates of smartphone ownership in young people, widespread use of smartphone tools to track PA, tools being less burdensome than traditional measures (eg, standalone step counter, heart rate monitor, or pen and paper), and ability to be used in remote locations.

If step counts and exercise duration and intensity are to be effectively used as reference values for achieving recommended PA levels, we need to ensure that the tools we are using to measure activity are accurate for the population in which we are measuring them. The Lancet PA Series Working Group's [10] current recommendation for continued improvement in monitoring PA to help guide policies to increase activity levels further highlights the need for accuracy and reliability of PA monitors via smartphone technology, which has developed significantly over the past decade. Previous research has assessed the accuracy of PA data measured by smartphones to influence PA [11]. However, to our knowledge, there is no systematic review of the literature with respect to the validity and reliability of using smartphones to quantify PA levels.

The promotion of PA, and tools available to monitor it, has emerged as an important area of research that has drawn increasing interest from researchers and health professionals in a variety of fields. The growing availability of inexpensive parts and equipment has led to the development of mobile devices such as smartphones, providing platforms for new opportunities in health care and the promotion of PA. Accelerometers have emerged as the most useful and extensive tool to capture and assess human physical activities in a continuous, unobtrusive, and reliable manner, but they are expensive and not practical in all settings. Recent evidence [9] has shown that using apps or PA trackers is effective at promoting PA; however, if PA recommendations for optimal health are based on minutes per week and at least 60 minutes per day for young people, we need to ensure that these measures are accurate and reliable.

The primary aim of this review was to assess the accuracy and reliability of using smartphones to measure PA levels in young people aged 10-24 years. This sample was chosen to avoid issues regarding generalizability and to enable a clearer understanding on the reliability of these measures to be established first (ie, by focusing on healthy young people with less variability caused by ageing, comorbidities, and health history); it was also chosen because of behavioral differences in populations (eg, smartphone use in younger compared to older adults).

The specific objectives of this review were to (1) identify and describe the ways in which PA has been measured using smartphones in young people; (2) describe and critically evaluate the available evidence on the measurement properties and feasibility of these measurement processes; and (3) provide recommendations on the most suitable and effective ways of measuring PA.

Methods

This review was conducted according to a published protocol [12] and in line with the 2020 PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analysis) reporting guidelines [13]. This review forms part of a larger review that has been registered with the International Register of Systematic Reviews in PROSPERO (CRD42019122242) [12,14]. This

XSL•FO
RenderX

larger review aimed to examine the measurement properties of smartphone approaches to assess 6 key health behaviors (ie, PA, sedentary activity, sleep, diet, alcohol use, and tobacco use, also known as the 'big six') that are recognized risk factors related to the development of chronic diseases [15]. However, studies identified for inclusion were heterogeneous, recruiting different populations and using different measurement methods across various health behaviors. Therefore, only those studies that examined PA, specifically in healthy young people, were included in this review. Findings regarding alcohol use, tobacco use, and diet are described in Thornton et al [12,14]. Methods for this review continued to follow those outlined in Thornton et al [14], with any differences in methods presented in the following sections. Only articles published between 2007 to 2022 were searched because smartphones with large touchscreens (ie, where users input directly using their finger) were not available before.

Search Strategy and Selection Criteria

A research librarian searched 8 web-based databases—Ovid MEDLINE, Embase (Elsevier), Cochrane Library (Wiley), PsychINFO (EBSCOhost), CINAHL (EBSCOHost), Web of Science (Clarivate), SPORTDiscus (EBSCOhost), and IEEE Xplore Digital Library database—as per the search strategy published in previous studies [12,14], using particular search terms (Table S1 in Multimedia Appendix 1). For inclusion in the current review, studies were required to describe a smartphone-based approach to assess PA in healthy young people and to report on at least one measurement property of this approach identified in the Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN) taxonomy of measurement properties [16]. The COSMIN checklist assesses the methodological quality of studies investigating the measurement properties of patient-reported outcome measures. Measurement properties assessed in this study include internal consistency, reliability, measurement error, content, construct, criterion validity, and responsiveness; definitions are outlined in Table 1 according to COSMIN. Healthy young people were defined according to the World Health Organization as persons aged between 10 and 24 years [17] with no known chronic conditions. Outcomes of interest included measurement effectiveness (ie, accuracy and reliability) of a smartphone to measure PA, reported as total PA time (min/day or min/week), total moderate to vigorous PA per week, daily step count, intensity measure (heart rate or rate of perceived exertion), or frequency measure (days per week).

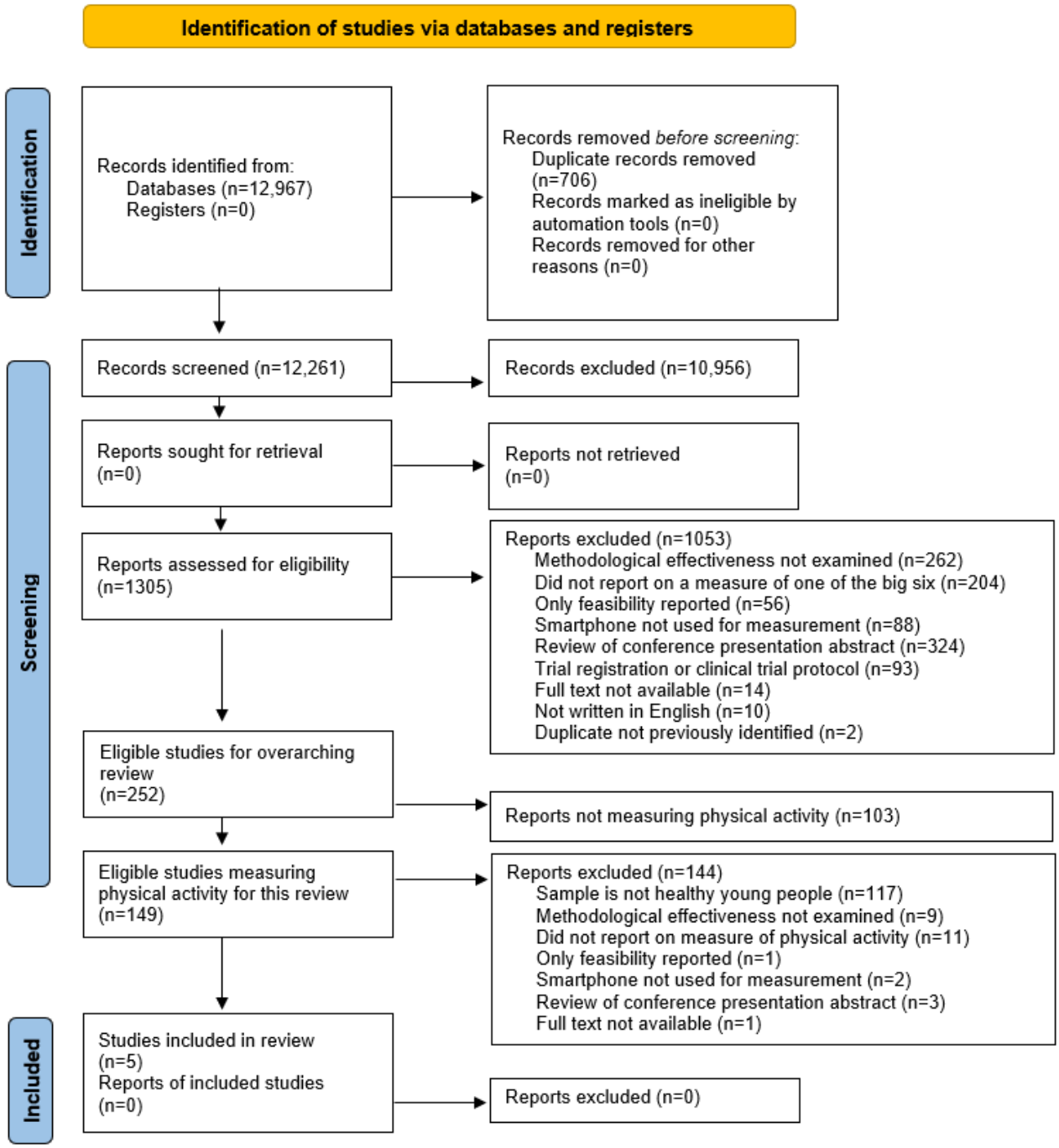
Studies were excluded if participants were not healthy young adults; they were also excluded if they did not examine methodological effectiveness; did not report on PA and feasibility; did not use smartphones; were not published in English; if they described feasibility of the measurement approach only; described measurement properties of using text messaging only to measure behaviors; and described the measurement properties of a wearable device (eg, Fitbit) alone (Figure 1).

Table 1. Measurement validity criterion types and their definition according to Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN) assessed in this study.

Criterion	Definition
Internal consistency	The degree of interrelatedness among devices
Reliability	The degree to which the measurement is free from measurement error
Measurement error	The difference between a measured quantity and its true value
Content validity	The degree to which the device being assessed is an adequate representation of the construct being measured
Construct validity	The degree to which the results are consistent with the hypothesis, based on the assumption that the device validly measures the construct to be measured
Criterion validity	The degree to which the results are adequate reflections of a 'gold standard'
Responsiveness	The ability of a device to detect change over time in the construct being measured

XSL•FO
RenderX

Figure 1. PRISMA (preferred reporting items for systematic review and meta-analysis) flow chart of the search strategy outcomes, including those of the larger review conducted by Thornton et al [12,14], followed by the narrowed search on studies measuring physical activity in healthy young people. Big six: 6 identified risk factors related to the development of chronic disease [15]: physical inactivity, sedentary time, poor sleep, poor diet, and alcohol and tobacco use.



Data Extraction

All identified studies were exported into Endnote (version X9) for removal of duplicates. Records were then uploaded to Covidence Systematic Review software (Veritas Health Innovation) for screening. Authors participating in the screening, full-text review, and data extraction process attended training sessions, where multiple reviewers independently reviewed and discussed the same selection of articles to help ensure consistency across reviewers. As described in Thornton et al [12,14] for the larger review, titles and abstracts were screened by one reviewer (OG, RV, JW, CS, LT, BO, LB, LG, OG, ZB, KC, or BP) and then the full text of potentially relevant studies was independently assessed for eligibility by at least two members of this group.

Titles and abstracts of included studies were then screened according to the specific inclusion criteria of this review by one reviewer (BP). The full texts of potentially relevant studies were independently assessed for eligibility by 2 authors of the research team (CS and JW), with any disagreements resolved with the assistance of a third or fourth researcher (LT and BP). CS and JW independently extracted data in duplicate using a standardized extraction form to ensure that it adequately

XSL•FO
RenderX

captured trial data, and they agreed on final reported data. Further details of data extraction are included in a previously published protocol [12].

Physical Activity Measures, Measurement Validity, and Risk of Bias

The primary outcomes of interest for this review were PA measurement properties of smartphone-based approaches to assess PA. Specifically, we investigated, as reported in included studies, the reliability, measurement error, content validity, construct validity (including convergent validity, structural validity, and cross-cultural validity), criterion validity, and responsiveness of the identified approaches. In this review, for studies to be classified as measuring criterion validity, the smartphone-based approach of interest must have been compared to an objective measure of PA that had already been tested for reliability (eg, step count using a pedometer with an internal spring that moves up and down with hip motion or an Actigraph accelerometer that uses small motion sensors to measure acceleration along 3 axes) [18]. Where the smartphone-based approach was compared to a self-report measure, even if it was described as the gold standard method by the studies' authors, the paper was classified as investigating construct and specifically convergent validity. For this review, the gold standard measurement tool for PA was classed as the Actigraph accelerometer [19,20]. Risk of bias of the included studies was assessed using the COSMIN risk of bias checklist [16,21,22].

Results

Summary of Included Studies

Of the 12,967 records identified through the search strategy (Figure 1), 149 studies met the PA smartphone criteria for this review without restricting studies by age. In summary, articles were excluded for not examining the methodological effectiveness of the measurement approach, not reporting on at least one of the 'big six' [15] or on feasibility, and not using a smartphone (Figure 1). When applying the inclusion criteria for participants to be healthy young people, only 5 studies [23-27] were eligible. The key characteristics of the included study populations are presented in Table 2. In total, 322 young people were included in the studies, and 176 were female, with a mean age of 14.3 (SD 3.2) years. Only 2 studies reported the age range of participants. Jongprasithporn et al [25] included participants aged 18-23 year, whereas Dunton et al [24] studied participants aged 9-13 years. As 90% of the participants were aged 10-14 years in the study by Dunton et al, the authors decided to include it in this review. Two studies reported the study was conducted with young people who were overweight or obese with a mean BMI of 32.7 kg/m^2 [27] and 31.3 kg/m^2, respectively [26].

Two studies were conducted in the United States [23,24], two in Germany [26,27], and one in Thailand [25]. Of the included 5 studies, all were assessing accuracy of smartphone apps collecting PA data, compared to the more subjective self-reporting or objective accelerometers.

The following 3 studies measured criterion validity, comparing phone data to objectively measured and automatically collected data: Bruening et al [23] completed an accuracy study to test the validity of a mobile ecological momentary assessment (EMA) methodology to assess PA levels ; Dunton et al [24] conducted a validity and feasibility study investigating electronic EMA of PA via smartphone-administered surveys; and the final study [25] examined the accuracy of the Thai 3 Axis (or TH3AX) smartphone app to assess PA. The 2 studies that measured construct validity [26,27] were pilot studies conducted on a group participating in an existing hospital treatment program for weight loss. Both studies assessed the feasibility and acceptability of modern electronic health care technology in treatment compared to self-report, where the user manually recorded bouts of PA across the study period.

Characteristics of included studies are outlined in Table 3. With respect to the apps used to record the PA data, 3 studies did not report the name of the app used, one used a mobile EMA app called devilSPARC [23], and one study [25] used the TH3AX app, which records PA data in real time. Four studies [23-25,27] did not report if the apps were available to the public. Schiel et al [26] reported the app was available to the public, but it could not be located in the Apple or Google Play app stores.

In 2 studies [23,24], participants were sent real time prompts throughout the day to answer questions about the activity they were undertaking. The 3 remaining studies [25-27] used real-time monitoring of PA though mobile phone motion. Only one of the studies [25] that used passive motion sensing reported where the participants were instructed to wear the phone; this study asked participants to attach the phone to their right hip in the first test and then the right anterior superior iliac spine in the second test. Of the 5 studies, 4 monitored activities over an average of 4 days. The remaining study [25] monitored activities over a range of performance trials, where participants randomly took part in standing, walking, and running activities. All 5 studies examined PA measurement on Android phones, with only 1 study [23] looking at the feasibility of PA measures on the Apple iOS. In this study, to those participants who were interested in participating but did not own an Android or iOS mobile phone, a Motorola Moto G was loaned to be used for the duration of the study. Only two [23,24] of the 5 studies mentioned the actual type and brand of the phones used. Of the 5 studies, 2 reported that they offered financial incentives (US $80 and US $40) for participating in the study [23,25].

Two studies [23,24] required the user to actively enter bouts of PA data. In the 3 remaining studies [25-27], PA data were automatically collected through the motion or movement of the mobile phone. Two studies [26,27] used a mobile motion sensor that was integrated within the phone. One study [25] used an accelerometer that was built into the phone. All 3 of these studies [25-27] reported the algorithms used to compute PA behavior; however, they did not report whether the algorithm used was accessible via open source. None of the studies reported if participants wore a PA tracking monitor for the duration of the study. Sedentary behavior was measured in 4 of the 5 studies, with 2 [26,27] recording it objectively and automatically through the phone, and 2 through subjective self-report measures [23,24].

XSL•FO
RenderX

Table 2. Characteristics of the included study populations.

Author and year	Population in the study					Instrument administration			
	Sample size (n)	Age (years), mean (SD)	Age (years), range	Gender (female), n (%)	Other characteristics	Setting	Country	Language	Device
Schiel (2010) [27]	30	14 (3)	NR[a]	14 (47)	Overweight or obese (BMI 32.7 kg/m^2)	Hospital treatment program	Germany	English	Android smartphone
Schiel (2012) [26]	124	13.5 (2.8)	NR	69 (56)	Overweight or obese (BMI mean 31.3, SD 5.2 kg/m^2)	Weight reduction program	Germany	English	Android smartphone
Dunton (2013) [24]	121	11 (NA[b])	9-13	59 (49)	38% risk of overweight or overweight	4th-8th grade	United States	English	Android smartphone
Bruening (2016) [23]	41	18.7 (0.5)	NR	30 (73)	Friendship networks and weight-related behaviors	College freshman	United States	English	Android or iOS smartphone
Jongprasithporn (2017) [25]	6	NR	18-23	4 (67)	Normal weight (BMI 21.5 kg/m^2)	Healthy young adults	Thailand	English	Android smartphone

[a]NR: not reported.

[b]Not available.

Table 3. Key characteristics of studies examining the measurement of physical activity via a smartphone.

Author and year	App name	Risk of bias properties assessed									
		Publicly available	Measurement approach	Internal consistency	Reliability	Measurement error	Content validity	Construct validity	Criterion validity	Responsiveness	Overall rating
Schiel (2010) [27]	NR[a]	NR	Passive[b]-objective	No	No	No	No	Yes	No	No	Inadequate
Schiel (2012) [26]	NR	Yes	Passive-objective	No	No	No	No	Yes	No	No	Inadequate
Dunton (2013) [24]	NR	NR	Active[c] self-report	No	No	No	No	No	Yes	No	Very good
Bruening (2016) [23]	devilSPARC	NR	Active self-report	No	No	No	No	No	Yes	No	Very good
Jongprasithporn (2017) [25]	Thai 3 Axis	NR	Passive-objective	No	No	Yes	No	No	Yes	No	Very good

[a]NR: not reported.

[b]Passive: data automatically collected.

[c]Active: requires user to do something.

Risk of Bias Measurement Properties

Overall, the use of measurement properties to assess reliability, validity, and responsiveness was poor, with each study only reporting on one measurement property completed. No studies looked at the reliability of their results through repeated measures. Only 1 study [25] reported on an overall measurement error, which was low at 0.12. No studies looked at structural validity, that is, the degree to which studies measured PA levels. No studies looked at cross-cultural validity and the applicability of results across other cultures.

Two studies reported on construct validity, both of which were conducted by the same research team [26,27]. In the first study [26], there were significant differences between self-reported PA and the PA measured by motion sensors in the phone. In general, the duration of PA documented by children and adolescents was much higher than the duration measured via motion sensors. Correlation analyses, however, revealed

XSL•FO
RenderX

moderate to strong significant correlations between the calculated duration of PA and the time spent in activities such as cycling (r=0.67; P<.01), the calculated duration of PA and the total amount of activity units (r=0.89; P<.01), and the calculated duration of PA and energy expenditure (r=0.82; P<.01) [27]. This demonstrated the strong consistency between PA assessment via a mobile motion sensor board and reality [26]. In the second study [27], the duration of PA estimated by children and adolescents was also significantly higher compared to the measured values for walking and running in the first study [26]. There was no difference between the estimates for cycling in the two studies. There were also weak to moderate significant correlations between the total calculated duration of PA and the time spent in some of the different activities, such as cycling (r=0.67; P=.001), driving (r=0.46, P=.01). There was a strong significant correlation between the measured duration of PA and the total number of activity units (r=0.89, P=.001). There was also a strong significant correlation between the measured duration of PA and the estimated energy expenditure (r=0.82, P=.001) [27].

The 3 remaining studies [23-25] specifically investigated criterion validity, where the app was compared to a previously validated objective measurement [22]. Although, one study [24] reported it as construct validity. In Dunton et al [24], the methodology of EMA activity responses was tested by examining differences in the mean number of steps (measured by an accelerometer) across EMA-reported activity categories. We have, therefore, reported this under criterion validity. Across both weight status groups, steps were significantly higher for EMA surveys reporting active play, sports, or exercise compared to any other type of activity. In addition, the mean number of steps recorded while talking on the phone, doing chores, riding in a car, and something else were significantly greater than mean steps recorded while reading, using a computer, doing homework, watching TV or movies, and playing video games [24]. However, this study concluded that it is acceptable and valid to use mobile phone EMA technology to measure PA and sedentary behavior in children aged 9-13 years during leisure time. In the second study [23], the odds of a participant having their accelerometer-derived activity level match their reported PA level were significant for mobile-based EMA-reported sedentary PA, light PA, and moderate PA. Due to only one participant having vigorous accelerometer values, odds were not computed for vigorous activity. The match rates were highest for EMA-reported sedentary and light PA (340/565, 60.3% and 37/63, 58.7%, respectively) and lowest for moderate PA (9/40, 22.5%) and vigorous PA (1/26, 3.8%). This study concluded that the devilSPARC mobile EMA app is valid for assessing the presence of sedentary activities during the day [23]. The third study [25] on criterion validity reported that the lowest sensitivity (0.975) of the TH3AX app was computed during standing activity. The highest sensitivity (0.988), specificity, and accuracy were all identified during running activity. The average sensitivity, specificity, and accuracy of TH3AX for standing, walking, and running were 0.981, 0.988, and 0.986, respectively. Based on these results, the authors validated the use of the smartphone app for activity recognition in young people. No studies looked at the responsiveness of the apps used and how well they detected change over time in PA levels.

Discussion

The primary aim of this review was to assess the accuracy and reliability of using smartphones to measure PA levels in young people aged 10-24 years. In summary, 5 studies met the inclusion criteria (including 322 young people, with a mean age of 14.3, SD 3.2 years) and objectively assessed the accuracy of smartphone apps to collect PA data. Data were either collected automatically via phone movement or manually by the user. The overall rating for 3 studies was considered ‘very good,’ and the remaining 2 studies were rated ‘Inadequate’ (Table 3). Only one study reported that the app was available to the public [26], but it could not be located in the Apple or Google Play app stores.

Results from this review suggest that much more research is needed on validating these apps against the gold standard tools such as the Actigraph accelerometer. With only 5 studies eligible for inclusion in this review, and only 3 of those studies comparing mobile apps to more traditional objective measures of PA, more research is needed on the efficacy and reliability of mobile phone tracking and monitoring of PA in young people. Furthermore, when reporting on the outcomes of these studies, more information should be provided on participant characteristics. Most of the studies in this review only reported on age and gender, with 3 of them providing information on weight status. Since only over 18% of children and adolescents in the world were reported as overweight or obese in 2016 [28], future studies on PA should start reporting weight status of young people to assist with identifying interventions that may reduce weight in this population.

It is crucial that studies report the most accurate placement site for general community use, as most people tend to carry their mobile phones in their pockets or handbags. For objective monitoring, the attachment site of the phone and the place of the accelerometer to which the phone was compared is important information for people relying on mobile phone data for PA monitoring. Previous work from our group showed that some motion senses and activity trackers work better and more accurate on specific parts of the body [29].

Of the 5 apps examined in this review, only 1 app is publicly available, and 3 of the 5 studies did not report on the public availability of the apps the researchers used. PA tracking apps should be publicly and freely available to have the greatest impact on improving PA levels and preventing the development of chronic diseases. In addition, with the Apple iPhone being the most popular smartphone on the market, more studies are needed on the efficacy of PA measurements using this system.

Limitations

Due to the small number of studies examining the efficacy, reliability, and validity of mobile apps measuring and tracking PA we were unable to complete a meta-analysis on the data. Each study’s methodology was so different, and it was not practical to combine study data. Some studies did not report on the actual measurement property used to examine reliability, validity, or responsiveness. In this study, we analyzed the information provided by one group under criterion validity,

XSL•FO
RenderX

even though it was reported in the study as construct. However, as the methods were so detailed, we were able to identify the correct measurement property used. The range of sample sizes across the 5 studies was large, from 6 to 124, and sample size calculations were not reported. In addition, the measurement approach of each study also ranged widely between using a passive-objective or active self-report approach.

Conclusions

Few studies have examined the accuracy, validity, or reliability of smartphones measuring and tracking PA levels in healthy young people. Of the 3 studies that measured validity against an objective measure of PA, such as the Actigraph accelerometer, all concluded that mobile phones were acceptable and valid tools. However, more research is needed that focuses on population characteristics, such as gender, different age groups, disability, and chronic diseases. Establishing the validity and reliability of smartphones to measure PA levels in young people will allow further research to investigate their use to increase PA in this population (ie, prevention strategies for developing chronic diseases or treating them), as well as identifying suitability for use in other populations, such as older adults and people currently living with chronic conditions.

Acknowledgments

This study was funded by the Paul Ramsay Foundation and the Australian National Health and Medical Research Council (fellowship to KC [APP1120641]). The funders of the study had no role in the study design, data collection, data analysis, or data interpretation.

Authors' Contributions

BP and LT led the development of the manuscript. KC, NN, TS, CC, BP, DRL, MS, and LT secured funding for the study. ABW led the literature search. BP, LT, LAG, BO, OG, CS, RV, KC, JW, ZB, and LB conducted the data extraction. CB assisted with the editing, interpretation, and write-up of the results and discussion. All authors contributed to the development of protocols for the study and reviewed, edited, and approved the final version of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Table S1 in Multimedia Appendix 1.
[DOCX File , 18 KB - mhealth_v10i10e39085_app1.docx]

References

1. Australia's health 2018. Australian Institute of Health and Welfare. URL: https://www.aihw.gov.au/reports/australias-health/australias-health-2018/contents/table-of-contents [accessed 2022-09-25]
2. Kohl HW, Craig CL, Lambert EV, Inoue S, Alkandari JR, Leetongin G, et al. The pandemic of physical inactivity: global action for public health. The Lancet 2012 Jul;380(9838):294-305. [doi: 10.1016/s0140-6736(12)60898-8]
3. Aubert S, Barnes JD, Abdeta C, Abi Nader P, Adeniyi AF, Aguilar-Farias N, et al. Global Matrix 3.0 physical activity report card grades for children and youth: results and analysis from 49 countries. J Phys Act Health 2018 Nov 01;15(S2):S251-S273. [doi: 10.1123/jpah.2018-0472] [Medline: 30475137]
4. Guthold R, Stevens GA, Riley LM, Bull FC. Global trends in insufficient physical activity among adolescents: a pooled analysis of 298 population-based surveys with 1·6 million participants. Lancet Child Adolesc Health 2020 Jan;4(1):23-35. [doi: 10.1016/s2352-4642(19)30323-2]
5. Patterson R, McNamara E, Tainio M, de Sá TH, Smith AD, Sharp SJ, et al. Sedentary behaviour and risk of all-cause, cardiovascular and cancer mortality, and incident type 2 diabetes: a systematic review and dose response meta-analysis. Eur J Epidemiol 2018 Sep 28;33(9):811-829 [FREE Full text] [doi: 10.1007/s10654-018-0380-1] [Medline: 29589226]
6. Boers E, Afzali MH, Newton N, Conrod P. Association of screen time and depression in adolescence. JAMA Pediatr 2019 Sep 01;173(9):853-859 [FREE Full text] [doi: 10.1001/jamapediatrics.2019.1759] [Medline: 31305878]
7. Champion KE, Chapman C, Gardner LA, Sunderland M, Newton NC, Smout S, Heath4Life team. Lifestyle risks for chronic disease among Australian adolescents: a cross-sectional survey. Med J Aust 2022 Feb 21;216(3):156-157. [doi: 10.5694/mja2.51333] [Medline: 34747039]
8. Ding D, Rogers K, van der Ploeg H, Stamatakis E, Bauman AE. Traditional and emerging lifestyle risk behaviors and all-cause mortality in middle-aged and older adults: evidence from a large population-based Australian cohort. PLoS Med 2015 Dec 8;12(12):e1001917 [FREE Full text] [doi: 10.1371/journal.pmed.1001917] [Medline: 26645683]
9. Laranjo L, Ding D, Heleno B, Kocaballi B, Quiroz JC, Tong HL, et al. Do smartphone applications and activity trackers increase physical activity in adults? Systematic review, meta-analysis and metaregression. Br J Sports Med 2021 Apr 21;55(8):422-432. [doi: 10.1136/bjsports-2020-102892] [Medline: 33355160]

XSL•FO
RenderX

10. Hallal PC, Andersen LB, Bull FC, Guthold R, Haskell W, Ekelund U. Global physical activity levels: surveillance progress, pitfalls, and prospects. The Lancet 2012 Jul;380(9838):247-257. [doi: 10.1016/s0140-6736(12)60646-1]
11. Bort-Roig J, Gilson ND, Puig-Ribera A, Contreras RS, Trost SG. Measuring and influencing physical activity with smartphone technology: a systematic review. Sports Med 2014 May 5;44(5):671-686. [doi: 10.1007/s40279-014-0142-5] [Medline: 24497157]
12. Thornton L, Osman B, Wescott AB, Sunderland M, Champion K, Green O, et al. Measurement properties of smartphone approaches to assess key lifestyle behaviours: protocol of a systematic review. Syst Rev 2020 Jun 03;9(1):127. [doi: 10.1186/s13643-020-01375-w]
13. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
14. Thornton L, Osman B, Champion K, Green O, Wescott AB, Gardner LA, et al. Measurement properties of smartphone approaches to assess diet, alcohol use, and tobacco use: systematic review. JMIR Mhealth Uhealth 2022 Feb 17;10(2):e27337 [FREE Full text] [doi: 10.2196/27337] [Medline: 35175212]
15. Teesson M, Champion KE, Newton NC, Kay-Lambkin F, Chapman C, Thornton L, Health4Life Team. Study protocol of the Health4Life initiative: a cluster randomised controlled trial of an eHealth school-based program targeting multiple lifestyle risk behaviours among young Australians. BMJ Open 2020 Jul 13;10(7):e035662 [FREE Full text] [doi: 10.1136/bmjopen-2019-035662] [Medline: 32665344]
16. Terwee CB, Prinsen CAC, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res 2018 May 17;27(5):1159-1170 [FREE Full text] [doi: 10.1007/s11136-018-1829-0] [Medline: 29550964]
17. World HO. Adolescent and young adult health. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/adolescents-health-risks-and-solutions [accessed 2022-09-16]
18. Arvidsson D, Fridolfsson J, Börjesson M. Measurement of physical activity in clinical practice using accelerometers. J Intern Med 2019 Aug 16;286(2):137-153 [FREE Full text] [doi: 10.1111/joim.12908] [Medline: 30993807]
19. Aadland E, Ylvisåker E. Reliability of the Actigraph GT3X+ accelerometer in adults under free-living conditions. PLoS One 2015 Aug 14;10(8):e0134606 [FREE Full text] [doi: 10.1371/journal.pone.0134606] [Medline: 26274586]
20. Kim Y, Lochbaum M. Comparison of polar active watch and waist- and wrist-worn ActiGraph accelerometers for measuring children's physical activity levels during unstructured afterschool programs. Int J Environ Res Public Health 2018 Oct 16;15(10):2268 [FREE Full text] [doi: 10.3390/ijerph15102268] [Medline: 30332785]
21. COSMIN definitions of domains, measurement properties, and aspects of measurement properties. COSMIN. URL: https://cosmin.nl/wp-content/uploads/COSMIN-definitions-domains-measurement-properties.pdf [accessed 2022-09-25]
22. Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. Qual Life Res 2018 May 19;27(5):1171-1179 [FREE Full text] [doi: 10.1007/s11136-017-1765-4] [Medline: 29260445]
23. Bruening M, van Woerden I, Todd M, Brennhofer S, Laska MN, Dunton G. AMobile ecological momentary assessment tool (devilSPARC) for nutrition and physical activity behaviors in college students: a validation study. J Med Internet Res 2016 Jul 27;18(7):e209 [FREE Full text] [doi: 10.2196/jmir.5969] [Medline: 27465701]
24. Dunton G, Liao Y, Intille SS, Spruijt-Metz D, Pentz M. Investigating children's physical activity and sedentary behavior using ecological momentary assessment with mobile phones. Obesity (Silver Spring) 2011 Jun;19(6):1205-1212 [FREE Full text] [doi: 10.1038/oby.2010.302] [Medline: 21164502]
25. Jongprasithporn M. A smartphone-based real-time simple activity recognition. 2017 Presented at: 3rd International Conference on Control, Automation and Robotics (ICCAR); 24-26 April; Nagoya, Japan p. 539-542. [doi: 10.1109/iccar.2017.7942755]
26. Schiel R, Kaps A, Bieber G. Electronic health technology for the assessment of physical activity and eating habits in children and adolescents with overweight and obesity IDA. Appetite 2012 Apr;58(2):432-437. [doi: 10.1016/j.appet.2011.11.021] [Medline: 22155072]
27. Schiel R, Kaps A, Bieber G, Kramer G, Seebach H, Hoffmeyer A. Identification of determinants for weight reduction in overweight and obese children and adolescents. J Telemed Telecare 2010 Aug 02;16(7):368-373. [doi: 10.1258/jtt.2010.091005] [Medline: 20679405]
28. World health organization newsroom fact sheets: obesity and overweight. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [accessed 2022-09-25]
29. Chow JJ, Thom JM, Wewege MA, Ward RE, Parmenter BJ. Accuracy of step count measured by physical activity monitors: the effect of gait speed and anatomical placement site. Gait Posture 2017 Sep;57:199-203. [doi: 10.1016/j.gaitpost.2017.06.012] [Medline: 28666177]

Abbreviations

EMA: ecological momentary assessment
PA: physical activity

XSL•FO
RenderX

PRISMA: preferred reporting items for systematic review and meta-analysis

Edited by L Buis; submitted 27.04.22; peer-reviewed by P Atorkey, M Swartz; comments to author 02.09.22; revised version received 16.09.22; accepted 16.09.22; published 21.10.22.

Please cite as:
Parmenter B, Burley C, Stewart C, Whife J, Champion K, Osman B, Newton N, Green O, Wescott AB, Gardner LA, Visontay R, Birrell L, Bryant Z, Chapman C, Lubans DR, Sunderland M, Slade T, Thornton L
Measurement Properties of Smartphone Approaches to Assess Physical Activity in Healthy Young People: Systematic Review
JMIR Mhealth Uhealth 2022;10(10):e39085
URL: https://mhealth.jmir.org/2022/10/e39085
doi:10.2196/39085
PMID:36269659

©Belinda Parmenter, Claire Burley, Courtney Stewart, Jesse Whife, Katrina Champion, Bridie Osman, Nicola Newton, Olivia Green, Annie B Wescott, Lauren A Gardner, Rachel Visontay, Louise Birrell, Zachary Bryant, Cath Chapman, David R Lubans, Matthew Sunderland, Tim Slade, Louise Thornton. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 21.10.2022. 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.

XSL•FO
RenderX

Review

Digital Biomarker–Based Studies: Scoping Review of Systematic Reviews

Hossein Motahari-Nezhad[1], MSc; Meriem Fgaier[2], MSc; Mohamed Mahdi Abid[3], MSc; Márta Péntek[4], MD, DSc; László Gulácsi[4,5], MD, DSc; Zsombor Zrubka[4,5], MBA, MD, PhD

[1]Doctoral School of Business and Management, Corvinus University of Budapest, Budapest, Hungary

[2]Doctoral School of Applied Informatics and Applied Mathematics, Óbuda University, Budapest, Hungary

[3]Research Center of Epidemiology and Statistics Sorbonne Paris Cité, Paris University, Paris, France

[4]Health Economics Research Center, University Research and Innovation Center, Óbuda University, Budapest, Hungary

[5]Corvinus Institute for Advanced Studies, Corvinus University of Budapest, Budapest, Hungary

Corresponding Author:
Hossein Motahari-Nezhad, MSc
Doctoral School of Business and Management, Corvinus University of Budapest
Fővám tér 8
Budapest, 1093
Hungary
Phone: 36 702097967
Email: h.motahari.lib@gmail.com

Abstract

Background: Sensors and digital devices have revolutionized the measurement, collection, and storage of behavioral and physiological data, leading to the new term *digital biomarkers*.

Objective: This study aimed to investigate the scope of clinical evidence covered by systematic reviews (SRs) of randomized controlled trials involving digital biomarkers.

Methods: This scoping review was organized using the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. With the search limited to English publications, full-text SRs of digital biomarkers included randomized controlled trials that involved a human population and reported changes in participants' health status. PubMed and the Cochrane Library were searched with time frames limited to 2019 and 2020. The World Health Organization's classification systems for diseases (International Classification of Diseases, Eleventh Revision), health interventions (International Classification of Health Interventions), and bodily functions (International Classification of Functioning, Disability, and Health [ICF]) were used to classify populations, interventions, and outcomes, respectively.

Results: A total of 31 SRs met the inclusion criteria. The majority of SRs studied patients with circulatory system diseases (19/31, 61%) and respiratory system diseases (9/31, 29%). Most of the prevalent interventions focused on physical activity behavior (16/31, 52%) and conversion of cardiac rhythm (4/31, 13%). Looking after one's health (physical activity; 15/31, 48%), walking (12/31, 39%), heart rhythm functions (8/31, 26%), and mortality (7/31, 23%) were the most commonly reported outcomes. In total, 16 physiological and behavioral data groups were identified using the ICF tool, such as looking after one's health (physical activity; 14/31, 45%), walking (11/31, 36%), heart rhythm (7/31, 23%), and weight maintenance functions (7/31, 23%). Various digital devices were also studied to collect these data in the included reviews, such as smart glasses, smartwatches, smart bracelets, smart shoes, and smart socks for measuring heart functions, gait pattern functions, and temperature. A substantial number (24/31, 77%) of digital biomarkers were used as interventions. Moreover, wearables (22/31, 71%) were the most common types of digital devices. Position sensors (21/31, 68%) and heart rate sensors and pulse rate sensors (12/31, 39%) were the most prevalent types of sensors used to acquire behavioral and physiological data in the SRs.

Conclusions: In recent years, the clinical evidence concerning digital biomarkers has been systematically reviewed in a wide range of study populations, interventions, digital devices, and sensor technologies, with the dominance of physical activity and cardiac monitors. We used the World Health Organization's ICF tool for classifying behavioral and physiological data, which seemed to be an applicable tool to categorize the broad scope of digital biomarkers identified in this review. To understand the clinical value of digital biomarkers, the strength and quality of the evidence on their health consequences need to be systematically evaluated.

XSL•FO
RenderX

(JMIR Mhealth Uhealth 2022;10(10):e35722) doi:10.2196/35722

KEYWORDS

scoping review; digital biomarkers; health; behavioral data; physiological data; digital health; remote monitoring; wearable; implantable; digestible; portable; sensor; digital health; mobile phone

Introduction

Background

In health care systems, the use of digital devices has become an accelerating trend [1], and their application was sped up by the COVID-19 pandemic [2]. The emergence of new sensor-based devices and wearables has revolutionized measuring, collecting, and storing clinical data, which has definite consequences for clinical decision-making [3]. A new notion, digital biomarkers, has emerged in medicine: "objective, quantifiable, physiological and behavioral measures collected using digital devices that are portable, wearable, implantable or digestible" [4]. In addition to their clinical value, digital biomarkers enable new health care value chains [5]. According to published reports, the global digital biomarkers market size was valued at >US $727 million in 2019 and is predicted to grow at a compound annual growth rate of 40% to reach approximately US $10.38 billion by 2027 [6].

Digital biomarkers are measured across multiple layers of the hardware (eg, sensors) and software of medical devices that capture signals (behavioral and physiological data) from patients [7]. Digital biomarkers can increase diagnostic and therapeutic precision in the modern health care system by remotely and continuously measuring reliable clinical data and allowing continuous monitoring and evaluation [8,9]. Captured by wearable, implantable, and digestible devices and sensors, digital biomarkers can be used at home to provide clinical data, collecting data that is not possible in the clinical setting [10]. This information can improve physicians' and patients' decisions, personalize the treatment, and predict diseases' current and future status [11]. Continuous evaluation allows personalized therapy [12]; for instance, continuous blood glucose monitoring by sensors in diabetes can be linked with patients' physical activity and food intake data, which can tailor insulin dose adjustments and generate predictive alerts for critically low blood glucose levels [13]. In addition, digital biomarkers play an essential role in the recognition of disease-related symptoms [14], are commonly used in clinical trials to evaluate different therapies [15], and offer better treatment, especially when combined with other interventions [16]. Overall, digital biomarkers play a significant role in precision medicine [17], can reduce clinical mistakes, improve the accuracy of diagnostic methods, and support personalized clinical decisions [18].

Several systematic reviews (SRs) have been published on digital biomarkers. However, most of them focused on a specific technology or disease area; for instance, studies reviewed the health impacts of wearable activity trackers on a general population [19] and in patients with Parkinson disease [20], to name a few that covered specific technologies in specific disease areas [21-23]. Scoping reviews aim to capture the main concepts of a research area and the available primary sources and categories of evidence in a formal, rigorous, and transparent manner [24]. Digital biomarkers cover various clinical areas such as Fitbit devices, activity trackers, and implantable cardiac defibrillators. The potential value of digital biomarkers in effective, technologically enhanced, safe, and user-centered care pathways [25,26] has been suggested by a plethora of published SRs on their clinical benefits in various clinical areas. However, no scoping review has covered all digital biomarkers in a single, complete study. An overview of the scope of clinical evidence can highlight clinical areas where the evidence supports the integration of digital biomarkers into health systems and areas with gaps in the evidence synthesis. Therefore, a scoping review of SRs on digital biomarkers may help readers grasp the breadth of the accumulated clinical evidence in the field. As the Cochrane Handbook states, reviews of reviews address the need for broad evidence synthesis by covering multiple interventions for the same condition as well as numerous reviews of the same intervention for different disease areas [27].

Objectives

Given the rapid accumulation of clinical evidence partly driven by the COVID-19 pandemic and the new European Medical Device Regulation that took effect in May 2021, this scoping review includes SRs published in 2019-2020 to determine in which clinical domains digital biomarkers and sensors were making progress before the new regulation took effect. Specifically, this scoping review aimed to explore the following:

1. The characteristics of SRs of digital biomarkers in terms of populations, interventions, and outcomes.
2. The characteristics of digital biomarkers in terms of behavioral and physiological data types, the digital devices and sensors used, and their role in the treatment pathway in the SRs.

The purpose of this scoping review was to categorize the building blocks of the research questions, not to synthesize or evaluate the quality of clinical evidence on digital biomarkers; this will be addressed in a separate SR of SRs of digital biomarker–based interventions, which will assess the methodological quality and quality of evidence for digital biomarkers in meta-analyses using A Measurement Tool to Assess Systematic Reviews-2 and Grading of Recommendations Assessment, Development, and Evaluation, respectively [28].

Methods

We followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [29].

Eligibility Criteria

According to the definition [4], digital biomarkers are behavioral and physiological data such as heart rate, physical activity, and

XSL•FO
RenderX

step counts collected using digital devices such as smartwatches [30]. Accordingly, in this study, we identified digital biomarkers as behavioral and physiological data that were measured using digital devices. Digital technologies that do not objectively quantify physiological or behavioral data were excluded from this study. To focus on the evidence relevant to clinical care, we included SRs that involved randomized controlled trials (RCTs), indicating that the review synthesized causal evidence concerning health outcomes [31]. Studies that did not report changes in participants' health status were excluded. In addition, SRs containing only observational studies were inappropriate for this study. Following the general definition of SRs [32], we included studies that used a systematic search strategy in electronic databases and had a predefined and clear research question, inclusion and exclusion criteria, screening, and data analysis and synthesis methods. Reviews lacking the critical appraisal of the included studies were considered SRs if other criteria were met [32]. We did not restrict our scoping review to a specific population. Human studies in any clinical setting and any age group or sex were eligible for this study. We considered all interventions that intentionally or unintentionally influence the health status of participants and involve the use of at least one digital biomarker for any purpose related to diagnosing patients, monitoring outcomes, or affecting the delivery of the therapeutic intervention or for prognostic purposes. We did not limit the scoping review to any specific type of comparator group. Full-text English-language SRs that considered any kind of health outcome (eg, change in the health status of individuals or a population due to an intervention) [31] were eligible for this scoping review.

Exclusion Criteria

Studies were excluded if (1) they were not SRs; (2) all included studies in the SR were not RCTs; (3) they were not human studies; (4) they did not use at least one digital biomarker to diagnose patients, monitor outcomes, or influence the delivery of the therapeutic intervention or for prognostic purposes; (5) they did not use at least a wearable, implantable, portable, or digestible device to measure behavioral or physiological data; (6) they did not report health outcomes (ie, they did not report a change in population health status due to the use of an intervention); (7) they were not published in full text written in English; and (8) they had not been published in 2019 or 2020.

Search Strategy

A comprehensive strategy for searching published SRs was established, including the following steps. First, the PubMed electronic database was searched using keywords related to the definition of digital biomarkers in the title or abstract, as well as applicable Medical Subject Headings terms [4], combined with the National Library of Medicine's filter for SRs [33]. Second, the Cochrane Library database of SRs was also searched using keywords related to digital biomarkers. The search was limited to studies published in 2019 or 2020. Finally, during the review process, an additional investigation was conducted into the reference lists of identified studies.

We used the following digital biomarker–related search terms: "digital biomarker" OR "digital biomarkers" OR "implantable" OR "implantables" OR "wearable" OR "wearables" OR "portable" OR "portables" OR "digestible" OR "digestibles" [28] (refer to Multimedia Appendix 1 for details of the search strategy).

Screening and Selection of Studies

Microsoft Excel was used to manage articles and remove duplicate references according to their digital object identifier numbers. Two independent reviewers selected the reviews in 2 phases as follows:

1. Titles and abstracts of retrieved records were screened to identify relevant studies based on the following 2 inclusion criteria: Is this an SR study (yes or uncertain, no)? and Is it a digital biomarker–based study (yes or uncertain, no)? The studies for which the answer to both questions was yes or uncertain were considered eligible for the next step.
2. The full texts of articles that met the criteria for title and abstract were assessed based on the following binary (yes or no) factors: whether the study was published in 2019 or 2020; whether it was written in English; whether it was a human study; whether the study included only RCTs; whether health outcomes were reported; and whether there was at least one digital biomarker in the study for diagnosing patients, monitoring outcomes, and influencing therapeutic intervention or use of a wearable, implantable, portable, or digestible device for prognostic purposes. Filtering the answers of all questions to *yes* identified all eligible studies. At each screening stage, disagreements were discussed between the two reviewers and resolved by consensus (HM-N, MMA, and MF). At each screening stage, the interrater agreement between the reviewers was calculated using the Cohen κ statistic using Microsoft Excel. The substantial agreement rate was considered to be $\kappa>0.6$ [34]. In case of low agreement ($\kappa<0.6$), the reviewers were retrained before entering the full-text phase.

Data Charting

Overview

Two review authors independently extracted data from the included reviews and discussed their findings to ensure consistency. All entries were cross-checked. We used charting data forms to extract data. Where possible, the data were copied and pasted directly from the text to avoid misinterpretation. Regarding the agreement rate, Cohen κ [34] was calculated using Microsoft Excel. In terms of countries, populations; interventions; outcomes; behavioral and physiological data; role of digital biomarkers; type of sensor technology; and descriptive statistics, including frequency and percentage, were calculated using Stata statistical software (version 16.0; StataCorp LLC) and Microsoft Excel. Regardless of country, type of digital device, and role of digital biomarkers, the total frequency and percentage of the other variables mentioned do not correspond to the total number of SRs included (31% and 100%, respectively) because an SR may have more than one category of these variables, as shown in the reported results. R statistical software (version 4.1.3; The R Foundation for Statistical Computing) was used to visualize the graphs. We did not assess the quality of the included reviews because this is not essential in scoping reviews [29].

XSL•FO
RenderX

The Characteristics of SRs

We used the classification systems developed by the World Health Organization (WHO) to characterize populations, interventions, and outcomes. The International Classification of Diseases, Eleventh Revision (ICD-11), is the latest disease classification system [35]. The International Classification of Health Interventions (ICHI) is a tool proposed for statistical classification, documenting, and analyzing of health interventions [36]. The ICHI encompasses diagnostic, medical, surgical, mental health, primary care, and allied health services; practical support; rehabilitation; traditional medicine; and public health initiatives offered by a wide range of professionals throughout the spectrum of health care systems. The International Classification of Functioning, Impairment, and Health (ICF) is a framework for assessing health and disability at the individual and population levels [37]. The populations studied in the included SRs were categorized using the ICD-11. In addition, the populations' age range (children, adults, older adults, other, or unspecified) and sex (male, female, or both male and female) were extracted. The interventions and outcomes were also grouped using the ICHI and ICF instruments, respectively. The ICF instrument was also used to classify physiological and behavioral data. In addition, the comparison types were collected from the included SRs. We also extracted the number of included RCTs in the SRs.

The Characteristics of Digital Biomarkers

The characteristics of digital biomarkers were recorded, including their role in the SR (intervention, measure of outcome, diagnostic tool, prognostic tool, or other), the type of physiological and behavioral data gathered by digital devices using the ICF tool [37], the type of digital device (implantable, portable, wearable, or digestible), and the type of applied sensor technology (biosensor, chemical sensor, flow sensor, fingerprint sensor, force sensor, heart rate sensor or pulse rate sensor, humidity sensor, hour monitor sensor, infrared sensor, image sensor, level sensor, muscle sensor, position sensor, pressure sensor, thermistor sensor, or temperature sensor) [38].

We identified sensors such as heart rate sensors and pulse rate sensors if they were involved in sensing cardiac rhythm and function (heart rate sensors) and blood pressure (pulse rate sensors). By contrast, position sensors were assigned to those reviews assessing physical activities, walking, running, or gait functions. In addition, sensors related to smoking behavior were grouped into flow sensors. The sensors that monitored body temperature were categorized as temperature sensors.

Evidence Synthesis

This scoping review used descriptive-analytical methods, including frequency, percentage, and data charting using Stata statistical software (version 16.0). The screening process was evaluated by calculating Cohen κ between the independent pairs of reviewers. The graphs were designed using R statistical software (version 4.1.3).

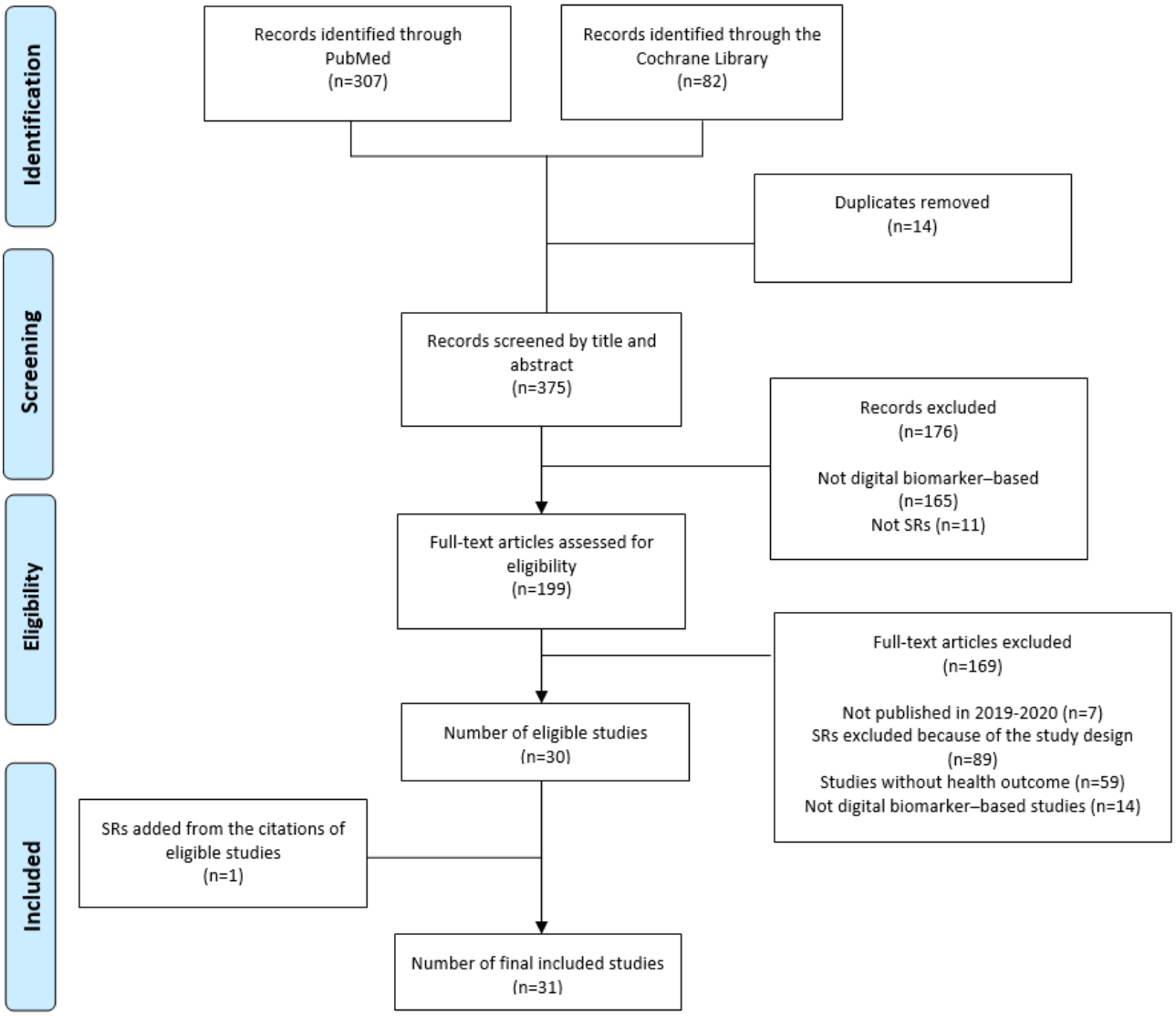
Results

Screening and Selection of Studies

From the computerized searches, 389 records were identified: 307 (78.9%) and 82 (21.1%) records in the PubMed and Cochrane Library databases, respectively. After removing duplicates, of the 389 records, 375 (96.4%) were screened for titles and abstracts. During title and abstract screening, there were 87 disagreements between the reviewers (Cohen κ=0.54). Therefore, the reviewers were retrained to reach a higher level of agreement. Consequently, they entered the discussion phase to resolve the discrepancies. In the screening phases of the titles and abstracts, 94% (82/87) of the disagreements were associated with *digital biomarker* and 6% (5/87) with *systematic review*. Of the 375 papers screened for titles and abstracts, 199 (53.1%) full-text papers were selected for the evaluation of eligibility. After resolving 42 disputes in the study selection phase (Cohen κ=0.76; n=17, 40%, disagreements on *health outcome*; n=14, 33%, disagreements on *digital biomarker–based studies*; n=9, 21%, disagreements about *RCTs*; and n=2, 5%, disagreements regarding *published in 2019-2020*), 44.7% (89/199) of the SRs were excluded at the full-text screening phase because of the study design (Multimedia Appendix 2). Of the 110 remaining SRs, 30 (27.3%) matched the inclusion criteria. After checking the reference lists of the qualifying SRs, one more record was included, bringing the total number of SRs that fit the inclusion criteria to 31. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart is shown in Figure 1. The characteristics of the studies are summarized in Multimedia Appendix 3 [39-69]. (Refer to Multimedia Appendix 4 for the list of excluded studies and the reasons for exclusion).

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram of the selection and screening process. SR: systematic review.



Characteristics of the Included SRs

The included SRs were published by authors from 14 different countries, most of them from Australia (6/31, 19%) [39-44], followed by Canada (4/31, 13%) [45-48]; the United States (4/31, 13%) [49-52]; the United Kingdom (3/31, 10%) [53-55]; Hong Kong (3/31, 10%) [56-58]; Taiwan (2/31, 7%) [59,60]; and Belgium (2/31, 7%) [61,62]. The remaining 7 studies were published by Chinese [63], French [64], Japanese [65], Portuguese [66], Italian [67], Dutch [68], and Danish [69] researchers.

Populations

Participants' disease areas covered 13 ICD-11 chapters (Table 1). The majority of SRs studied participants with circulatory system diseases (19/31, 61%) [40,41,45,46,49,51,53-55,57,60,62-69], followed by respiratory system diseases (9/31, 29%) [43,46,50,57,61-65]; endocrine, nutritional, or metabolic diseases (7/31, 23%) [40,45,50,57,62,64,65]; sleep-wake disorders (4/31, 13%) [50,57,63,64]; diseases of the nervous system (4/31, 13%) [40,57,63,64]; neoplasms (3/31, 10%) [40,52,64]; factors influencing health status or contact with health services (3/31, 10%) [40,56,60]; mental, behavioral, or neurodevelopmental disorders (2/31, 7%) [63,64]; diseases of the genitourinary system (2/31, 7%) [58,64]; and diseases of the musculoskeletal system or connective tissue (2/31, 7%) [40,64]. Moreover, the study by Lu et al [63] included patients with visual system diseases (1/31, 3%). Injury, poisoning, or certain other consequences of external causes as well as patients with skin diseases were the eligible included populations in another SR [64]. In 19% (6/31) of the studies, the included populations were only nonclinical and general participants [39,42,44,47,48,59]. In spite of comprising patients with the aforementioned specific clinical conditions, some SRs also included general populations without an applicable ICD-11 category, such as employees (3/31, 10%) [40,56,61]; students (3/31, 10%) [40,46,50]; healthy participants (2/31, 7%) [40,53]; and office workers (1/31, 3%) [40].

XSL•FO
RenderX

Table 1. Disease areas were identified using the International Classification of Diseases, Eleventh Revision, tool (N=31).

Populations	Values, n (%)
Diseases of the circulatory system	19 (61)
Diseases of the respiratory system	9 (29)
Endocrine, nutritional, or metabolic diseases	7 (23)
Diseases of the nervous system	4 (13)
Sleep-wake disorders	4 (13)
Neoplasms	3 (10)
Factors influencing health status	3 (10)
Mental, behavioral, or neurodevelopmental disorders	2 (6)
Diseases of the genitourinary system	2 (6)
Diseases of the musculoskeletal system	2 (6)
Diseases of the visual system	1 (3)
Diseases of the skin	1 (3)
Consequences of external causes	1 (3)

Interventions

According to the ICHI classification (Table 2), a high proportion of the interventions focused on physical activity behavior (16/31, 52%) [39-46,48,52,55-57,61,62,68] and conversion of cardiac rhythm (4/31, 13%) [49,53,66,67]. Percutaneous transluminal destruction of the arrhythmia circuit was covered by 7% (2/31) of the SRs [51,54]. Other SRs concerned assessment of weight maintenance functions (2/31, 7%) [50,59]; cardiac electrophysiological monitoring (1/31, 3%) [60]; assisting or leading exercise for functions of the cardiovascular system (1/31, 3%) [69]; assisting or leading exercise for functions related to pregnancy (1/31, 3%) [58]; blood pressure functions (1/31, 3%) [50]; noneconomic incentives to encourage improved physical activity (1/31, 3%) [65]; and economic incentives to encourage improved physical activity (1/31, 3%) [47]. The study by Jo et al [50] included 4 different types of interventions, but only 2 of these interventions could be categorized by the ICHI instrument: weight maintenance function and blood pressure function. The other 2 intervention types—blood cholesterol monitoring and wearable blood glucose monitoring systems—could not be categorized with the ICHI tool. Besides, 7% (2/31) of the SRs [63,64] did not assign an intervention category because they did not define a specific type of intervention; therefore, assignment to a specific intervention was not possible with the ICHI tool.

Table 2. Categorization of interventions using the International Classification of Health Interventions tool (N=31).

Interventions	Values, n (%)
Assessment of physical activity behaviors	16 (52)
Conversion of cardiac rhythm	4 (13)
Percutaneous transluminal destruction of arrhythmia circuit	2 (6)
Assessment of weight maintenance functions	2 (6)
Cardiac electrophysiological monitoring	1 (3)
Assisting or leading exercise for functions of the cardiovascular system	1 (3)
Assisting or leading exercise for functions related to pregnancy	1 (3)
Blood pressure function	1 (3)
Noneconomic incentives to encourage improved physical activity	1 (3)
Economic incentives to encourage improved physical activity	1 (3)

Outcomes

The reported outcomes fell into 13 unique categories of the ICF. Looking after one's health (physical activity; 15/31, 48%) [39-41,43-48,52,55-57,64,68]; walking (12/31, 39%) [39-43,46,52,56,57,61,62,68]; heart rhythm (8/31, 26%) [49,51,53,54,60,66,67,69]; demographic change (mortality; 7/31, 23%) [49,51,53,54,66,67,69]; and weight maintenance functions (7/31, 23%) [42,46-48,50,55,59] were the most commonly reported outcomes in the studies. Blood pressure functions (3/31, 10%) [50,55,64] and heart functions (3/31, 10%) [54,63,69] were the primary types of outcomes in 3 distinct SRs each. In addition, the following classifications were each assigned to one review: aerobic capacity (1/31, 3%) [41]; functions related to pregnancy (1/31, 3%) [58]; sleep functions

XSL•FO
RenderX

(1/31, 3%) [64]; and heart rate (1/31, 3%) [64]. Refer to Table 3 for further details.

Table 3. Categorization of outcomes using the International Classification of Functioning, Disability, and Health tool (N=31).

Outcomes	Values, n (%)
Looking after one's health	15 (48)
Walking	12 (39)
Heart rhythm	8 (26)
Demographic change (mortality)	7 (23)
Weight maintenance functions	7 (23)
Blood pressure functions	3 (10)
Heart functions	3 (10)
Hematological system functions	2 (6)
Exercise tolerance functions	2 (6)
Aerobic capacity	1 (3)
Functions related to pregnancy	1 (3)
Sleep functions	1 (3)
Heart rate	1 (3)

Characteristics of Digital Biomarkers

The behavioral and physiological data characteristics, digital devices, and sensors are summarized in Multimedia Appendix 5 [39-69].

Behavioral and Physiological Data and Digital Devices

Digital biomarkers were extracted from the included SRs. In total, 16 physiological and behavioral data groups were identified using the ICF tool, such as looking after one's health (physical activity; 14/31, 45%) [39,40,43-48,52,55-57,64,68]; walking (11/31, 36%) [39-43,52,56,57,61,62,68]; heart rhythm (7/31, 23%) [49,51,53,54,66,67,69]; and weight maintenance functions (7/31, 23%) [42,46-48,50,55,59]. The other identified data can be found in Multimedia Appendices 5 and 6. Besides, various digital devices were also used to collect these data when assessing other interventions; for example, an implantable cardiac defibrillator to gather heart function data [60]; a Fitbit device for capturing running activity [46]; Yorbody and AiperMotion for capturing physical activity [45]; and smart glasses, smartwatches, smart bracelets, smart shoes, and smart socks for capturing data related to heart function, gait pattern function, and temperature [63]. For more information, refer to Multimedia Appendix 5.

Role of Digital Biomarkers in Clinical Care

A substantial number of digital biomarkers were used as interventions in the SRs (24/31, 77%) [39-46,48-50,52,53,55-62,65-67]. By contrast, digital biomarkers were used to measure outcomes in 10% (3/31) of the studies [47,64,68]. In addition, in the review by Lu et al [63], digital biomarkers were used as intervention as well as outcome measurement and diagnostic tools. The remaining studies (3/31, 10%) [51,54,69] did not use digital biomarkers as intervention or diagnostic tools, as prognostic tools, or to measure outcomes; we categorized the role of digital biomarkers as *other*. In these studies, the included populations were patients with digital biomarkers (implantable cardiac defibrillators).

Types of Sensor Technologies

Wearables (22/31, 71%) were the most common types of digital devices [39-48,50,52,55,57-59,61-65,68], followed by implantable devices (8/31, 26%) [49,51,53,54,60,66,67,69]. The study by Liu et al [56] included both wearable and portable digital devices. Position sensors (21/31, 68%) [39-48,50,52,55-59,61-63,68,69] and heart rate sensors and pulse rate sensors (12/31, 39%) [49-51,53-55,60,63,64,66,67,69] were identified as the most prevalent types of sensors used to acquire behavioral and physiological data in the reviews. Flow (1/31, 3%) [64] and temperature (1/31, 3%) [63] sensors were used in 1 review each. Multimedia Appendix 5 shows the included studies and the role and types of digital devices, sensors, and physiological and behavioral data.

Discussion

Principal Findings

This scoping review of SRs of digital biomarkers published in 2019-2020 aimed to determine the scope of the literature in terms of populations, interventions, outcomes, technologies used, behavioral and physiological data, device types, and sensors. The search yielded 31 SRs that met the inclusion criteria and were published primarily by Australian, Canadian, American, and British researchers. The results showed that most of the populations studied were patients with circulatory, respiratory, endocrine, nutritional, or metabolic diseases. Intervention types were also predominantly used to assess physical activity behaviors and cardiac rhythm conversion. Wearables were the most common types of digital devices, mainly as interventions in the form of position and heart rate sensors.

XSL•FO
RenderX

There are numerous scoping reviews in this area, divided into 2 categories. First, scoping reviews focus on a specific type of digital device; for example, the study by Brognara et al [70] examined wearable sensors for assessing gait and postural alterations in patients with diabetes. Another scoping review highlighted the scope of wearable technologies in field hockey competitions [71]. Second, some scoping reviews considered only a specific behavioral or physiological data type; for example, the scoping review by Youn et al [21] examined digital biomarkers for neuromuscular disorders. Another study also reported the capabilities of artificial intelligence–aided digital biomarkers to aid in the early detection of dementia [72]. Therefore, we did not restrict the study to a particular digital device or behavioral or physiological data type to establish comprehensive results on digital biomarkers.

Populations, Interventions, and Outcomes

According to the findings, the populations, interventions, and outcomes studied in the SRs predominantly fall into 2 groups: physical activity and cardiovascular diseases. Although only 13 chapters of the ICD-11 and 10 categories of the ICHI were included in the SRs, because of the rapid pace of developments in the field and the fact that digital health and digital biomarkers are in transition [73,74], it is expected that the number of studies in other categories will increase. In addition, new devices, digital biomarkers, and sensors are expected to be introduced in health care systems and various disease areas because of the new advancements [75].

SRs are essential to evidence-based practice and health care decision-making [27,76]. According to the Cochrane Handbook, formulating a research topic based on population, intervention, comparison, and outcomes is one of the most significant requirements for SRs [27]. Although 61% (19/31) of the included SRs explicitly described a particular group with a clinical condition such as chronic obstructive pulmonary disease [43], others (6/31, 19%) included populations without clinical disorders [39,42,44,47,48,59]. By contrast, 19% (6/31) of the studies did not restrict their targeted population to a particular therapeutic area; rather, they encompassed diverse disease areas [46,50,57,62-64]. As evidenced by the findings, several reviews included general populations (patients with nonclinical conditions) to whom ICD-11 codes could not be assigned. As the use of wearable devices and sensors is spreading to broad populations such as students, employees, and office workers, and certain studies (RCTs) have included these populations, ICD-11 coding should include these populations. This issue also applies to interventions. Some (2/31, 6%) of the SRs did not clearly define the type of intervention in their study. Wearable health devices in health care settings, as well as pharmacologic and nonpharmacologic interventions, are not specific enough to be categorized using the ICHI tool. Therefore, researchers in this area are advised to consider this issue when formulating their research questions. In addition, 2 types of interventions (blood cholesterol monitoring and wearable blood glucose monitoring systems) could not be categorized with the ICHI tool, which should be considered by the developers of the tool.

Although some (11/31, 35%) of the studies evaluated the impact of a single kind of digital technology on the population, such as a Fitbit device [46], an implanted cardiac defibrillator [53], or a pedometer [61], others (20/31, 65%) included a variety of technologies [40,63-65]. In this context, the Cochrane Handbook suggests that reviews of reviews may be one way to address the need for breadth in evidence synthesis because they may combine multiple reviews of different interventions for the same condition or numerous reviews of the same intervention for different types of participants [27]. This is particularly true for the digital biomarker literature because this area encompasses a wide variety of populations and therapies, as demonstrated by this study's results.

Digital Devices, Physiological and Behavioral Data, and Sensors

Although numerous studies have shown that the accuracy of digital devices in measuring behavioral and physiological data may vary [77-80], most (20/31, 65%) of the included studies used different digital devices to synthesize qualitative or quantitative findings, which can be considered a gap in the literature on digital biomarkers; for example, the SR by Hannan et al [41] included various wearable digital devices (Garmin Forerunner, Fitbit Charge, My Wellness Key accelerometer, Yamax Digiwalker pedometer, Gex vital signs sensor, Nokia smartphone, and SenseWear Mini Armband) as interventions to quantitatively summarize the evidence for cardiac rehabilitation. By contrast, another study used only a Fitbit device to generate a meta-analysis for physical activity [46]. Despite the growing acceptance of wearables, the widespread adoption of wearables in clinical practice is still hampered by several barriers, including concerns about device accuracy and cost. To overcome these barriers, multiple stakeholders must collaborate in developing comprehensive assessment frameworks, clinical trials, and medical education programs. However, companies developing digital health technologies should consider the importance of evidence generation and validation for digital devices [7], considering that verification and validation of digital biomarkers require a multidisciplinary approach that includes engineering, data science, health information technology, and clinical research [8].

As shown in Multimedia Appendix 5, various digital devices are used to collect the same behavioral and physiological data; for example, an implantable cardiac defibrillator, iPhone-based rhythm monitoring device [60], and Cardio First Angel [81] to capture cardiac functions. Health economics research should evaluate the cost-effectiveness of each device in collecting behavioral and physiological data to determine the most cost-effective digital device for collecting specific data; for example, a study examined the potential cost-effectiveness of a wearable cardioverter defibrillator for patients with implantable cardiac defibrillator explant in a high-income Chinese city. It concluded that the cost-effectiveness of the wearable cardioverter defibrillator was highly dependent on the daily cost of the device in China [81]. Another study that examined the cost-effectiveness of portable devices for stroke diagnosis found no evidence of the cost-effectiveness under consideration for stroke diagnosis [82]. Another question that might arise from our results is the clinical effectiveness as well

XSL•FO
RenderX

as cost-effectiveness of different digital devices; for example, portable devices (tablet computers) and wearables (Jawbone UP24) were used to measure physical activity [56], or implantable cardiac monitors and portable devices (Holter electrocardiogram) were applied to record cardiac parameters [83]. Still, the question of comparative assessment of clinical effectiveness and cost-effectiveness between wearables and portables has remained unanswered. Digital health is undergoing rapid change, and new digital devices are being integrated into health care systems to facilitate it [84]. However, research shows that in medicine, more is not necessarily better [85], and rigorous evaluation of such therapies will become increasingly important in the future [84]. As a result, the cost-effectiveness of similar digital technologies could help clinicians and policy makers improve health care quality and reduce clinical costs.

We also used a simple search syntax derived from the definition of digital biomarkers. Still, the precision and specificity of this search formula to find all relevant studies have not been determined. Hence, the question regarding the development of a comprehensive and authoritative search formula with high precision and specificity has remained unanswered and is beyond the scope of this research. One of the significant challenges in digital health or biomarkers is the lack of a standard definition and mechanism for researchers to use when formulating a search syntax for reviews [82]. Standardizing definition terms is paramount to enhancing information retrieval and evidence synthesis. Accordingly, the quality of evidence synthesis of digital biomarkers may be compromised by publication bias, resulting in lower certainty of the evidence [83]. Digital biomarkers are an emerging field in flux and encompass various technologies. As we explained in the Methods section, digital biomarkers, by definition, are physiological and behavioral data collected using digital devices, including wearables as well as implantable, portable, and digestible devices. As can be inferred from the definition, there are 2 concepts to consider with digital biomarkers: behavioral and physiological data and digital devices that capture these data. However, no SR of portable digital devices that satisfied the inclusion requirements for this scoping study was discovered. This issue may result from the search formula obtained from the definition or a gap in the literature.

The varieties of digital biomarkers used in different health domains require specialized definitions, standards, and methodologies for achieving integration [84,85]. We used the ICF tool in this study to identify and classify behavioral and physiological data and outcomes in digital biomarkers and concluded that this tool has the potential to be used as a system for recognizing and categorizing behavioral and physiological data in the field. As there is no categorization scheme for digital biomarkers, we propose that researchers use this tool. This review also used other coding systems such as the ICD-11 and ICHI. The former allows systematic classification of the population, whereas the latter supports categorization of interventions. Using these systematized techniques, we can place digital biomarker research into a relevant population, intervention, comparison, and outcomes query. This may help to ensure the consistency and advancement of digital biomarker research [86] because the new Medical Device Regulation in the European Union [87] has increased the need for clinical evidence to support medical device approval; hence, the number of industry-sponsored SRs in this area is expected to increase in the future.

Implications

This scoping review examined SRs of digital biomarker–based studies regarding population, intervention, and outcomes. To our knowledge, this is the first scoping review of SRs of RCTs involving digital biomarkers. Therefore, these results may help clinicians and researchers to keep updated about the scope of the literature concerning digital biomarkers. In addition, we highlighted the behavioral and physiological data types as well as digital devices and sensors used in SRs of digital biomarkers. The aforementioned findings could also inform researchers about the field's gaps, as examined in the Digital Devices, Physiological and Behavioral Data, and Sensors subsection the Discussion section. In addition, as mentioned earlier, we have proposed that the ICF tool can be used by digital biomarker researchers as a standard tool for categorizing behavioral and physiological data.

Limitations

This study's findings should be considered in light of its limitations. First, we searched for studies on digital biomarkers using a mix of "digital biomarkers, wearable, implantable, portable, digestible" terms. We did not test the search strategy's precision and specificity in finding all field-related research, but we hypothesized that relevant publications might be found using this method. Nonetheless, this search formula may have missed some SRs relating to this topic. Second, the short time period (2019-2020) of the study is one of its possible weaknesses. Due to the broad scope of the topic, we chose a shorter time period. However, given the new European medical device legislation proposed in 2017 [87], we felt that this would be a critical time period for reviewing clinical data before regulation.

Moreover, we expected that because the studies in our analysis were SRs, we could incorporate all relevant studies if we included SRs. It is possible that non–English-language reviews were ignored because we limited the scope of this study to English-language research. Finally, the ICHI and ICF tools used in this study to categorize interventions and outcomes are not officially authorized (they are still being developed by the WHO) because there are no established definition systems for digital technologies. We did not examine the SRs for overlap among RCTs. Therefore, some results may overlap.

Strengths

To our knowledge, our scoping review is the most thorough presentation of SRs of digital biomarkers. Several scoping reviews have already been published on digital biomarkers, such as the use of accelerometers to measure physical activity [88], the use of wearable and mobile technology to measure and promote healthy sleep patterns in adolescents [89], the use of wearable inertial sensors in work-related activities [90], and the use of wearable sensor technology to detect shock impacts in sports and occupational settings [91], all of which relate to a specific type of digital device or population. By contrast, this

XSL•FO
RenderX

scoping review was conducted as comprehensive research to demonstrate the full spectrum of the topic and how digital biomarkers are already being integrated into health care systems. Our research described the scope of SRs of digital biomarkers without limitation to a specific type of patient or digital biomarker. Another strength of our study is that we used the WHO's classification tools (ICD-11, ICHI, and ICF) to identify and categorize the included studies' patients, interventions, and outcomes. In addition, physiological and behavioral data (digital biomarkers) were classified using the ICF tool, which was considered reliable for this purpose.

Conclusions

Our scoping review revealed that clinical evidence for a wide range of study populations, interventions, digital biomarkers, and sensor technologies has been systematically reviewed in recent years. Still, some clinical areas dominate, and notable unexplored fields exist. Understanding the clinical value of digital biomarkers requires a systematic assessment of the strength and quality of the evidence for their health effects. Understanding the breadth and quality of clinical evidence will inform clinical and health policy decision-makers about which areas are ripe for widespread adoption and evidence-based use of digital biomarkers and in which areas evidence gaps remain to be filled. Given the volume of literature on digital biomarkers across many health domains, specific definitions, standards, and methods for integration seem to be needed. We used the ICF tool to categorize behavioral and physiological data (digital biomarkers) in this study because there is no standard measurement in this area. The results suggest that this approach's categorization of behavioral and physiological data is applicable to digital biomarkers.

Acknowledgments

Project number TKP2020-NKA-02 has been implemented with the support provided from the National Research, Development, and Innovation Fund of Hungary, financed under the Tématerületi Kiválósági Program funding scheme. For conducting this research, HMN and MP received funding from project number TKP2020-NKA-02, implemented with the support provided from the National Research, Development, and Innovation Fund of Hungary, financed under the Tématerületi Kiválósági Program funding scheme. During this research, MP received funding from project number 2019-1.3.1-KK-2019-00007, implemented with the support provided from the National Research, Development, and Innovation Fund of Hungary, financed under the 2019-1.3.1-KK funding scheme. During this research, LG received funding from project number TKP2020-NKA-02, implemented with the support provided from the National Research, Development, and Innovation Fund of Hungary, financed under the Tématerületi Kiválósági Program funding scheme. During this research, ZZ received funding from project number TKP2020-NKA-02, implemented with the support provided from the National Research, Development, and Innovation Fund of Hungary, financed under the Tématerületi Kiválósági Program funding scheme; from the European Research Council under the European Union's Horizon 2020 research and innovation program (679681); and from project number 2019-1.3.1-KK-2019-00007, implemented with the support provided from the National Research, Development, and Innovation Fund of Hungary, financed under the 2019-1.3.1-KK funding scheme.

Authors' Contributions

HMN, LG, MP, and ZZ developed the concept. HMN wrote the first manuscript draft. MMA, HMN, and MF performed the screening and data extraction. All authors have commented on and approved the final manuscript. ZZ supervised the research.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 14 KB - mhealth_v10i10e35722_app1.docx]

Multimedia Appendix 2
Systematic reviews that were excluded because of the study design.
[DOCX File , 118 KB - mhealth_v10i10e35722_app2.docx]

Multimedia Appendix 3
Characteristic of the included studies.
[DOCX File , 26 KB - mhealth_v10i10e35722_app3.docx]

Multimedia Appendix 4
List of excluded studies with reasons for exclusion.
[DOCX File , 216 KB - mhealth_v10i10e35722_app4.docx]

XSL•FO
RenderX

Multimedia Appendix 5
Roles and types of digital biomarkers, sensors, and physiological and behavioral data, as well as digital devices.
[DOCX File , 62 KB - mhealth_v10i10e35722_app5.docx]

Multimedia Appendix 6
Physiological and behavioral data categories using the International Classification of Functioning, Disability, and Health tool.
[PNG File , 79 KB - mhealth_v10i10e35722_app6.png]

References

1. Silver L. Smartphone ownership is growing rapidly around the world, but not always equally. Pew Research Center. 2019 Feb 25. URL: https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/ [accessed 2021-02-28]
2. Crawford A, Serhal E. Digital health equity and COVID-19: the innovation curve cannot reinforce the social gradient of health. J Med Internet Res 2020 Jun 02;22(6):e19361 [FREE Full text] [doi: 10.2196/19361] [Medline: 32452816]
3. Kołodziej J, González-Vélez H. Survey High-Performance Modelling and Simulation for Big Data Applications: Selected Results of the COST Action IC1406 cHiPSet. Cham, Switzerland: Springer; 2019.
4. Babrak LM, Menetski J, Rebhan M, Nisato G, Zinggeler M, Brasier N, et al. Traditional and digital biomarkers: two worlds apart? Digit Biomark 2019 Aug 16;3(2):92-102 [FREE Full text] [doi: 10.1159/000502000] [Medline: 32095769]
5. Meister S, Deiters W, Becker S. Digital health and digital biomarkers – enabling value chains on health data. Curr Dir Biomed Eng 2016 Sep 30;2(1):577-581. [doi: 10.1515/cdbme-2016-0128]
6. Digital Biomarkers Market By System Component, By Therapeutic Area, By End-Users, Forecasts to 2027. Emergen Research. 2020 Nov. URL: https://www.emergenresearch.com/industry-report/digital-biomarkers-market [accessed 2021-10-25]
7. Coravos A, Khozin S, Mandl KD. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. NPJ Digit Med 2019;2(1):14 [FREE Full text] [doi: 10.1038/s41746-019-0090-4] [Medline: 30868107]
8. Lipsmeier F, Taylor KI, Kilchenmann T, Wolf D, Scotland A, Schjodt-Eriksen J, et al. Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. Mov Disord 2018 Aug;33(8):1287-1297 [FREE Full text] [doi: 10.1002/mds.27376] [Medline: 29701258]
9. Robb MA, McInnes PM, Califf RM. Biomarkers and surrogate endpoints: developing common terminology and definitions. JAMA 2016 Mar 15;315(11):1107-1108. [doi: 10.1001/jama.2016.2240] [Medline: 26978201]
10. Byrom B, Watson C, Doll H, Coons SJ, Eremenco S, Ballinger R, ePRO Consortium. Selection of and evidentiary considerations for wearable devices and their measurements for use in regulatory decision making: recommendations from the ePRO consortium. Value Health 2018 Jun;21(6):631-639 [FREE Full text] [doi: 10.1016/j.jval.2017.09.012] [Medline: 29909867]
11. Shin EK, Mahajan R, Akbilgic O, Shaban-Nejad A. Sociomarkers and biomarkers: predictive modeling in identifying pediatric asthma patients at risk of hospital revisits. NPJ Digit Med 2018 Oct 2;1:50 [FREE Full text] [doi: 10.1038/s41746-018-0056-y] [Medline: 31304329]
12. Coravos A, Khozin S, Mandl KD. Erratum: Author Correction: Developing and adopting safe and effective digital biomarkers to improve patient outcomes. NPJ Digit Med 2019 May 10;2:40 [FREE Full text] [doi: 10.1038/s41746-019-0119-8] [Medline: 31304386]
13. Vettoretti M, Cappon G, Facchinetti A, Sparacino G. Advanced diabetes management using artificial intelligence and continuous glucose monitoring sensors. Sensors (Basel) 2020 Jul 10;20(14):3870 [FREE Full text] [doi: 10.3390/s20143870] [Medline: 32664432]
14. Golini E, Rigamonti M, Iannello F, De Rosa C, Scavizzi F, Raspa M, et al. A non-invasive digital biomarker for the detection of rest disturbances in the SOD1G93A mouse model of ALS. Front Neurosci 2020 Sep 1;14:896 [FREE Full text] [doi: 10.3389/fnins.2020.00896] [Medline: 32982678]
15. Dorsey ER, Papapetropoulos S, Xiong M, Kieburtz K. The first frontier: digital biomarkers for neurodegenerative disorders. Digit Biomark 2017 Jul 4;1(1):6-13 [FREE Full text] [doi: 10.1159/000477383] [Medline: 32095743]
16. Guthrie NL, Carpenter J, Edwards KL, Appelbaum KJ, Dey S, Eisenberg DM, et al. Emergence of digital biomarkers to predict and modify treatment efficacy: machine learning study. BMJ Open 2019 Jul 23;9(7):e030710 [FREE Full text] [doi: 10.1136/bmjopen-2019-030710] [Medline: 31337662]
17. Jeong IC, Bychkov D, Searson PC. Wearable devices for precision medicine and health state monitoring. IEEE Trans Biomed Eng 2019 May;66(5):1242-1258. [doi: 10.1109/TBME.2018.2871638] [Medline: 31021744]
18. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA 2017 Oct 03;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
19. Shin G, Jarrahi MH, Fei Y, Karami A, Gafinowitz N, Byun A, et al. Wearable activity trackers, accuracy, adoption, acceptance and health impact: a systematic literature review. J Biomed Inform 2019 May;93:103153 [FREE Full text] [doi: 10.1016/j.jbi.2019.103153] [Medline: 30910623]

XSL•FO
RenderX

20. Corrà MF, Warmerdam E, Vila-Chã N, Maetzler W, Maia L. Wearable health technology to quantify the functional impact of peripheral neuropathy on mobility in Parkinson's disease: a systematic review. Sensors (Basel) 2020 Nov 19;20(22):6627 [FREE Full text] [doi: 10.3390/s20226627] [Medline: 33228056]
21. Youn BY, Ko Y, Moon S, Lee J, Ko SG, Kim JY. Digital biomarkers for neuromuscular disorders: a systematic scoping review. Diagnostics (Basel) 2021 Jul 15;11(7):1275 [FREE Full text] [doi: 10.3390/diagnostics11071275] [Medline: 34359358]
22. Wang C, Goel R, Noun M, Ghanta RK, Najafi B. Wearable sensor-based digital biomarker to estimate chest expansion during sit-to-stand transitions-a practical tool to improve sternal precautions in patients undergoing median sternotomy. IEEE Trans Neural Syst Rehabil Eng 2020 Jan;28(1):165-173 [FREE Full text] [doi: 10.1109/TNSRE.2019.2952076] [Medline: 31714229]
23. Alexander S, Peryer G, Gray E, Barkhof F, Chataway J. Wearable technologies to measure clinical outcomes in multiple sclerosis: a scoping review. Mult Scler 2021 Oct;27(11):1643-1656 [FREE Full text] [doi: 10.1177/1352458520946005] [Medline: 32749928]
24. Mays N, Roberts EP, Popay J. Chapter 12: Synthesising research evidence. In: Allen P, Black N, Clarke A, Fulop N, Anderson S, editors. Studying the Organisation and Delivery of Health Services: Research Methods. London, UK: Routledge; 2001.
25. Rajasekar V, Jayapaul P, Krishnamoorthi S, Saračević M. Secure remote user authentication scheme on health care, IoT and cloud applications: a multilayer systematic survey. Acta Polytech Hung 2021;18(3):87-106. [doi: 10.12700/aph.18.3.2021.3.5]
26. Chaczko Z, Klempous R, Rozenblit J, Adegbija T, Chiu C, Kluwak K, et al. Biomimetic middleware design principles for IoT infrastructures. Acta Polytech Hung 2020;17(5):135-150. [doi: 10.12700/aph.17.5.2020.5.7]
27. Becker LA, Oxman AD. Chapter 22: Overview of reviews. In: Higgins JP, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. Hoboken, NJ, USA: John Wiley & Sons; 2008:187-235.
28. Motahari-Nezhad H, Péntek M, Gulácsi L, Zrubka Z. Outcomes of digital biomarker-based interventions: protocol for a systematic review of systematic reviews. JMIR Res Protoc 2021 Nov 24;10(11):e28204 [FREE Full text] [doi: 10.2196/28204] [Medline: 34821568]
29. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
30. Nam KH, Kim DH, Choi BK, Han IH. Internet of things, digital biomarker, and artificial intelligence in spine: current and future perspectives. Neurospine 2019 Dec;16(4):705-711 [FREE Full text] [doi: 10.14245/ns.1938388.194] [Medline: 31905461]
31. Sedgley C. The responsibilities of being a physiotherapist. In: Porter SB, editor. Tidy's Physiotherapy. 15th edition. London, UK: Churchill Livingstone; 2013:1-21.
32. Krnic Martinic M, Pieper D, Glatt A, Puljak L. Definition of a systematic review used in overviews of systematic reviews, meta-epidemiological studies and textbooks. BMC Med Res Methodol 2019 Nov 04;19(1):203 [FREE Full text] [doi: 10.1186/s12874-019-0855-0] [Medline: 31684874]
33. Search Strategy Used to Create the PubMed Systematic Reviews Filter. National Library of Medicine, National Institutes of Health. 2018. URL: https://www.nlm.nih.gov/bsd/pubmed_subsets/sysreviews_strategy.html [accessed 2021-07-12]
34. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
35. International Classification of Diseases. 11th Revision. The global standard for diagnostic health information. World Health Organization. 2020. URL: https://icd.who.int/en [accessed 2020-09-12]
36. International Classification of Health Interventions (ICHI). World Health Organization. 2020. URL: https://www.who.int/standards/classifications/international-classification-of-health-interventions [accessed 2020-09-10]
37. ICF browser. World Health Organization. 2017. URL: https://apps.who.int/classifications/icfbrowser/ [accessed 2020-09-05]
38. Shinde M, Gupta J, Gawade S. Overview of different types of sensors used in eHealth environment. Int J Infinite Innov Technol 2013;2(2):1-5.
39. Brickwood KJ, Watson G, O'Brien J, Williams AD. Consumer-based wearable activity trackers increase physical activity participation: systematic review and meta-analysis. JMIR Mhealth Uhealth 2019 Apr 12;7(4):e11819 [FREE Full text] [doi: 10.2196/11819] [Medline: 30977740]
40. Lynch C, Bird S, Lythgo N, Selva-Raj I. Changing the physical activity behavior of adults with fitness trackers: a systematic review and meta-analysis. Am J Health Promot 2020 May;34(4):418-430. [doi: 10.1177/0890117119895204] [Medline: 31858812]
41. Hannan AL, Harders MP, Hing W, Climstein M, Coombes JS, Furness J. Impact of wearable physical activity monitoring devices with exercise prescription or advice in the maintenance phase of cardiac rehabilitation: systematic review and meta-analysis. BMC Sports Sci Med Rehabil 2019 Jul 30;11:14 [FREE Full text] [doi: 10.1186/s13102-019-0126-8] [Medline: 31384474]

XSL•FO
RenderX

42. Tang MS, Moore K, McGavigan A, Clark RA, Ganesan AN. Effectiveness of wearable trackers on physical activity in healthy adults: systematic review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2020 Jul 22;8(7):e15576 [FREE Full text] [doi: 10.2196/15576] [Medline: 32706685]
43. Burge AT, Cox NS, Abramson MJ, Holland AE. Interventions for promoting physical activity in people with chronic obstructive pulmonary disease (COPD). Cochrane Database Syst Rev 2020 Apr 16;4(4):CD012626 [FREE Full text] [doi: 10.1002/14651858.CD012626.pub2] [Medline: 32297320]
44. Freak-Poli RL, Cumpston M, Albarqouni L, Clemes SA, Peeters A. Workplace pedometer interventions for increasing physical activity. Cochrane Database Syst Rev 2020 Jul 21;7(7):CD009209 [FREE Full text] [doi: 10.1002/14651858.CD009209.pub3] [Medline: 32700325]
45. Kirk MA, Amiri M, Pirbaglou M, Ritvo P. Wearable technology and physical activity behavior change in adults with chronic cardiometabolic disease: a systematic review and meta-analysis. Am J Health Promot 2019 Jun;33(5):778-791. [doi: 10.1177/0890117118816278] [Medline: 30586996]
46. Ringeval M, Wagner G, Denford J, Paré G, Kitsiou S. Fitbit-based interventions for healthy lifestyle outcomes: systematic review and meta-analysis. J Med Internet Res 2020 Oct 12;22(10):e23954 [FREE Full text] [doi: 10.2196/23954] [Medline: 33044175]
47. Mitchell MS, Orstad SL, Biswas A, Oh PI, Jay M, Pakosh MT, et al. Financial incentives for physical activity in adults: systematic review and meta-analysis. Br J Sports Med 2020 Nov;54(21):1259-1268. [doi: 10.1136/bjsports-2019-100633] [Medline: 31092399]
48. Sypes EE, Newton G, Lewis ZH. Investigating the use of an electronic activity monitor system as a component of physical activity and weight-loss interventions in nonclinical populations: a systematic review. J Phys Act Health 2019 Apr 01;16(4):294-302. [doi: 10.1123/jpah.2017-0660] [Medline: 30849927]
49. Tseng AS, Kunze KL, Lee JZ, Amin M, Neville MR, Almader-Douglas D, et al. Efficacy of pharmacologic and cardiac implantable electronic device therapies in patients with heart failure and reduced ejection fraction: a systematic review and network meta-analysis. Circ Arrhythm Electrophysiol 2019 Jun;12(6):e006951. [doi: 10.1161/CIRCEP.118.006951] [Medline: 31159582]
50. Jo A, Coronel BD, Coakes CE, Mainous AG. Is there a benefit to patients using wearable devices such as Fitbit or health apps on mobiles? A systematic review. Am J Med 2019 Dec;132(12):1394-400.e1. [doi: 10.1016/j.amjmed.2019.06.018] [Medline: 31302077]
51. Kheiri B, Barbarawi M, Zayed Y, Hicks M, Osman M, Rashdan L, et al. Antiarrhythmic drugs or catheter ablation in the management of ventricular tachyarrhythmias in patients with implantable cardioverter-defibrillators: a systematic review and meta-analysis of randomized controlled trials. Circ Arrhythm Electrophysiol 2019 Nov;12(11):e007600. [doi: 10.1161/CIRCEP.119.007600] [Medline: 31698933]
52. Schaffer K, Panneerselvam N, Loh KP, Herrmann R, Kleckner IR, Dunne RF, et al. Systematic review of randomized controlled trials of exercise interventions using digital activity trackers in patients with cancer. J Natl Compr Canc Netw 2019 Jan;17(1):57-63 [FREE Full text] [doi: 10.6004/jnccn.2018.7082] [Medline: 30659130]
53. Alotaibi S, Hernandez-Montfort J, Ali OE, El-Chilali K, Perez BA. Remote monitoring of implantable cardiac devices in heart failure patients: a systematic review and meta-analysis of randomized controlled trials. Heart Fail Rev 2020 May;25(3):469-479. [doi: 10.1007/s10741-020-09923-1] [Medline: 32002732]
54. Martinez BK, Baker WL, Konopka A, Giannelli D, Coleman CI, Kluger J, et al. Systematic review and meta-analysis of catheter ablation of ventricular tachycardia in ischemic heart disease. Heart Rhythm 2020 Jan;17(1):e206-e219. [doi: 10.1016/j.hrthm.2019.04.024] [Medline: 31082362]
55. Hodkinson A, Kontopantelis E, Adeniji C, van Marwijk H, McMillan B, Bower P, et al. Accelerometer- and pedometer-based physical activity interventions among adults with cardiometabolic conditions: a systematic review and meta-analysis. JAMA Netw Open 2019 Oct 02;2(10):e1912895. Retracted in: JAMA Netw Open. December 1, 2020. doi: 10.1001/jamanetworkopen.2020.32700 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.12895] [Medline: 31596494]
56. Liu JY, Kor PP, Chan CP, Kwan RY, Sze-Ki D. The effectiveness of a wearable activity tracker (WAT)-based intervention to improve physical activity levels in sedentary older adults: a systematic review and meta-analysis. Arch Gerontol Geriatr 2020 Jul 26;91:104211. [doi: 10.1016/j.archger.2020.104211] [Medline: 32739713]
57. Kwan RY, Salihu D, Lee PH, Tse M, Cheung DS, Roopsawang I, et al. The effect of e-health interventions promoting physical activity in older people: a systematic review and meta-analysis. Eur Rev Aging Phys Act 2020 Apr 21;17(1):7 [FREE Full text] [doi: 10.1186/s11556-020-00239-5] [Medline: 32336996]
58. Chan KL, Chen M. Effects of social media and mobile health apps on pregnancy care: meta-analysis. JMIR Mhealth Uhealth 2019 Jan 30;7(1):e11836 [FREE Full text] [doi: 10.2196/11836] [Medline: 30698533]
59. Yen HY, Chiu HL. The effectiveness of wearable technologies as physical activity interventions in weight control: a systematic review and meta-analysis of randomized controlled trials. Obes Rev 2019 Oct;20(10):1485-1493. [doi: 10.1111/obr.12909] [Medline: 31342646]
60. Jang JP, Lin HT, Chen YJ, Hsieh MH, Huang YC. Role of remote monitoring in detection of atrial arrhythmia, stroke reduction, and use of anticoagulation therapy - a systematic review and meta-analysis. Circ J 2020 Oct 23;84(11):1922-1930 [FREE Full text] [doi: 10.1253/circj.CJ-20-0633] [Medline: 33012748]

XSL•FO
RenderX

61. Armstrong M, Winnard A, Chynkiamis N, Boyle S, Burtin C, Vogiatzis I. Use of pedometers as a tool to promote daily physical activity levels in patients with COPD: a systematic review and meta-analysis. Eur Respir Rev 2019 Dec 31;28(154):190039 [FREE Full text] [doi: 10.1183/16000617.0039-2019] [Medline: 31722891]
62. Franssen WM, Franssen GH, Spaas J, Solmi F, Eijnde BO. Can consumer wearable activity tracker-based interventions improve physical activity and cardiometabolic health in patients with chronic diseases? A systematic review and meta-analysis of randomised controlled trials. Int J Behav Nutr Phys Act 2020 May 11;17(1):57 [FREE Full text] [doi: 10.1186/s12966-020-00955-2] [Medline: 32393357]
63. Lu L, Zhang J, Xie Y, Gao F, Xu S, Wu X, et al. Wearable health devices in health care: narrative systematic review. JMIR Mhealth Uhealth 2020 Nov 09;8(11):e18907 [FREE Full text] [doi: 10.2196/18907] [Medline: 33164904]
64. Graña Possamai C, Ravaud P, Ghosn L, Tran VT. Use of wearable biometric monitoring devices to measure outcomes in randomized clinical trials: a methodological systematic review. BMC Med 2020 Nov 06;18(1):310 [FREE Full text] [doi: 10.1186/s12916-020-01773-w] [Medline: 33153462]
65. Kamei T, Kanamori T, Yamamoto Y, Edirippulige S. The use of wearable devices in chronic disease management to enhance adherence and improve telehealth outcomes: a systematic review and meta-analysis. J Telemed Telecare 2022 Jun;28(5):342-359. [doi: 10.1177/1357633X20937573] [Medline: 32819184]
66. Gama F, Ferreira J, Carmo J, Costa FM, Carvalho S, Carmo P, et al. Implantable cardioverter-defibrillators in trials of drug therapy for heart failure: a systematic review and meta-analysis. J Am Heart Assoc 2020 Apr 21;9(8):e015177 [FREE Full text] [doi: 10.1161/JAHA.119.015177] [Medline: 32290732]
67. Disertori M, Masè M, Rigoni M, Nollo G, Ravelli F. Declining clinical benefit of ICD in heart failure patients: temporal trend of mortality outcomes from randomized controlled trials. J Cardiol 2020 Feb;75(2):148-154 [FREE Full text] [doi: 10.1016/j.jjcc.2019.06.001] [Medline: 31526627]
68. Braakhuis HE, Berger MA, Bussmann JB. Effectiveness of healthcare interventions using objective feedback on physical activity: a systematic review and meta-analysis. J Rehabil Med 2019 Mar 13;51(3):151-159 [FREE Full text] [doi: 10.2340/16501977-2522] [Medline: 30843082]
69. Nielsen KM, Zwisler AD, Taylor RS, Svendsen JH, Lindschou J, Anderson L, et al. Exercise-based cardiac rehabilitation for adult patients with an implantable cardioverter defibrillator. Cochrane Database Syst Rev 2019 Feb 12;2(2):CD011828 [FREE Full text] [doi: 10.1002/14651858.CD011828.pub2] [Medline: 30746679]
70. Brognara L, Mazzotti A, Di Martino A, Faldini C, Cauli O. Wearable sensor for assessing gait and postural alterations in patients with diabetes: a scoping review. Medicina (Kaunas) 2021 Oct 22;57(11):1145 [FREE Full text] [doi: 10.3390/medicina57111145] [Medline: 34833363]
71. Lim JZ, Sim A, Kong PW. Wearable technologies in field hockey competitions: a scoping review. Sensors (Basel) 2021 Aug 03;21(15):5242 [FREE Full text] [doi: 10.3390/s21155242] [Medline: 34372481]
72. Li R, Wang X, Lawler K, Garg S, Bai Q, Alty J. Applications of artificial intelligence to aid early detection of dementia: a scoping review on current capabilities and future directions. J Biomed Inform 2022 Mar;127:104030. [doi: 10.1016/j.jbi.2022.104030] [Medline: 35183766]
73. Vasudevan S, Saha A, Tarver ME, Patel B. Digital biomarkers: convergence of digital health technologies and biomarkers. NPJ Digit Med 2022 Mar 25;5(1):36 [FREE Full text] [doi: 10.1038/s41746-022-00583-z] [Medline: 35338234]
74. Tarakji KG, Silva J, Chen LY, Turakhia MP, Perez M, Attia ZI, et al. Digital health and the care of the patient with arrhythmia: what every electrophysiologist needs to know. Circ Arrhythm Electrophysiol 2020 Nov;13(11):e007953. [doi: 10.1161/CIRCEP.120.007953] [Medline: 33021815]
75. Stern AD, Brönneke J, Debatin JF, Hagen J, Matthies H, Patel S, et al. Advancing digital health applications: priorities for innovation in real-world evidence generation. Lancet Digit Health 2022 Mar;4(3):e200-e206 [FREE Full text] [doi: 10.1016/S2589-7500(21)00292-2] [Medline: 35216754]
76. Stevens KR. Systematic reviews: the heart of evidence-based practice. AACN Clin Issues 2001 Nov;12(4):529-538. [doi: 10.1097/00044067-200111000-00009] [Medline: 11759425]
77. Piccinini F, Martinelli G, Carbonaro A. Accuracy of mobile applications versus wearable devices in long-term step measurements. Sensors (Basel) 2020 Nov 05;20(21):6293 [FREE Full text] [doi: 10.3390/s20216293] [Medline: 33167361]
78. Bayoumy K, Gaber M, Elshafeey A, Mhaimeed O, Dineen EH, Marvel FA, et al. Smart wearable devices in cardiovascular care: where we are and how to move forward. Nat Rev Cardiol 2021 Aug;18(8):581-599 [FREE Full text] [doi: 10.1038/s41569-021-00522-7] [Medline: 33664502]
79. Shei RJ, Holder IG, Oumsang AS, Paris BA, Paris HL. Wearable activity trackers-advanced technology or advanced marketing? Eur J Appl Physiol 2022 Sep;122(9):1975-1990 [FREE Full text] [doi: 10.1007/s00421-022-04951-1] [Medline: 35445837]
80. Germini F, Noronha N, Borg Debono V, Abraham Philip B, Pete D, Navarro T, et al. Accuracy and acceptability of wrist-wearable activity-tracking devices: systematic review of the literature. J Med Internet Res 2022 Jan 21;24(1):e30791 [FREE Full text] [doi: 10.2196/30791] [Medline: 35060915]
81. Wang SA, Su CP, Fan HY, Hou WH, Chen YC. Effects of real-time feedback on cardiopulmonary resuscitation quality on outcomes in adult patients with cardiac arrest: a systematic review and meta-analysis. Resuscitation 2020 Oct;155:82-90. [doi: 10.1016/j.resuscitation.2020.07.024] [Medline: 32755666]

XSL•FO
RenderX

82. Zrubka Z, Burrell A, Sharkawy MN, Pfeiffer CM, Janodia MD, Saylan M, et al. PP424 Piloting a comprehensive search for eHealth definitions in the grey literature: preliminary results from a systematic scoping review. Int J Technol Assess Health Care 2021 Dec 03;37(S1):36. [doi: 10.1017/s0266462321001562]
83. Faggion Jr CM, Wu YC, Tu YK, Wasiak J. Quality of search strategies reported in systematic reviews published in stereotactic radiosurgery. Br J Radiol 2016 Jun;89(1062):20150878 [FREE Full text] [doi: 10.1259/bjr.20150878] [Medline: 26986458]
84. Puleo V, Gentili A, Failla G, Melnyk A, Di Tanna G, Ricciardi W, et al. Digital health technologies: a systematic review of their cost-effectiveness. Eur J Public Health 2021 Oct 20;31(Supplement_3):2020-2021. [doi: 10.1093/eurpub/ckab164.273]
85. Naoum P, Pavi E, Athanasakis K. Economic evaluation of digital health interventions in palliative care: a systematic review of the literature. Front Digit Health 2021 Nov 3;3:730755 [FREE Full text] [doi: 10.3389/fdgth.2021.730755] [Medline: 34806076]
86. Burrell A, Zrubka Z, Redekop K, Kolasa K, Vinuesa L, Zah V, et al. PNS192 Defining the new age of digital health? Results of a systematic scoping literature review. Value in Health 2020 Dec 1;23(Supplement 2):S673. [doi: 10.1016/j.jval.2020.08.1636]
87. Thacker N, Singh P. REGULATION (EU) 2017/745 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EE. Official Journal of the European Union. 2017. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745 [accessed 2021-10-20]
88. Baldwin CE, Parry SM, Norton L, Williams J, Lewis LK. A scoping review of interventions using accelerometers to measure physical activity or sedentary behaviour during hospitalization. Clin Rehabil 2020 Sep;34(9):1157-1172. [doi: 10.1177/0269215520932965] [Medline: 32517508]
89. Beck AJ, Duffett-Leger L, Bright KS, Keys EM, Hayden A, Ward TM, et al. Using wearable and mobile technology to measure and promote healthy sleep behaviors in adolescents: a scoping review protocol. JBI Evid Synth 2021 May 10;19(10):2760-2769. [doi: 10.11124/JBIES-20-00293] [Medline: 34645774]
90. Lee R, James C, Edwards S, Skinner G, Young JL, Snodgrass SJ. Evidence for the effectiveness of feedback from wearable inertial sensors during work-related activities: a scoping review. Sensors (Basel) 2021 Sep 24;21(19):6377 [FREE Full text] [doi: 10.3390/s21196377] [Medline: 34640695]
91. Eitzen I, Renberg K, Færevik H. The use of wearable sensor technology to detect shock impacts in sports and occupational settings: a scoping review. Sensors (Basel) 2021 Jul 21;21(15):4962 [FREE Full text] [doi: 10.3390/s21154962] [Medline: 34372198]

Abbreviations

ICD-11: International Classification of Diseases, Eleventh Revision
ICF: International Classification of Functioning, Disability, and Health
ICHI: International Classification of Health Interventions
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RCT: randomized controlled trial
SR: systematic review
WHO: World Health Organization

Edited by L Buis; submitted 15.12.21; peer-reviewed by YH Kwan, P Altmann, J Gamble, K Matthias; comments to author 23.02.22; revised version received 20.04.22; accepted 26.07.22; published 24.10.22.

Please cite as:
Motahari-Nezhad H, Fgaier M, Mahdi Abid M, Péntek M, Gulácsi L, Zrubka Z
Digital Biomarker–Based Studies: Scoping Review of Systematic Reviews
JMIR Mhealth Uhealth 2022;10(10):e35722
URL: https://mhealth.jmir.org/2022/10/e35722
doi:10.2196/35722
PMID:36279171

©Hossein Motahari-Nezhad, Meriem Fgaier, Mohamed Mahdi Abid, Márta Péntek, László Gulácsi, Zsombor Zrubka. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 24.10.2022. 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

XSL•FO
RenderX

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.

XSL•FO
RenderX

Review

The Definitions of Health Apps and Medical Apps From the Perspective of Public Health and Law: Qualitative Analysis of an Interdisciplinary Literature Overview

Laura Maaß[1,2], BA, MA; Merle Freye[2,3], First State Examination; Chen-Chia Pan[2,4,5], BSc, MA; Hans-Henrik Dassow[2,6], BSc, MA; Jasmin Niess[2,7], MSc, PhD; Tina Jahnel[2,8], BA, MA, PhD

[1]Department of Health, Long-Term Care and Pensions, Research Center on Inequality and Social Policy, University of Bremen, Bremen, Germany

[2]Leibniz ScienceCampus Digital Public Health, Bremen, Germany

[3]Institute for Information, Health and Medical Law, University of Bremen, Bremen, Germany

[4]Unit of Lifestyle-Related Disorders, Department of Epidemiological Methods and Etiological Research, Leibniz Institute for Prevention and Epidemiology, Bremen, Germany

[5]Department of Prevention and Health Promotion, Institute for Public Health and Care Research, University of Bremen, Bremen, Germany

[6]Institute for Philosophy, University of Bremen, Bremen, Germany

[7]Human Computer Interaction Research Group, Institute of Computer Science, University of St. Gallen, St. Gallen, Switzerland

[8]Department of Health Services Research, Institute for Public Health and Care Research, University of Bremen, Bremen, Germany

Corresponding Author:
Laura Maaß, BA, MA
Department of Health, Long-Term Care and Pensions
Research Center on Inequality and Social Policy
University of Bremen
Unicom-Building, Room 3.4200
Mary-Somerville-Straße 3
Bremen, 28359
Germany
Phone: 49 421 218 58610
Email: laura.maass@uni-bremen.de

Abstract

Background: The terms *health app* and *medical app* are often used interchangeably but do not necessarily mean the same thing. To better understand these terms and better regulate such technologies, we need distinct definitions of health and medical apps.

Objective: This study aimed to provide an overview of the definitions of health and medical apps from an interdisciplinary perspective. We summarized the core elements of the identified definitions for their holistic understanding in the context of digital public health.

Methods: The legal frameworks for medical device regulation in the United States, the European Union, and Germany formed the basis of this study. We then searched 6 databases for articles defining health or medical apps from an interdisciplinary perspective. The narrative literature review was supported by a forward and backward snowball search for more original definitions of health and medical apps. A qualitative analysis was conducted on the identified relevant aspects and core elements of each definition. On the basis of these findings, we developed a holistic definition of health and medical apps and created a decision flowchart to highlight the differences between the 2 types.

Results: The legal framework showed that medical apps could be regulated as mobile medical devices, whereas there is no legal term for health apps. Our narrative literature review identified 204 peer-reviewed publications that offered a definition of health and medical apps. After screening for original definitions and applying the snowball method, 11.8% (24/204) of the publications were included in the qualitative analysis. Of these 24 publications, 22 (88%) provided an original definition of health apps and 11 (44%) described medical apps. The literature suggests that medical apps are a part of health apps. To describe health or medical apps, most definitions used the user group, a description of health, the device, the legal regulation, collected data, or technological functions. However, the regulation should not be a distinction criterion as it requires legal knowledge, which is neither suitable nor practical. An app's intended medical or health use enables a clear differentiation between health and medical apps. Ultimately, the health aim of an app and its main target group are the only distinction criteria.

XSL•FO
RenderX

Conclusions: Health apps are software programs on mobile devices that process health-related data on or for their users. They can be used by every health-conscious person to maintain, improve, or manage the health of an individual or the community. As an umbrella term, health apps include medical apps. Medical apps share the same technological functions and devices. Health professionals, patients, and family caregivers are the main user groups. Medical apps are intended for clinical and medical purposes and can be legally regulated as mobile medical devices.

(JMIR Mhealth Uhealth 2022;10(10):e37980) doi:10.2196/37980

KEYWORDS

mobile health; health app; medical app; digital health; regulation; mobile medical device; digital health applications; DiGA; digital care applications; DiPA; snowball search; mobile phone

Introduction

Background

The way people monitor their health has changed rapidly in the last decades owing to the smartphone and its widespread accessibility [1,2]. According to the World Health Organization (WHO), mobile health (mHealth) is a general term that covers public health and medical practice through the use of smartphones, sensors, personal digital assistants, wireless monitoring devices, or other wireless devices [3]. It can enable and improve the delivery of health care [4,5]. Mobile apps are defined as software programs that run on smartphones or tablet platforms [6]. Such apps can promote health and primary disease prevention [7,8]. At the same time, apps can support people with chronic illnesses in managing their medical conditions [8,9] or improve treatment adherence [10]. Furthermore, apps offer the opportunity to increase the autonomy of patients without necessarily needing to include physicians [11]. Not only may apps help in improving or monitoring one's health, but they can also play an important role in health economics as they can help in saving costs and increasing utility for health care systems [4,12-14].

The use of apps in health can have downsides. First, it is unclear whether and how mHealth contributes to a health care system with improved cost-effectiveness. Second, the evidence of a positive health impact through apps is not always available. Some studies suggest a lack of (long-term) evidence for health apps, which indicates a potential risk to the health of mHealth users [7,9,15]. These potential risks may cause unintended consequences for users, their social environment, and the overall health care system [16]. Although some studies have reported positive long-term effects on users' health [7-11,15], a clear directive on whether apps are an effective tool for diagnosing and treating health issues is yet to be established. Hence, there is a need to disentangle some of the conceptual unclarities in this domain and further advance research.

The potential of apps on the one hand and their risks on the other have drawn the attention of legislators as well. Several countries have started to regulate mobile apps in their function as medical devices—in 2021, the European Union (EU) Medical Device Regulation (MDR) started to uniformly regulate apps as medical devices in Europe [17]. The MDR is not an app-specific regulation as it addresses medical devices in general, including apps and other medical devices (eg, cardiac pacemakers, catheters, and filtering facepiece 2 masks). Apps can be affected by the areas of, for example, data protection, data security, consumer protection, medicine, and social law. App-specific laws, on the contrary, are rare [18,19]. In Germany, apps as medical devices have been officially made available for prescription since October 2020. This was introduced in the Digital Healthcare Act. The Digital Healthcare Act explicitly considers apps as *digital health applications* (DiGA), which, therefore, fall under app-specific law [20]. Despite all this, regulation—regardless of being app-specific or not—shares no common understanding of apps in the field of health. Although the law is highly dependent on accurate definitions, there is no legal definition of apps in the field of health. On the basis of a standard definition of apps in health, we argue that app developers will be nudged to use evidence-based approaches to develop apps in health. This ensures user safety and reduces the risk of harm [21].

The number of different stakeholders is mainly what causes the current unsatisfying situation. The diversity of stakeholders shows that apps in this field are not geared exclusively toward health care providers but also toward the general public and policy makers. When developing a new app, developers of mHealth apps need to consider patients, clinicians, families, researchers, politicians, providers, and payers alike [22]. Especially in health care systems with public funding, payers and users do not necessarily have to be the same person. In some countries (eg, Germany), health insurance companies have started to reimburse their clients for specific mHealth apps once prescribed by a physician [23].

Objectives

To better understand apps in health and aim at a more precise regulation of such technologies, exact definitions of these terms are needed. The goal of our literature review was to explore existing definitions of health and medical apps in the academic literature, specifically from a public health and law perspective. In this study, we focused on the differences and similarities in definitions and how health and medical apps relate to each other. We summarized the findings, resulting in more holistic definitions of health and medical apps generatively. Furthermore, we presented the differences between the definitions and terms used in research and legal regulation in the United States, the EU, and Germany. To date, existing frameworks have only focused on, for example, user groups [24-26] or technical functions [27-29]. Our aim was to follow a holistic approach that combines the user group, functions, and health aim of the apps. This approach provided us with a starting point that may also support researchers in building a more in-depth understanding of what constitutes health and medical

XSL•FO
RenderX

apps. Consequently, we hope that this will help progress the discourse on regulating such apps and provide a basis for further research in this domain.

Methods

To address our study's objective, we first conducted a backward and forward snowball search of the literature published in 6 different databases. Second, we analyzed the literature by inductively applying a thematic content analysis approach [30].

Inclusion Criteria and Search Strategy

We applied an inclusive, multistep approach to chart interdisciplinary research on health and medical apps. Owing to this paper's interdisciplinary nature, we searched for publications in the following academic databases that cover scientific literature from various disciplines: MEDLINE (PubMed), The Cochrane Library, Web of Science, Beck-Online, Juris, and Google Scholar. Through this approach, we were able to cover different disciplines (eg, psychology, human-computer interaction, epidemiology, ethics, public health, law, and health economics) that are involved in the field of mHealth. We included and excluded papers based on the criteria described in Textbox 1.

We formed an initial search strategy that we applied to all the databases. Our basic search syntax included terms and their synonyms, such as "health application," "medical application," and "definition" (Textbox 2).

We did not conduct a systematic literature review but rather a backward and forward snowball search to better connect different research fields and their understanding of health and medical apps (see the flowchart in Figure 1). The initial database search was conducted on January 21, 2021, and produced 60 papers that provided a definition of health apps or medical apps. All of these papers (60/60, 100%) were used to conduct the snowball search (see the following section for further explanation). Of these initially identified 60 publications, 20 (33%) provided an original definition and, thus, were included (n=6, 30% were identified through the database search and n=14, 70% were identified through snowballing).

Owing to the rapidly developing field of health and medical apps and the corresponding literature being published, we reran the search for updates on February 3, 2022. We identified 144 more publications that offered a definition of health and medical apps during a database search. However, only 2.8% (4/144) provided an original definition, and no other publications were identified during the snowball search. This resulted in a total of 24 publications that were included in our qualitative analysis. Of these 24 publications, 22 (92%) provided an original definition of health apps and 11 (46%) defined medical apps.

Textbox 1. Inclusion and exclusion criteria of the search.

Inclusion criteria

- Peer-reviewed publications
- Full text available in English, Chinese, or German (as these languages are spoken fluently by at least one author each)
- Full text focused on mobile health apps or mobile medical apps

Exclusion criteria

- No named concrete definitions or frameworks for apps in health
- Published in another language
- Evaluated the effectiveness of a single app without giving a description
- No available full text

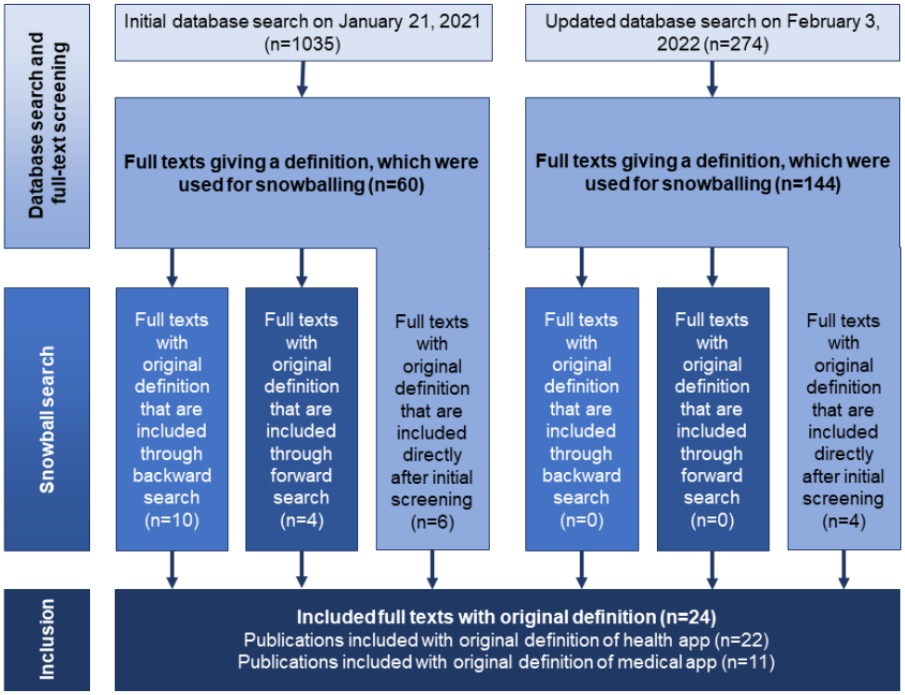
Textbox 2. Search terms in PubMed for the initial search.

Example search terms in PubMed

- Synonym health app or medical app: ("health application"[Title/Abstract] OR "health app"[Title/Abstract] OR "medical application"[Title/Abstract] OR "medical app"[Title/Abstract])
- AND
- Synonym definition: character*[Title/abstract] OR defin*[Title/abstract] OR concept*[Title/abstract] OR outlin*[Title/Abstract] OR mean*[Title/abstract] OR descri*[Title/abstract] OR terminology[Title/abstract] OR glossary[Title/abstract] OR framework[Title/abstract])

XSL•FO
RenderX

Figure 1. Flowchart of the search strategy.



Snowball Method: Backward Search

We formed a citation network using the snowball method after identifying the first set of potential publications from the narrative database search. The snowball search method allows for linking various scientific fields. This is essential as they all have different search terms, leading to overseen but relevant papers in systematic literature searches with a specific search string [31-33]. Our approach followed the guidelines for snowballing in systematic literature studies by Wohlin [32]. The already identified papers (identified through keyword search) served as seed articles (level 1). We screened the reference lists of these seed articles to identify other relevant publications (level 1). As a next step, relevant papers identified through the reference lists of level 1 publications were included (level 2). We stopped after level 2 screening as publications became less relevant to the seed publication the further we continued the search.

Snowball Method: Forward Search

To avoid the bias of citing only older literature (based on the literature search solely through reference lists), we used the platform Connected Papers [34]. This website connects databases from different fields to form holistic citation trees around a chosen paper. This way, Connected Papers allows for forward snowballing (ie, which other reports cited the selected article?) and backward snowballing (ie, which references did the chosen paper use?). Connected Papers displays all papers that cite the publication (level 1) and that cite level 1 publications (level 2) in a tree-like graphic. The multilevel multidisciplinary tree of connected references helped us identify our topic's core papers. We screened the publications shown in the tree and included them if they met our inclusion criteria.

Qualitative Analysis

One author collected all definitions in a table in preparation for the data synthesis. In total, 2 German authors independently translated German definitions into English following the translation by Kramer [35] that *Gesundheits-Apps* in German are health apps and *Medizin-Apps* in German are medical apps. A third author translated them from English back to German to check if the meaning was the same. Disagreements between the authors were resolved in a group discussion between all authors, where a final consensus was reached.

After collecting and translating the definitions from the retrieved literature, we conducted qualitative data analysis for evidence synthesis. Qualitative methods helped us better understand the rationale, perspectives, assumptions, statements, and approaches that researchers used to define health or medical apps [36]. In synthesizing the search results, we opted for an inductive bottom-up approach (conventional content analysis [37]). We developed categories for the qualitative content assessment from the material we screened. This was an iterative process, with new categories being added whenever indicated by the search results. We applied a thematic content analysis approach to the original and translated definitions to categorize the recurrent or common topics [30]. One author coded the definition themes using the software MAXQDA 2020 (VERBI GmbH).

Results

Overview

The basis for the qualitative content analysis was formed by a review of the legal landscape from the United States, the EU, and the particular case of Germany. The given nature of law is to form precise definitions in which every word has a concrete meaning. Although legal definitions need to be exact enough to be used, they also need to be abstract enough to suit individual

XSL•FO
RenderX

cases—such as medical apps. We applied the legal definitions of medical apps as mobile medical devices as an understanding lens to later analyze our scientific findings.

Legal Definitions of Medical Apps as Medical Devices

Overview

Apps are not entirely unregulated, but app-specific law is still rare [18,19]. For example, data protection law requires a legal basis for the data processing of an app [38], and the MDR sets high standards of quality and safety for software entering the market as a medical device [39]. Nevertheless, these laws do not use the term *app* and are not intended to regulate apps only. These laws use broader terms such as *data processing* or *medical device*, thus fitting more product categories than just apps. Consequently, apps can be affected by all these laws. Only the German regulation of DiGA and digital care applications (DiPA) is app-specific in the health context. The following sections provide an overview of the legal definitions of medical devices, mobile medical apps, DiGA, and DiPA and their thematic overlap.

Medical Devices

Both the United States and the EU regulate some apps through the control and supervision of *medical devices*. Although in the United States only the Food and Drug Administration (FDA) is responsible for regulating and approving, in the EU, several national conformity assessment bodies evaluate whether medical devices meet the standards of the EU MDR [40]. Although they are 2 different legal frameworks, the US and EU definitions of medical devices resemble each other—they concentrate on particular intended purposes. Related to the Federal Food, Drug, and Cosmetic (FD&C) Act, the US medical device is “an instrument, apparatus, implement, machine, contrivance, implant, in vitro reagent, or other similar or related article, including a component, part, or accessory, which is: [...] intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease, in man or other animals [...]” [41].

Similarly, the European MDR depicts medical devices as “any instrument, apparatus, appliance, software, implant, reagent, material or other article intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purposes [...]” [17].

Comparing these laws, the definition of medical devices focuses on their intended use for medical purposes.

Mobile Medical Apps

The United States and European countries have recognized various apps and their potential benefits and risks to public health. Consequently, they have published orientation guides and policies defining which apps are considered medical devices, creating the term *mobile medical apps*. According to the definition of the FDA, a mobile medical app is “a mobile app that incorporates device software functionality that meets the definition of device in section 201(h) of the FD&C Act 11; and either is intended to be used as an accessory to a regulated medical device; or to transform a mobile platform into a regulated medical device” [42].

This definition suggests that a mobile medical app is inevitably a medical device from a legal perspective and, thus, is intended to be used for the same medicinal purposes as a medical device. In parallel, the German Federal Institute for Drugs and Medical Devices orientation guide on medical apps also qualifies medical apps as medical devices, referring to particular intended purposes [43]:

> *Stand alone software like smartphone apps can indeed be classified as a medical device. In order for this to be the case, the software must be intended by the manufacturer to be used for humans and for at least one of the following purposes pursuant to Section 3 number 1 [Act on Medical Devices] MPG:*

- *diagnosis, prevention, monitoring, treatment or alleviation of disease,*
- *diagnosis, monitoring, treatment, alleviation or compensation of injuries or handicaps,*
- *investigation, replacement or modification of the anatomy or of a physiological process,*
- *control of conception.*

As this guidance was published before the current MDR entered into force, Article 2 (1) of the MDR [17] replaces the outdated purposes of the Act on Medical Devices. In summary, based on the legal definitions, mobile medical apps are medical devices and, therefore, are intended to be used for medical purposes. Thus, the key phrase is the intended use [44].

DiGA Definition

Closely related to medical apps are DiGA, which are officially available for prescription in Germany since October 2020. The law itself defines DiGA as “medical devices of lower risk class [intended to] support the detection, monitoring, treatment or alleviation of disease or the detection, treatment, alleviation of, or compensation for, an injury or disability of insured people” (own translation [45]).

Referring to medical devices, the definition of DiGA reveals that the intended use is again essential for their definition. Hence, DiGA are a small subset of mobile medical apps and medical devices. Spoken more abstractly, all DiGA are medical apps—and, therefore, medical devices—but not all medical devices and medical apps are DiGA. DiGA are predominantly regulated by the German Medical Device law and are only offered when they fulfill the Digital Health Applications Ordinance requirements regarding, for example, quality and safety, data protection, and data security.

DiPA Definition

Shortly after introducing DiGA, the German legislature regulated so-called DiPA in January 2022. According to the law, DiPA are “applications that are essentially based on digital technologies and are used by those in need of care or the interaction of those in need of care, relatives, and approved outpatient care facilities to reduce impairments to the independence or abilities of the person in need of care and to counteract a worsening of the need for care, insofar as the application is not to be provided by the health insurance or other responsible service providers due to illness or disability (digital care applications)” (own translation [46]).

XSL•FO
RenderX

Unlike DiGA, DiPA can but do not necessarily have to be medical devices. This affects the health aim of DiPA as well—DiGA need to be medical devices and, thus, have to be intended by the manufacturer to be used for medical purposes. However, DiPA do not need this health aim as they focus on care and the reduction of impairments or abilities of care recipients. According to the legislator, a DiPA can sometimes cover the same purposes as a DiGA [46], which implies that DiPA can be used for medical purposes. Through this statement, the legislator classifies care as a possible medical purpose.

Nevertheless, delimitation questions appear when an app addresses medical and care purposes. In addition, DiPA can include medical apps that cannot be reimbursed as DiGA. DiGA are limited to medical devices of a lower-risk class (according to the risk classes of the MDR), whereas DiPA can also include those of higher-risk classes. As the DiPA regulation was just passed at the time this paper was written, the distinction between DiPA and DiGA is not clear yet. It is presumed that it will be part of the former DiPA ordinance—the equivalent of the Digital Health Applications Ordinance.

Summary of the Legal Landscape

Legal norms and guidance only contain the terms *medical devices*, *mobile medical apps*, *DiGA*, and *DiPA*. They do not include the term *health app*. Consequently, there are only legal definitions of *medical devices*, *mobile medical apps*, *DiGA*, and *DiPA*. Except for DiPA, these definitions directly share one key characteristic: the intended use for medical purposes. As the legislator implies that caring is a medical purpose, DiPA indirectly share the intended use for medical purposes.

However, the question remains as to which aspects are regulated by the named laws. The regulation of medical devices, mobile medical apps, and DiGA is part of the Medical Device Law. Accordingly, the Federal Food, Drug, and Cosmetic Act and the MDR set high quality and safety standards for medical devices to meet everyday safety concerns. Medical devices need to fulfill special requirements to enter the market. For example, the MDR stipulates the supervision of notified bodies, conformity assessment procedures, clinical investigations, clinical evaluation, vigilance, market surveillance, transparency, and traceability [47].

Scientific Definitions

Overview

Our search produced a total of 22 papers providing a definition of health apps and 11 papers that defined medical apps in English or German (the total number of included papers was 24). No original definitions in Chinese were identified (the only definition coming from a Chinese publication [48] was published in English). We identified more papers providing a description. For example, Karakoyun et al [49] referred to the FDA definition of medical apps [42]. In contrast, Azad-Khaneghah et al [50] cited the definition of health apps by Morse et al [51], and Sun et al [52] used the description of health apps by Aitken and Lyle [53]. However, all these other papers referenced definitions from the papers included in our corpus. Hence, they were not included in our final data set.

The included original definitions are presented in Table 1. Where publications provided a definition of both health and medical apps, the definitions are displayed next to each other to make differences more visible.

XSL•FO
RenderX

Table 1. Overview of the included original definitions of health apps and medical apps sorted by publication[a].

Author, year, country	Definition of health app	Definition of medical app
Klasnja and Pratt [54], 2012, United States	"At the core of many mobile phone health applications is a single strategy: using the phone to track health-related behaviors, physiological states, symptoms, and other parameters relevant to health. [...] In addition to automatic detection of health-related behaviors, a number of phone-based health applications can connect, often wirelessly, to devices for measuring and uploading physiological data."	"Medical apps are used to diagnose, self-manage, monitor and treat conditions, provide decision-support, and collect health-related information."
Scherenberg and Kramer [55], 2013, Germany[b]	"Transferring the established health definition of the WHO from 1946, health apps can be described as mobile applications that aim to positively and sustainably influence physical, mental and social well-being on the basis of scientific evidence."	—[c]
Albrecht et al [21], 2014, Germany	"To better define the term 'health app,' we would like to suggest using the definition provided by the World Health Organization (WHO) in 1946 that defined health as 'a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity' (WHO, 1948). Apps that are in accordance with this definition of health—including apps that deal with wellness and fitness—can be summarized as 'health apps.' Health apps generally address healthy individuals who are simply interested in obtaining general information about their body and health status and want to keep fit and stay healthy."	"Apps dealing with the prevention of or aid with diagnostics and treatment of diseases as well as injuries could also be added to this category, but since they touch on areas typically covered by medical professionals, assigning the label 'medical app' seems more appropriate to underline the diagnostic and therapeutic aspects of such apps. Medical apps usually target health care professionals as well as patients that have already been diagnosed with a specific—often chronic—problem."
Aungst et al [56], 2014, United States	"Usually, apps designated for health and fitness are meant for daily use by patients, performing such functions as calorie and diet recorders, exercise assistants and patient diaries."	"[...] apps categorized as medical are usually designated for medical professionals or as supportive apps for patients with a medical condition. [...] What sets apart a medical app from others is that they can play a role as a clinical tool in medical practice. They are utilised by allied health professionals in medical education, at the point-of-care, through direct interaction with patients, and as clinical references."
Boudreaux et al [57], 2014, United States	"Phone application designed to promote health behavior or health maintenance [...], with the majority being related to diet and physical activity."	—
European Commission [58], 2014, Belgium	"Lifestyle and health apps are primarily apps that directly or indirectly promote healthy behaviors, quality of life, and well-being for individuals."	—
Seabrook et al [59], 2014, Canada	—	"Medical applications ('apps') for mobile devices such as smartphones and tablet computers provide health care professionals, patients, and the public with a growing number of specialized tools and resources."
Gehring et al [60], 2014, Germany[b]	"The term 'Health App' therefore refers to mobile applications that aim to positively and sustainably influence physical, mental, and social well-being on the basis of scientific findings. [...] In addition to apps that are intended to assist with relaxation (wellness), those that prescribe or accompany a physical work out (fitness) also belong to the category of 'health.'"	"Among the large group of 'health apps,' there are of course also apps that have their mission in the prevention, detection and treatment of diseases and injuries. These are genuine medical topics and a differentiation into 'medical apps' (apps with a medical purpose—derived from the Latin ars medicinae, 'medical art' or 'medicine' [...]) is appropriate here in order to emphasize the diagnostic and therapeutic aspect."
Powell et al [61], 2014, United States	"mHealth apps are mobile device applications intended to improve health outcomes, deliver health care services, or enable health research. [...] Because apps can be used to inexpensively promote wellness and manage chronic diseases, their appeal has increased with health reform and the increasing focus on value."	—
Aitken and Lyle [53], 2015, United States	"The availability of consumer apps continues to grow, particularly in the area of health care apps. Commonly referred to as mHealth apps, these apps assist consumers in self-management of overall wellness, disease prevention and disease management."	—

XSL•FO
RenderX

Author, year, country	Definition of health app	Definition of medical app
Lucht et al [62], 2015, Germany[b]	"Health apps aim at maintaining fitness and supporting a health-promoting lifestyle. These apps are offered in the 'Health and Fitness' category."	"In general, medical apps are apps for health care professionals that support their everyday work, as well as apps for patients to better manage mostly chronic diseases. These apps are mainly found in the 'Medicine' category of the two major app stores (Google Play, Apple iTunes)."
Albrecht and von Jan [63], 2016, Germany[b]	"Health apps can be defined as apps that provide users with functionalities for the areas of health, medicine, healing or wellness and thus, in a broader sense, transfer the WHO definition of health (WHO 1948) to the app area."	—
Kramer [35], 2017, Germany[b]	"Health apps address healthy users who want to use the app to support themselves in a health-promoting lifestyle, and who want to strengthen their resources—for relaxation, for a better understanding of health or illness, for a healthy, active lifestyle—with the help of an app."	"Medical apps address patients or their family members who are looking for support to better manage their daily lives, e.g., with a chronic disease. These apps are designed to strengthen the 'self-empowerment' of those affected to manage their disease. Medical apps also include apps that support health care professionals (physicians, nurses, therapists) in their daily practice or clinical routine, e.g., with reference works, dosage calculators, and medical decision-making aids. In the English-speaking world, medical apps are also referred to as medical apps."
Morse et al [51], 2018, Malaysia	"Mobile health (mHealth) applications (apps) are software that are incorporated into smartphones to improve health outcome, health research, and health care services."	—
Evers-Wölk et al [64], 2018, Germany[b]	"Basically, health apps are application programs for mobile devices, in particular smartphones and tablets, whose aim is to have a positive effect on the user's health."	"Medical apps that address either medical professionals or must be classified as medical devices due to their medical indication. Applications that are specifically intended to detect, cure or alleviate diseases, illnesses and physical injuries are classified as medicine and medical science."
Gregor-Haack [65], 2018, Germany[b]	"Health apps are applications for citizens and patients whose primary goal is health or health promotion. According to the WHO definition of health, health apps are mobile applications that 'positively and sustainably influence physical, mental, and social well-being on the basis of scientific findings.'"	"Medical apps are a smaller group of health apps that are specifically related to medicine. These include applications for service providers to support everyday professional life as well as applications for patients for the self-management of mostly chronic diseases."
Groß and Schmidt [66], 2018, Germany[b]	"A distinction must be made here between health apps, which are primarily aimed at medical laypersons interested in health, and medical apps, which are aimed at representatives of the health care professions. Health apps are much more common, mostly available free of charge in app stores and not subject to regular quality control. They range from pedometers, nutrition and weight control tools, fitness-related applications, the collection of (disease-related) user measurement data, to medication management."	"A distinction must be made here between health apps, which address primarily laypersons interested in health, and medical apps, which address representatives of the health care professions. Medical apps focus on medical and nursing functions and often relate to the areas of diagnostics and therapy; they are regulated by the Medical Devices Act. Examples include medical reference works, calculators (e.g. for drug dosages) or the presentation of medical documents and images."
Moshi et al [67], 2018, Australia	—	"One form of mHealth is mobile medical applications (MMAs) also known as 'apps.' These are a type of software available for mobile platforms (e.g., smartphone, tablet, smartwatch). In a medical context, MMAs may be used by patients to self-manage and/or screen medical conditions, rather than presenting at hospitals or clinics for additional appointments. MMAs may also allow for medical practitioners and/or allied health workers to remotely monitor, screen and manage their patients."

XSL•FO
RenderX

Author, year, country	Definition of health app	Definition of medical app
Heretleif et al [68], 2021, Germany[b]	"The delimitation is based on the consideration of three central perspectives: the user perspective, the technological perspective and the regulatory perspective. Health apps are primarily aimed at health-conscious users and offer support in the areas of prevention, education and health promotion. They are generally not aimed at healthcare professionals. From a technological perspective, their focus is on the collection, recording, processing and visualization of users' health-related data. A not insignificant proportion of apps focus purely on imparting knowledge. From a regulatory perspective, it is worth noting that, unlike medical apps or DiGA, health apps are not generally subject to any specific, binding regulation."	—
Wang and Qi [48], 2021, China	"Mobile health applications, the principal manifestation of mobile healthcare, refer to health applications based on mobile terminal systems such as Android and iOS that provide services such as medical information inquiry and symptom self-examination. Mobile health applications allow users not only to seek answers to health problems but also to gain access to healthcare, exercise and fitness, health management, and other related services anytime and anywhere. Mobile health applications alleviate the shortage of health information resources to a certain extent, provide a convenient way for users to obtain health information and services, and play an important role in spreading health knowledge and meeting the users' need for health consultation."	—
Volpi et al [69], 2021, Brazil	"Mobile health apps (mHealth) offer a way to monitor patient's health conditions, such as diet, body weight, blood pressure, mood, and sleep, among others, and can be used in combination with traditional health care to facilitate access to health information [...]. Thus, mHealth apps might increase awareness of needed behavioral changes and the adherence to healthy habits, along with the health care provider's awareness of what the patient is doing [...]. Moreover, mHealth apps can guide illness self-management, providing patients with psychological support and decision-making support, and facilitating collaboration between health professionals, patients, and their families."	—
Golden et al [70], 2021, United States	"In recent years, mobile health (mHealth) apps on smartphones have become ubiquitous tools for personal health management and behavior tracking. mHealth apps can provide individuals with continuous feedback on health status and progress, push notification reminders, and other useful engagement features."	—
Racioppi et al [71], 2021, United States	"The Health apps specifically have the potential to enhance patient/provider communication and assessment through active and passive evaluation and tracking. Smartphone mobile applications are able to record self-reported patient outcomes, whereas activity trackers such as the Apple Watch and Fitbit are able to collect real-time physiological data such as heart rate and step counts."	—
Tobias and Spanier [72], 2021, Israel	"Health apps can monitor health conditions and alert the patient or attending physician about deterioration."	—

[a]Of the 22 definitions of health apps, 9 (41%) were in German and translated into English. A total of 45% (10/22) of the publications came from Germany, and 32% (7/22) came from the United States. Other definitions were identified from Asia (3/22, 14%) [50,53,54] or South America (1/22, 5%) [55]. For medical apps, of the 11 definitions, 6 (55%) were published in German. As for health apps, most descriptions for medical apps came from Germany (7/11, 64%) or the United States (2/11, 18%), only 9% (1/11) came from Australia [71], and 9% (1/11) came from Canada [72].

[b]German publications translated by the authors (MF, HHD, and TJ) into English.

[c]No definition given.

XSL•FO
RenderX

Quality Analysis of Found Definitions From Scientific Papers

We analyzed the identified definitions using MAXQDA to form word clouds of the terms used for health apps and medical apps. Multimedia Appendix 1 provides a brief overview of the differences in wording of health and medical apps (for word cloud data, see also Multimedia Appendix 1). The following sections provide a more detailed analysis of the definitions' topics.

Health Apps

Our qualitative analysis identified the 7 most common topics (user, health aim, definition of health, device, regulation, data, and technological functions) addressed by most publications for health apps. For health apps, the 2 dominant topics were the health aim of health apps and the used or produced data. Only a few definitions (2/22, 9%) addressed the missing regulation and the functions of the apps (Multimedia Appendix 2 [21,35,48,51,53-58,60-66,68-72]).

Except for 9% (2/22) of the publications [54,72], every definition addressed the aim of health apps. A total of 45% (10/22) of the papers [35,51,56-58,62-65,68] named *health promotion* as a goal of health apps. Other frequently given aims included *improving the user's fitness* [21,56,60,62,66], *wellness* [21,35,53,60,61,63], and *mental and social well-being* [55,60,65]. Less frequently stated were the health aims *disease prevention* and *disease management* [53,69]; *illness or disease self-management* [53]; *behavior tracking* [70]; *access to health information*; *psychological suppor*t; and *decision-making support* to *facilitate collaboration* between the patient, their family, and their physician [69]. More generally, Tobias and Spanier [72] spoke about *health conditions*.

The user group was described in 64% (14/22) of the papers. Although 21% (3/14) of these papers mentioned *users* in general [48,63,64], 29% (4/14) limited this group to *health-conscious users* [68], (*healthy) users who are interested in health* [21,35], or *medical laypersons interested in health* [66]. Similarly, mentioned user groups of health apps were *citizens* [65] and *individuals* [58,63,70], *patients* [56,65,69,71,72], or the *family* as a potential user group [69]. Only 29% (4/14) of the definitions named professionals (ie, *non-healthcare professionals* [68], *health professionals* [69], or *physicians* [72]) as users of health apps.

A total of 64% (14/22) of the publications included data collection in their definition of health apps. However, the types of data collected were heterogeneous. In total, 36% (5/14) of the papers stated that health apps collect data on *health behavior* [35,54,57,62,70]. Other publications reported data on *physical activity* (4/14, 29%) [35,56,57,60] or *dietary data* (4/14, 29%) [56,57,66,69]. Some definitions restrained themselves from naming concrete fields and were phrased more generically as collecting *general information about their body* (1/14, 7%) [21], *other parameters related to health* (1/14, 7%) [54], *health outcomes* (1/14, 7%) [51], *health conditions* (2/14, 14%) [69,72], *physiological data* (1/14, 7%) [71], or *health-related data* (1/14, 7%) [68].

Concerning devices, of the 12 papers that defined health apps, 6 (50%) described health apps as *mobile applications* in their definitions [51,55,60,65,66,71]. Some publications described health apps as *mobile phone* (2/12, 17%) [54,57]; *smartphone* (3/12, 25%) [64,70,71]; or, more generally, as *mobile device based* (3/12, 25%) [54,61,64]. Only 8% (1/12) of the papers included *tablets* as devices for health apps [64].

In total, 41% (9/22) of the papers described the technical functions of health apps. These included *tracking* [54,66,70,71], *data collection* [66,68,71], *recording* [56,68,71], *exercise assistance* [56,60], *monitoring* [69,72], *detection*, *connecting*, *measuring*, and *uploading* [54]. A total of 5% (1/22) of the papers each added the role of a *diary* [56], *prescribing tools* [60], *providing continuous feedback* [70], *alerting* [72], or more general *tools for processing and visualizing data* [68].

In total, 18% (4/22) of the papers (all from German authors) included a *definition of health* in their description of health apps [21,55,63,65]. All of them (4/4, 100%) referred to the definition given by the WHO in 1948, which framed health as "a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity" [73]. The regulation of health apps was addressed by only 9% (2/22) of the definitions [66,68]. Both publications stated that health apps are not *subject to regular quality control* [66] or *not generally subject to any specific, binding regulation* [68].

Medical Apps

In the definitions of medical apps, the common topics were similar to those of health apps. In contrast to health apps, no publication on medical apps defined medicine. Therefore, only 6 recurring issues in the descriptions of medical apps were identified. The reports on medical apps most often stated the health aim of the app and the user groups. Only 9% (1/11) of the definitions specified the data and function of medical apps (Multimedia Appendix 2).

In total, 91% (10/11) of the definitions described that the health aim of medical apps is to serve as *clinical tools for detection, diagnosis, monitoring, and treatment of diseases or injuries* [21,35,54,56,60,62,64,66,67] and *decision support for medicine and nursing* [21,35,54]. A total of 36% (4/11) of the definitions also stated that medical apps serve as *self-management tools for chronic diseases and empowerment* of patients [35,54,62,65,67]. In total, 18% (2/11) of the definitions mentioned the *preventional purpose* of medical apps [21,60].

Of the 11 definitions, 9 (82%) addressed users of medical apps. Most of the definitions commonly indicated that the users of medical apps are *medical professionals* (3/11, 27%) [21,64,67], *medical service providers* (1/11, 9%) [65], *health care professionals* (7/11, 64%) [21,35,56,59,62,66,67], and *patients* (7/11, 64%) [21,35,56,59,62,65,67]. A total of 9% (1/11) of the definitions included the general public in the user group [59]. Another definition specifically pointed out that patients' *family members* (eg, in their function as guardians or caretakers) are also part of the medical app user group [35].

Only 27% (3/11) of the papers defined medical apps as *mobile applications* or *software* that operate on *mobile devices*, including *smartphones* and *tablets* [59,66,67], or *smartwatches*

XSL•FO
RenderX

[67]. Equally less often addressed was the data criterion. Only 9% (1/11) of the definitions stated that medical apps collect *health-related information* [54], whereas no other publication described the data or technological functions of medical apps. In total, 18% (2/11) of the papers stated that medical apps are *regulated as medical devices* in Germany and the United States [64,66]. We described the legal regulation of medical apps as medical devices in detail in the previous section.

Discussion

Many of the definitions identified through the literature search addressed the criteria user group, a specific understanding of the term *health*, a concrete device, the fragmentary regulation of medical apps (or the lack of regulation for most other types of health apps), data collected, or technological functions to gather or display content within the app. A definition of health and medical apps should reflect this structure by using these criteria.

Definition of Health Apps

The health aim of health apps can be summarized as maintaining, improving, or managing the user's health. However, it remains unclear what *health* means. Although none of the definitions of medical apps are built on a definition of health or medicine, some health app definitions (4/22, 18%) included their understanding of health. Although 18% (4/22) of the papers mentioned a *definition of health* [21,55,63,65], they all followed the WHO definition of health from 1948. On the basis of this definition, "health is a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity" [73]. Although there is no consensus on what aspects the term *health* entails, many authors see the WHO definition as outdated [74-77]. It is criticized for defining health as binary (complete health or no health) and, therefore, as a utopian status that cannot be achieved [77]. When we use the term *health* (aim), we consequently do not understand health as a binary construct, as the WHO defined it once. Instead, we see health as the "capability to cope with and to manage one's malaise and well-being conditions" [75]. This capability is affected by potential limitations of the social, biological, physical, or interpersonal environment [78].

Building on the literature, *health apps are software programs on mobile devices that process health-related data on or for their user. Every health-conscious individual can use them, be it medical laypersons, family caregivers, or health professionals, to maintain, improve, or manage the health of an individual as well as communities or the whole population of a country* (such as COVID-19–tracing apps). Data processing, in our understanding, includes any operation performed on personal data (eg, collecting; organizing; storing; adapting; visualizing; retrieving; disseminating or otherwise making it available; restricting; or erasing) according to the General Data Protection Regulation [79].

Definition of Medical Apps

Medical apps are generally used for clinical and medical purposes and can but do not have to be legally regulated as mobile medical devices. From a public health perspective, they focus on secondary (early diagnosis and treatment of acute diseases or injuries) and tertiary (rehabilitation and management of chronic diseases) prevention. Nevertheless, the health aim of medical apps leads to the problem that some health apps (eg, fitness trackers or diary apps) could be seen as medical apps when used for self-monitoring by people with chronic illnesses. This would turn the health app into a medical app. To avoid delimitation questions, we propose integrating the manufacturer's intention into the health aim of the app (see the legal definition of medical devices). Accordingly, medical apps are intended for medical purposes, which is stated in the app's description, privacy policies, or terms and conditions. Thus, we propose the following holistic definition of medical apps: *Medical apps are a subgroup of health apps that focus on secondary and tertiary prevention. They share the same technological functions (processing of health-related data) as health apps, and can be used on mobile devices. The main target groups (but not exclusively) are health professionals, patients, and family caregivers.*

Following the legal definitions of DiPA and DiGA, both are subgroups of medical apps. DiGA must necessarily be used for medical purposes as they are medical devices. As the legislator implies that caring is a medical purpose, DiPA indirectly share the intended use for medical purposes, too.

Health-Related Apps

We identified a third group of apps during the literature search for definitions of health and medical apps: health-related apps. This relatively vague term was used in different publications [4,25,80]. It remains unclear to what extent this type of app is related to health and whether apps within this group are homogeneous enough for an umbrella definition. We want to emphasize that the term *health-related app* should not be used as a category because of its fuzzy nature. This term displays the uncertainty within the scientific community of what a health app is and, therefore, expresses the need for a clear definition.

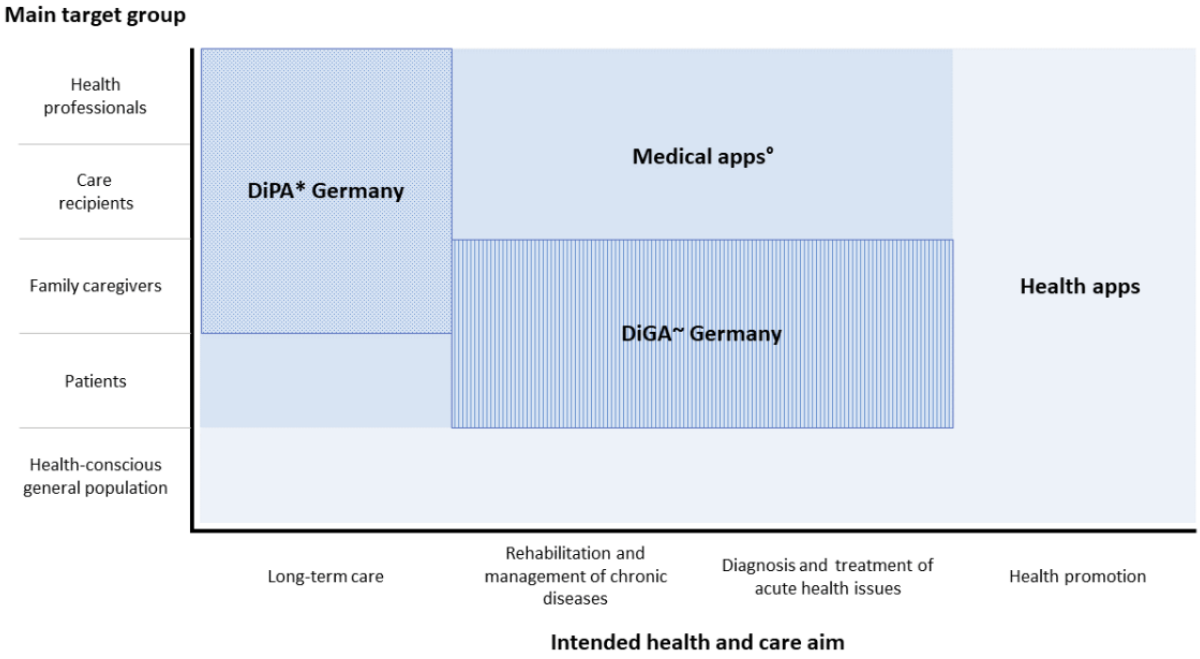
Relationship Between Health and Medical Apps

Our findings regarding the differentiation between health and medical apps (and the 2 exceptional German cases of DiGA and DiPA) are displayed in Figure 2. Comparing the definitions of health apps with those of medical apps, we recognized that some criteria are equally relevant for both types of apps (ie, definition criteria), whereas others are not (ie, distinction criteria). The criteria *device*, *collected data*, and *technological function* are the definition criteria as they are equal for health and medical apps. In addition, regulation should not be a distinction criterion either. Defining medical apps by the way they are regulated requires legal knowledge, which is neither suitable nor practical for a definition and cannot be expected from all stakeholders.

On the contrary, the user group is a distinction criterion for the *main* user group. Equally, the health aim of an app is a distinction criterion. This leads to our understanding of health apps as the umbrella term, which includes medical apps as a subgroup.

XSL•FO
RenderX

Figure 2. The connection between health apps, medical apps, digital health applications (DiGA), and digital care applications (DiPA). °Can be affected by medical device regulation because of the intended use for medical purposes. ~Regulated by section 33a of volume 5 of the German Social Insurance Code and DiGAV as low-risk mobile medical devices in Germany only. *Regulated by section 40a of volume 11 of the German Social Insurance Code in Germany only.



Decision Flowchart for Health and Medical Apps

On the basis of our analysis of the definitions derived from scientific literature and the legal regulations in the United States, the EU, and Germany, we designed a flowchart to classify health and medical apps (Multimedia Appendix 3).

Most of the definitions of medical apps given by scientific papers (6/11, 55%) matched the legal regulations in the United States and the EU insofar as they listed medical purposes (if they did not, it was because the scientific descriptions were not as specific as the legal papers). Owing to this finding, the flowchart offers insights on health and medical apps based on the legal definitions in the United States and the EU. However, it does not provide insights into whether an app is legally regulated. Instead, it showcases a specific app that fits the legal definitions in the named regions. However, other countries might have different legal regulations for mobile medical apps, so this flowchart can only report on the 2 regions given as case studies. The decision tree is intended to provide a low-threshold and practical way to differentiate between health and medical apps quickly. Therefore, it can be used by scientists and developers, politicians, or users of health-related apps. Owing to the recent introduction of DiPA, a concrete regulatory framework for these apps is still missing in Germany. As such, the flowchart does not include DiPA.

However, it is still unclear whether there will be more health-specific legal definitions once other countries begin regulating apps, similar to how Germany regulates DiGA and DiPA. Furthermore, it is unknown how app stores define and use the terms *health app* and *medical app* in their app categorization. Building on these questions, it is also unclear how the many different definitions affect the various stakeholders. It is yet to be examined what specific user groups expect from health and medical apps and how they would define both (eg, do health-conscious people or family caregivers want to use them in their daily lives?). It is also not evident whether prescribed apps affect the health care systems in any way and whether they improve health care. Finally, it remains unclear how we can ensure that health apps do not harm their users as most of them are unregulated. Further research needs to be conducted to answer all these questions.

Strengths and Limitations

To our knowledge, this is the first study to use an interdisciplinary approach to differentiate between health and medical apps from a legal and scientific point of view. By analyzing various sources, we provided an interdisciplinary and international overview of the multiple uses of the terms *health app* and *medical app*. The in-depth discussion of the scientific and legal perspective on health and medical apps offers a more holistic and deeper understanding of the similarities and differences between these terms. We conducted a qualitative analysis using the MAXQDA software to ensure the transparency of the analysis process of the identified definitions. To present the results of our approach, we incorporated our research findings into the creation of a decision chart (Multimedia Appendix 3). This may not only assist researchers in fostering a shared and clear understanding of the terms but also enable practitioners to interpret the findings of our study efficiently.

We do recognize that our research is prone to certain limitations. First, we applied the snowballing method in line with Wohlin [32]. However, although using a systematic collection technique [81] might have yielded different results, one of the key strengths of this review is its diversity of included disciplines and breadth of content. Hence, as the goal of our study was to be as inclusive as possible and analyze a wide range of fields

XSL•FO
RenderX

and databases, the snowballing method was the most feasible option. Furthermore, we excluded gray literature that was not cited in our references or displayed the legal regulation in regions other than the United States or the EU. The same applies to the descriptions of apps in the app stores. We deemed this approach appropriate as the focus was on using the terms in science. However, scientists and policy makers are not the only parties involved in developing apps for health. Other stakeholders, such as developers or users, may have a scientific education but may struggle to discern health and medical apps represented in scientific publications. Subsequent research should be dedicated primarily to these sources to incorporate practitioners and other legal perspectives into our definitions of health and medical apps. It should further be noted that we only included literature published in Chinese, English, or German. Economically speaking, a large part of the health and medical app market is in the Asian region (eg, Japan). Therefore, from a global perspective, our research is undoubtedly fragmentary. Finally, it should be noted that this study does not report whether other publications used the identified scientific definitions.

Conclusions

With our literature review and qualitative analysis of the scientific definitions of health and medical apps and the legal perspective on medical apps, we offered the first step toward defining health and medical apps more holistically—health apps are software programs on mobile devices that process health-related data on or for their users. Every health-conscious individual can use them, be it medical laypersons, family caregivers, or health professionals, to maintain, improve, or manage an individual's and the community's health. As an umbrella term, health apps include medical apps.

Medical apps can be defined as a subgroup of health apps that have the same technological functions (processing health-related data) as health apps and can be used on mobile devices. The main target user groups (not exclusively) include health professionals, patients, and family caregivers. Medical apps share the intended use for clinical and medical purposes and can be legally regulated as mobile medical devices. From a public health perspective, they focus on secondary (early diagnosis and treatment of acute diseases or injuries) and tertiary (rehabilitation and management of chronic diseases) prevention. For the special case of Germany, medical apps include DiGA and DiPA.

Acknowledgments

The authors would like to thank Felix Holl for his input and critical feedback on the manuscript and for strengthening the scope of this paper. Furthermore, the authors gratefully acknowledge the support of Leibniz ScienceCampus Bremen Digital Public Health, jointly funded by the Leibniz Association (W4/2018), the Federal State of Bremen, and the Leibniz Institute for Prevention Research and Epidemiology – BIPS.

Authors' Contributions

All authors screened the papers for definitions of health and medical apps from their disciplines. LM wrote the initial draft of the paper, developed its structure, designed the figure, and organized the manuscript. MF wrote the first draft of the introduction, wrote the legal section, structured the argumentation of the discussion, translated the German definitions into English, and designed a first draft of the flowchart. CCP conducted and interpreted the qualitative analysis of the included definitions and wrote the first draft of the qualitative analysis of the included definitions. HHD contributed to the first draft of the conclusions and discussion, structured the paper's argumentation, translated the German definitions into English, and contributed to writing the manuscript. JN edited the draft and oversaw the process. TJ edited the draft, translated the former German definitions into English for a quality check, and oversaw the process.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Word clouds of definitions for health apps and medical apps (including word count per term).
[PDF File (Adobe PDF File), 543 KB - mhealth_v10i10e37980_app1.pdf]

Multimedia Appendix 2
Overview of the criteria included in the definitions of health and medical apps.
[PDF File (Adobe PDF File), 212 KB - mhealth_v10i10e37980_app2.pdf]

Multimedia Appendix 3
Decision flowchart.
[PDF File (Adobe PDF File), 137 KB - mhealth_v10i10e37980_app3.pdf]

XSL•FO
RenderX

References

1. Rathbone AL, Prescott J. The use of mobile apps and SMS messaging as physical and mental health interventions: systematic review. J Med Internet Res 2017 Aug 24;19(8):e295 [FREE Full text] [doi: 10.2196/jmir.7740] [Medline: 28838887]
2. Pearson AL, Mack E, Namanya J. Mobile phones and mental well-being: initial evidence suggesting the importance of staying connected to family in rural, remote communities in Uganda. PLoS One 2017;12(1):e0169819 [FREE Full text] [doi: 10.1371/journal.pone.0169819] [Medline: 28095427]
3. mHealth: New Horizons for Health Through Mobile Technologies: Second Global Survey on eHealth. World Health Organization. 2011. URL: https://apps.who.int/iris/handle/10665/44607 [accessed 2022-10-21]
4. Becker S, Miron-Shatz T, Schumacher N, Krocza J, Diamantidis C, Albrecht U. mHealth 2.0: experiences, possibilities, and perspectives. JMIR Mhealth Uhealth 2014 May 16;2(2):e24 [FREE Full text] [doi: 10.2196/mhealth.3328] [Medline: 25099752]
5. Naslund JA, Shidhaye R, Patel V. Digital technology for building capacity of nonspecialist health workers for task sharing and scaling up mental health care globally. Harv Rev Psychiatry 2019;27(3):181-192 [FREE Full text] [doi: 10.1097/HRP.0000000000000217] [Medline: 30958400]
6. Van Ameringen M, Turna J, Khalesi Z, Pullia K, Patterson B. There is an app for that! The current state of mobile applications (apps) for DSM-5 obsessive-compulsive disorder, posttraumatic stress disorder, anxiety and mood disorders. Depress Anxiety 2017 Jun;34(6):526-539 [FREE Full text] [doi: 10.1002/da.22657] [Medline: 28569409]
7. Kampmeijer R, Pavlova M, Tambor M, Golinowska S, Groot W. The use of e-health and m-health tools in health promotion and primary prevention among older adults: a systematic literature review. BMC Health Serv Res 2016 Sep 05;16 Suppl 5:290 [FREE Full text] [doi: 10.1186/s12913-016-1522-3] [Medline: 27608677]
8. Biswas M, Tania MH, Kaiser MS, Kabir R, Mahmud M, Kemal AA. ACCU3RATE: a mobile health application rating scale based on user reviews. PLoS One 2021;16(12):e0258050 [FREE Full text] [doi: 10.1371/journal.pone.0258050] [Medline: 34914718]
9. Wu X, Guo X, Zhang Z. The efficacy of mobile phone apps for lifestyle modification in diabetes: systematic review and meta-analysis. JMIR Mhealth Uhealth 2019 Jan 15;7(1):e12297 [FREE Full text] [doi: 10.2196/12297] [Medline: 30664494]
10. Ahmed I, Ahmad NS, Ali S, Ali S, George A, Saleem Danish H, et al. Medication adherence apps: review and content analysis. JMIR Mhealth Uhealth 2018 Mar 16;6(3):e62 [FREE Full text] [doi: 10.2196/mhealth.6432] [Medline: 29549075]
11. Boulos MN, Wheeler S, Tavares C, Jones R. How smartphones are changing the face of mobile and participatory healthcare: an overview, with example from eCAALYX. Biomed Eng Online 2011 Apr 05;10:24 [FREE Full text] [doi: 10.1186/1475-925X-10-24] [Medline: 21466669]
12. Cano Martín JA, Martínez-Pérez B, de la Torre-Díez I, López-Coronado M. Economic impact assessment from the use of a mobile app for the self-management of heart diseases by patients with heart failure in a Spanish region. J Med Syst 2014 Sep;38(9):96. [doi: 10.1007/s10916-014-0096-z] [Medline: 24994514]
13. Luxton DD, Hansen RN, Stanfill K. Mobile app self-care versus in-office care for stress reduction: a cost minimization analysis. J Telemed Telecare 2014 Dec;20(8):431-435. [doi: 10.1177/1357633X14555616] [Medline: 25316037]
14. Nghiem N, Leung W, Cleghorn C, Blakely T, Wilson N. Mass media promotion of a smartphone smoking cessation app: modelled health and cost-saving impacts. BMC Public Health 2019 Mar 08;19(1):283 [FREE Full text] [doi: 10.1186/s12889-019-6605-8] [Medline: 30849943]
15. Mobasheri M, Johnston M, King D, Leff D, Thiruchelvam P, Darzi A. Smartphone breast applications - what's the evidence? Breast 2014 Oct;23(5):683-689 [FREE Full text] [doi: 10.1016/j.breast.2014.07.006] [Medline: 25153432]
16. Schüz B, Urban M. [Unintended consequences and side effects of digital health technology: a public health perspective]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020 Feb;63(2):192-198. [doi: 10.1007/s00103-019-03088-5] [Medline: 31950231]
17. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance). Publications Office of the European Union. 2017 May. URL: https://op.europa.eu/en/publication-detail/-/publication/83bdc18f-315d-11e7-9412-01aa75ed71a1/language-en/format-PDF/source-58036705 [accessed 2022-10-04]
18. von Zezschwitz F. Neue regulatorische Herausforderungen für Anbieter von Gesundheits-Apps. MedR 2020 Mar 11;38(3):196-201 [FREE Full text] [doi: 10.1007/s00350-020-5482-6]
19. Hensher M, Cooper P, Dona SW, Angeles MR, Nguyen D, Heynsbergh N, et al. Scoping review: development and assessment of evaluation frameworks of mobile health apps for recommendations to consumers. J Am Med Inform Assoc 2021 Jun 12;28(6):1318-1329 [FREE Full text] [doi: 10.1093/jamia/ocab041] [Medline: 33787894]
20. Münkler L. Health-Apps im gesundheitsrechtlichen Regierungsgefüge. NZS 2021:41 [FREE Full text]
21. Albrecht UV, Pramann O, von Jan U. Synopsis for health apps: transparency for trust and decision making. In: Househ M, Borycki E, Kushniruk A, editors. Social Media and Mobile Technologies for Healthcare. Hershey: IGI Global; 2014:94-108.
22. Petersen C, Adams SA, DeMuro PR. mHealth: don't forget all the stakeholders in the business case. Med 2 0 2015 Dec 31;4(2):e4 [FREE Full text] [doi: 10.2196/med20.4349] [Medline: 26720310]

XSL•FO
RenderX

23. DiGA. German Federal Institute for Drugs and Medical Devices. URL: https://www.bfarm.de/EN/Medical-devices/Tasks/DiGA-and-DiPA/Digital-Health-Applications/_node.html;jsessionid=21BD1B9C03EE2C32E6C7D01685394A27.intranet672 [accessed 2022-10-14]
24. Degelsegger-Márquez A, Trunner K, Piso B. Medication apps - a systematic search and classification. In: Schreier G, Hayn D, Eggerth A, editors. dHealth 2020 – Biomedical Informatics for Health and Care. Amsterdam: IOS Press; Jun 23, 2020.
25. BinDhim NF, Trevena L. There's an app for that: a guide for healthcare practitioners and researchers on smartphone technology. Online J Public Health Inform 2015;7(2):e218 [FREE Full text] [doi: 10.5210/ojphi.v7i2.5522] [Medline: 26392848]
26. Mosa AS, Yoo I, Sheets L. A systematic review of healthcare applications for smartphones. BMC Med Inform Decis Mak 2012 Jul 10;12:67 [FREE Full text] [doi: 10.1186/1472-6947-12-67] [Medline: 22781312]
27. Terry NP. Mobile health: assessing the barriers. Chest 2015 May;147(5):1429-1434 [FREE Full text] [doi: 10.1378/chest.14-2459] [Medline: 25940254]
28. Evidence standards framework for digital health technologies. National Institute for Health and Care Excellence (NICE). 2022. URL: https://www.nice.org.uk/guidance/ecd7/resources/evidence-standards-framework-for-digital-health-technologies-pdf-1124017457605 [accessed 2022-10-04]
29. Iribarren SJ, Cato K, Falzon L, Stone PW. What is the economic evidence for mHealth? A systematic review of economic evaluations of mHealth solutions. PLoS One 2017;12(2):e0170581 [FREE Full text] [doi: 10.1371/journal.pone.0170581] [Medline: 28152012]
30. Green J, Thorogood N. Qualitative Methods for Health Research. Thousand Oaks, California: SAGE Publications; Mar 22, 2018.
31. Lecy JD, Beatty KE. Representative literature reviews using constrained snowball sampling and citation network analysis. SSRN J 2012 Jan 27. [doi: 10.2139/ssrn.1992601]
32. Wohlin C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. 2014 Presented at: EASE '14: 18th International Conference on Evaluation and Assessment in Software Engineering; May 13 - 14, 2014; London England. [doi: 10.1145/2601248.2601268]
33. Badampudi D, Wohlin C, Petersen K. Experiences from using snowballing and database searches in systematic literature studies. In: Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering. 2015 Presented at: EASE '15: 19th International Conference on Evaluation and Assessment in Software Engineering; Apr 27 - 29, 2015; Nanjing China. [doi: 10.1145/2745802.2745818]
34. Eitan AT, Smolyansky E, Harpaz IK, Perets S. Explore connected papers in a visual graph. Conneced Papers. URL: https://www.connectedpapers.com/ [accessed 2022-10-04]
35. Kramer U. Wie gut sind Gesundheits-Apps? Aktuel Ernahrungsmed 2017 Jun 20;42(03):193-205. [doi: 10.1055/s-0043-109130]
36. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: implications for conducting a qualitative descriptive study. Nurs Health Sci 2013 Sep;15(3):398-405 [FREE Full text] [doi: 10.1111/nhs.12048] [Medline: 23480423]
37. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
38. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Official Journal of the European Union. 2016 May 4. URL: https://op.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1 [accessed 2022-10-04]
39. -Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. European Union. 2017 May. URL: https://op.europa.eu/en/publication-detail/-/publication/83bdc18f-315d-11e7-9412-01aa75ed71a1/language-en/format-PDF/source-58036705 [accessed 2022-10-04]
40. 53 MDR Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. European Union. 2017 May. URL: https://op.europa.eu/en/publication-detail/-/publication/83bdc18f-315d-11e7-9412-01aa75ed71a1/language-en/format-PDF/source-58036705 [accessed 2022-10-04]
41. 21 USC 321: Definitions; generally. Federal Food, Drug, and Cosmetic Act (FD&C Act). URL: https://uscode.house.gov/view.xhtml?req=granuleid:USC-prelim-title21-section321&num=0&edition=prelim [accessed 2022-10-04]
42. Policy for device software functions and mobile medical applications guidance for industry and food and drug administration staff. U.S. Food & Drug Administration. 2022 Sep. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/policy-device-software-functions-and-mobile-medical-applications [accessed 2022-10-04]
43. Medical devices. Differentiation and classification. German Federal Institute for Drugs and Medical Devices (BfArM). 2022. URL: https://www.bfarm.de/EN/Medical-devices/Tasks/Differentiation-and-classification/_node.html;jsessionid=41CB9196DC8D464C7991BE2CB09567D3.intranet382 [accessed 2022-10-04]

XSL•FO
RenderX

44. Second draft of guidelines EU guidelines on assessment of the reliability of mobile health applications. European Commission. 2016. URL: https://www.ospi.es/export/sites/ospi/documents/documentos/servicios-publicos-digitales/mHealth-Assessment_Guidelines_Apps_2_Draft_20160530.pdf [accessed 2022-10-04]
45. Social Insurance Code 5. § 33a Sect. 1 Digital Health Applications. Social Health Insurance. URL: https://www.gesetze-im-internet.de/sgb_5/__33a.html [accessed 2022-10-04]
46. Social Insurance Code 11. §40a Digital Care Applications. Social Care Insurance. URL: https://www.gesetze-im-internet.de/sgb_11/__40a.html [accessed 2022-10-04]
47. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (Text with EEA relevance.). Publications Office of the European Union. URL: https://op.europa.eu/en/publication-detail/-/publication/83bdc18f-315d-11e7-9412-01aa75ed71a1/language-en/format-PDF/source-58036705 [accessed 2022-10-04]
48. Wang C, Qi H. Influencing factors of acceptance and use behavior of mobile health application users: systematic review. Healthcare (Basel) 2021 Mar 22;9(3):357 [FREE Full text] [doi: 10.3390/healthcare9030357] [Medline: 33809828]
49. Karakoyun T, Podhaisky HP, Frenz A, Schuhmann-Giampieri G, Ushikusa T, Schröder D, et al. Digital medical device companion (MyIUS) for new users of intrauterine systems: app development study. JMIR Med Inform 2021 Jul 13;9(7):e24633 [FREE Full text] [doi: 10.2196/24633] [Medline: 34255688]
50. Azad-Khaneghah P, Neubauer N, Miguel Cruz A, Liu L. Mobile health app usability and quality rating scales: a systematic review. Disabil Rehabil Assist Technol 2021 Oct;16(7):712-721. [doi: 10.1080/17483107.2019.1701103] [Medline: 31910687]
51. Morse SS, Murugiah MK, Soh YC, Wong TW, Ming LC. Mobile health applications for pediatric care: review and comparison. Ther Innov Regul Sci 2018 May;52(3):383-391 [FREE Full text] [doi: 10.1177/2168479017725557] [Medline: 29714538]
52. Sun RT, Han W, Chang HL, Shaw MJ. Motivating adherence to exercise plans through a personalized mobile health app: enhanced action design research approach. JMIR Mhealth Uhealth 2021 Jun 02;9(6):e19941 [FREE Full text] [doi: 10.2196/19941] [Medline: 34076580]
53. Aitken M, Lyle J. Patient adoption of mHealth: use, evidence and remaining barriers to mainstream acceptance. IMS Institute for Healthcare Informatics. 2015 Sep. URL: https://www.iqvia.com/-/media/iqvia/pdfs/institute-reports/patient-adoption-of-mhealth.pdf [accessed 2022-10-04]
54. Klasnja P, Pratt W. Healthcare in the pocket: mapping the space of mobile-phone health interventions. J Biomed Inform 2012 Feb;45(1):184-198 [FREE Full text] [doi: 10.1016/j.jbi.2011.08.017] [Medline: 21925288]
55. Scherenberg V, Kramer U. Schöne neue Welt: Gesünder mit Health-Apps? Hintergründe, Handlungsbedarf und schlummernde Potenziale. In: Strahlendorf P, editor. Jahrbuch Healthcare Marketing. Hamburg: New Business; 2013.
56. Aungst TD, Clauson KA, Misra S, Lewis TL, Husain I. How to identify, assess and utilise mobile medical applications in clinical practice. Int J Clin Pract 2014 Feb;68(2):155-162 [FREE Full text] [doi: 10.1111/ijcp.12375] [Medline: 24460614]
57. Boudreaux ED, Waring ME, Hayes RB, Sadasivam RS, Mullen S, Pagoto S. Evaluating and selecting mobile health apps: strategies for healthcare providers and healthcare organizations. Behav Med Pract Policy Res 2014 Sep 24;4(4):363-371 [FREE Full text] [doi: 10.1007/s13142-014-0293-9]
58. Green Paper on mobile health ("mHealth"). European Commission. 2014 Apr 10. URL: https://digital-strategy.ec.europa.eu/en/library/green-paper-mobile-health-mhealth [accessed 2022-10-04]
59. Seabrook HJ, Stromer JN, Shevkenek C, Bharwani A, de Grood J, Ghali WA. Medical applications: a database and characterization of apps in Apple iOS and Android platforms. BMC Res Notes 2014 Aug 27;7:573 [FREE Full text] [doi: 10.1186/1756-0500-7-573] [Medline: 25167765]
60. Gehring H, Pramann O, Imhoff M, Albrecht UV. [Future trend medical apps. From the apps store directly into medical practice?]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2014 Dec;57(12):1402-1410 [FREE Full text] [doi: 10.1007/s00103-014-2061-x] [Medline: 25388438]
61. Powell AC, Landman AB, Bates DW. In search of a few good apps. JAMA 2014 May 14;311(18):1851-1852. [doi: 10.1001/jama.2014.2564] [Medline: 24664278]
62. Lucht M, Bredenkamp R, Boeker M, Kramer U. Gesundheits- und Versorgungs-Apps. Hintergründe zu deren Entwicklung und Einsatz. Freiburg: Universitätsklinikum Freiburg; 2015.
63. Albrecht UV, von Jahn U. Einführung und Begriffsbestimmungen. In: Albrecht UV, editor. Chancen und Risiken von Gesundheits-Apps (CHARISMHA). Braunschweig: Technische Universität Braunschweig; 2016.
64. Evers-Wölk M, Oertel B, Sonk M, Jacobs M. Gesundheits-Apps. Innovationsanalyse. TAB-Arbeitsbericht Nr 179. Berlin: Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag; 2018.
65. Gregor-Haack J. [Reimbursement of health apps by the German statutory health insurance]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2018 Mar;61(3):328-333 [FREE Full text] [doi: 10.1007/s00103-018-2689-z] [Medline: 29374299]

XSL•FO
RenderX

66. Groß D, Schmidt M. [Ethical perspectives on E-health and health apps : is all that is achievable desirable?]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2018 Mar;61(3):349-357 [FREE Full text] [doi: 10.1007/s00103-018-2697-z] [Medline: 29383392]
67. Moshi MR, Tooher R, Merlin T. Suitability of current evaluation frameworks for use in the health technology assessment of mobile medical applications: a systematic review. Int J Technol Assess Health Care 2018 Sep 11;34(5):464-475 [FREE Full text] [doi: 10.1017/s026646231800051x]
68. Heretleif M, Huptas D, Neureither J, Schmidt A, Schulze M, Stöcker N, Federal Office of Information Security Germany (BSI). IT-Sicherheit auf dem digitalen Verbrauchermarkt: Fokus Gesundheits-Apps. Federal Office of Information Security Germany (BSI). 2021 May 17. URL: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/DVS-Berichte/gesundheitsapps.pdf?__blob=publicationFile&v=2 [accessed 2022-10-04]
69. Volpi SS, Biduski D, Bellei EA, Tefili D, McCleary L, Alves AL, et al. Using a mobile health app to improve patients' adherence to hypertension treatment: a non-randomized clinical trial. PeerJ 2021;9:e11491 [FREE Full text] [doi: 10.7717/peerj.11491] [Medline: 34123593]
70. Golden EA, Zweig M, Danieletto M, Landell K, Nadkarni G, Bottinger E, et al. A resilience-building app to support the mental health of health care workers in the COVID-19 era: design process, distribution, and evaluation. JMIR Form Res 2021 May 05;5(5):e26590 [FREE Full text] [doi: 10.2196/26590] [Medline: 33872189]
71. Racioppi A, Dalton T, Ramalingam S, Romero K, Ren Y, Bohannon L, et al. Assessing the feasibility of a novel mHealth app in hematopoietic stem cell transplant patients. Transplant Cell Ther 2021 Feb;27(2):181.e1-181.e9. [doi: 10.1016/j.jtct.2020.10.017] [Medline: 33830035]
72. Tobias G, Spanier AB. Using an mhealth app (iGAM) to reduce gingivitis remotely (part 2): prospective observational study. JMIR Mhealth Uhealth 2021 Sep 16;9(9):e24955 [FREE Full text] [doi: 10.2196/24955] [Medline: 34528897]
73. World Health Organization (WHO). Constitution of the World Health Organization. 1948. URL: https://www.who.int/about/governance/constitution [accessed 2022-10-04]
74. Larson JS. The World Health Organization's definition of health: social versus spiritual health. Soc Indic Res 1996 Jun;38(2):181-192. [doi: 10.1007/bf00300458]
75. Leonardi F. The definition of health: towards new perspectives. Int J Health Serv 2018 Oct;48(4):735-748. [doi: 10.1177/0020731418782653] [Medline: 29902944]
76. Saracci R. The World Health Organisation needs to reconsider its definition of health. BMJ 1997 May 10;314(7091):1409-1410 [FREE Full text] [doi: 10.1136/bmj.314.7091.1409] [Medline: 9161320]
77. Segre M, Carvalho Ferraz F. The health's concept. Rev Saúde Pública 1997 Oct;31(5):538-542 [FREE Full text] [doi: 10.1590/S0034-89101997000600016]
78. Starfield B. Basic concepts in population health and health care. J Epidemiol Community Health 2001 Jul;55(7):452-454 [FREE Full text] [doi: 10.1136/jech.55.7.452] [Medline: 11413173]
79. 4 GDPR. Sect. 2 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). European Union. URL: https://op.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1 [accessed 2022-10-04]
80. van Kerkhof LW, van der Laar CW, de Jong C, Weda M, Hegger I. Characterization of apps and other e-tools for medication use: insights into possible benefits and risks. JMIR Mhealth Uhealth 2016 Apr 06;4(2):e34 [FREE Full text] [doi: 10.2196/mhealth.4149] [Medline: 27052946]
81. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al. Cochrane Handbook for systematic reviews of interventions. Cochrane Training. 2022. URL: https://training.cochrane.org/handbook/current [accessed 2022-10-04]

Abbreviations

DiGA: digital health applications
DiGAV: Digital Health Applications Regulation
DiPA: digital care applications
EU: European Union
FDA: Food and Drug Administration
MDR: Medical Device Regulation
mHealth: mobile health
WHO: World Health Organization

XSL•FO
RenderX

Edited by L Buis; submitted 14.03.22; peer-reviewed by R Kabir, H Qi, H Mehdizadeh, A Joseph; comments to author 05.05.22; revised version received 30.05.22; accepted 30.06.22; published 31.10.22.

Please cite as:
Maaß L, Freye M, Pan CC, Dassow HH, Niess J, Jahnel T
The Definitions of Health Apps and Medical Apps From the Perspective of Public Health and Law: Qualitative Analysis of an Interdisciplinary Literature Overview
JMIR Mhealth Uhealth 2022;10(10):e37980
URL: https://mhealth.jmir.org/2022/10/e37980
doi:10.2196/37980
PMID:36315221

©Laura Maaß, Merle Freye, Chen-Chia Pan, Hans-Henrik Dassow, Jasmin Niess, Tina Jahnel. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 31.10.2022. 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.

XSL•FO
RenderX

Original Paper

Electronic Health Diary Campaigns to Complement Longitudinal Assessments in Persons With Multiple Sclerosis: Nested Observational Study

Chloé Sieber[1,2], MSc; Deborah Chiavi[1,2], MSc; Christina Haag[1,2], PhD; Marco Kaufmann[1], PhD; Andrea B Horn[3,4,5], Dr rer nat; Holger Dressel[6], MD, Prof Dr; Chiara Zecca[7,8], MD, Prof Dr; Pasquale Calabrese[9,10], MD, Dr rer nat, Prof Dr; Caroline Pot[11], MD, Prof Dr; Christian Philipp Kamm[12,13], MD, Prof Dr; Viktor von Wyl[1,2], PhD, Prof Dr; Swiss Multiple Sclerosis Registry[14]

[1]Swiss Multiple Sclerosis Registry, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland
[2]Institute for Implementation Science in Health Care, University of Zurich, Zürich, Switzerland
[3]University Research Priority Program "Dynamics of Healthy Aging", University of Zurich, Zurich, Switzerland
[4]Competence Center of Gerontology, University of Zurich, Zurich, Switzerland
[5]Department of Psychology, University of Zurich, Zurich, Switzerland
[6]Division of Occupational and Environmental Medicine, University of Zurich and University Hospital Zurich, Zurich, Switzerland
[7]Multiple Sclerosis Center, Neurocenter of Southern Switzerland, Ente Ospedaliero Cantonale, Lugano, Switzerland
[8]Faculty of Biomedical Sciences, Università della Svizzera Italiana, Lugano, Switzerland
[9]Neuropsychology and Behavioral Neurology Unit, Division of Molecular and Cognitive Neuroscience, University of Basel, Basel, Switzerland
[10]Department of Neurology, University Clinic of Basel, Basel, Switzerland
[11]Service of Neurology, Department of Clinical Neurosciences, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland
[12]Department of Neurology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland
[13]Neurocenter, Cantonal Hospital Lucerne, Lucerne, Switzerland
[14]See Acknowledgments

Corresponding Author:
Viktor von Wyl, PhD, Prof Dr
Swiss Multiple Sclerosis Registry
Epidemiology, Biostatistics and Prevention Institute
University of Zurich
Hirschengraben 84
Zurich, 8001
Switzerland
Phone: 41 44 634 63 80
Email: viktor.vonwyl@uzh.ch

Abstract

Background: Electronic health diaries hold promise in complementing standardized surveys in prospective health studies but are fraught with numerous methodological challenges.

Objective: The study aimed to investigate participant characteristics and other factors associated with response to an electronic health diary campaign in persons with multiple sclerosis, identify recurrent topics in free-text diary entries, and assess the added value of structured diary entries with regard to current symptoms and medication intake when compared with survey-collected information.

Methods: Data were collected by the Swiss Multiple Sclerosis Registry during a nested electronic health diary campaign and during a regular semiannual Swiss Multiple Sclerosis Registry follow-up survey serving as comparator. The characteristics of campaign participants were descriptively compared with those of nonparticipants. Diary content was analyzed using the Linguistic Inquiry and Word Count 2015 software (Pennebaker Conglomerates, Inc) and descriptive keyword analyses. The similarities between structured diary data and follow-up survey data on health-related quality of life, symptoms, and medication intake were examined using the Jaccard index.

XSL•FO
RenderX

Results: Campaign participants (n=134; diary entries: n=815) were more often women, were not working full time, did not have a higher education degree, had a more advanced gait impairment, and were on average 5 years older (median age 52.5, IQR 43.25-59.75 years) than eligible nonparticipants (median age 47, IQR 38-55 years; n=524). Diary free-text entries (n=632; participants: n=100) most often contained references to the following standard Linguistic Inquiry and Word Count word categories: negative emotion (193/632, 30.5%), body parts or body functioning (191/632, 30.2%), health (94/632, 14.9%), or work (67/632, 10.6%). Analogously, the most frequently mentioned keywords (diary entries: n=526; participants: n=93) were "good," "day," and "work." Similarities between diary data and follow-up survey data, collected 14 months apart (median), were high for health-related quality of life and stable for slow-changing symptoms such as fatigue or gait disorder. Similarities were also comparatively high for drugs requiring a regular application, including interferon beta-1a (Avonex) and glatiramer acetate (Copaxone), and for modern oral therapies such as fingolimod (Gilenya) and teriflunomide (Aubagio).

Conclusions: Diary campaign participation seemed dependent on time availability and symptom burden and was enhanced by reminder emails. Electronic health diaries are a meaningful complement to regular structured surveys and can provide more detailed information regarding medication use and symptoms. However, they should ideally be embedded into promotional activities or tied to concrete research study tasks to enhance regular and long-term participation.

(JMIR Mhealth Uhealth 2022;10(10):e38709) doi:10.2196/38709

KEYWORDS

registry; multiple sclerosis; digital health; electronic health diary; diary; participation; adherence; patient-reported outcome; natural language processing; unstructured text

Introduction

Background

Electronic health diaries are consumer- or patient-facing electronic tools to record personal health-relevant information prospectively [1]. Flexibility and ease of use make electronic health diaries an attractive option for self-learning, patient empowerment, and health care management support [2]. Electronic health diaries are also used in research contexts where they offer more timely or even real-time reporting of health-related indicators and can thus meaningfully complement medical records or retrospective questionnaires [1]. In addition, free-text entries of diary studies can offer a window into a person's emotions [3-6] and daily-life contexts.

Electronic health diaries have also found application in multiple sclerosis (MS) research, where they have been used as logbooks of physical activity sessions [7] or to record disability- or health-related quality of life scores [8-10]. Furthermore, health diaries have also been explored in the context of self-management improvement programs for persons with MS [11,12], as well as to monitor and enhance medication adherence to disease-modifying treatments (eg, through reminder functionalities) [13,14].

However, numerous methodological and user challenges have been highlighted [1,15]. First, keeping health diaries over extended periods (ie, months or even years) demands continual intrinsic and extrinsic motivation, as well as substantial support from study investigators [1,15]. Second, studies relying on free-text entries require special analytical methods [1,15] that involve substantial manual data preprocessing. Finally, to increase compliance, the number of diary questions needs to be relatively small—thus limiting the depth of health information that can realistically be collected (eg, regarding a person's disease history).

Objectives

In light of these challenges, we aimed to explore the applicability and usability of electronic health diary data collected by the Swiss Multiple Sclerosis Registry (SMSR) during a week-long health diary campaign. Specifically, we aimed to compare the characteristics of the participants in the health diary campaign with eligible nonparticipants of the SMSR campaign (aim 1). Furthermore, we intended to analyze the content of the diary free-text field section by applying 2 different natural language processing methods (aim 2). Finally, we aimed to evaluate the diary-collected health-related quality of life, symptoms, and treatment information by comparing it with corresponding survey-based information (aim 3).

These investigations were guided by several literature- and research-based assumptions. Recent research has revealed that younger age is linked to a more frequent adoption of mobile health technologies [1,2]. Therefore, we hypothesized that the electronic health diary campaign participants would be younger than nonparticipants with access to the registry's web-based platform. Furthermore, although there is usually very good intraindividual consistency of stable symptoms, treatments, or side effects among semiannual surveys [16,17], the health diary may be better able to capture dynamic symptoms and treatment effects in a more comprehensive and fine-grained fashion.

Methods

Study Context

The electronic health diary data were collected by the SMSR as part of a nested campaign entitled "A week in the life of persons with MS," which ran from March 9, 2019, to March 17, 2019. The SMSR is a nationwide, patient-centered registry for persons with MS (trial registration: ClinicalTrials.gov NCT02980640) [18,19]. The registry was launched in June 2016 as a collaboration between the Swiss Multiple Sclerosis Society and the University of Zurich with the aim to create a

XSL•FO
RenderX

patient-centered registry for adult persons with MS living in Switzerland.

Since its launch in 2016, the SMSR has been collecting and updating standard information on MS symptoms, treatment histories, and health-related quality of life in semiannual surveys [19]. As of July 13, 2019, a total of 2350 participants had contributed data to the SMSR. Participants can either enroll for digital participation (approximately 80% of all participants), or they can choose to complete the surveys on paper (approximately 20%). The baseline assessment has 2 stages. First, participants complete a short survey (initial questionnaire) that collects basic sociodemographic characteristics and core data on MS diagnosis, as well as symptom and treatment histories. A second, more extensive survey (baseline questionnaire) collects current symptoms and treatments, as well as detailed information on education, work, health care use, comorbidities, and MS status [18,19].

The electronic health diary is a separate feature on the digital SMSR study platform, available in German, French, and Italian, and specifically developed to work on PCs and mobile devices. The diary was semistructured (Figures S1-S6 in Multimedia Appendix 1). It contained a free-text field where participants could describe their current health status and any other information that they deemed relevant. There were no restrictions on the number of characters. In addition, the diary included structured questions that mirrored corresponding survey items from the longitudinal surveys, namely the EQ-5D-5L and the EuroQol visual analog scale (EQ-VAS; 0-100 scale), as well as current MS symptoms (checkboxes) and current use of immunomodulatory and complementary therapies (checkboxes; Figures S5 and S6 in Multimedia Appendix 1). Visually, the diary was organized as a calendar sheet that allowed retrospective entries.

Description of the Nested Electronic Health Diary Campaign

To promote the release of the electronic health diary, the SMSR launched a campaign entitled “A week in the life of persons with MS,” which ran from March 9, 2019, to March 17, 2019. All web-based participants were invited to make daily diary entries in the free-text field and complete the structured questions on current symptoms and medication intake (Multimedia Appendix 2). Participants were encouraged to share positive and negative experiences as well as self-management strategies.

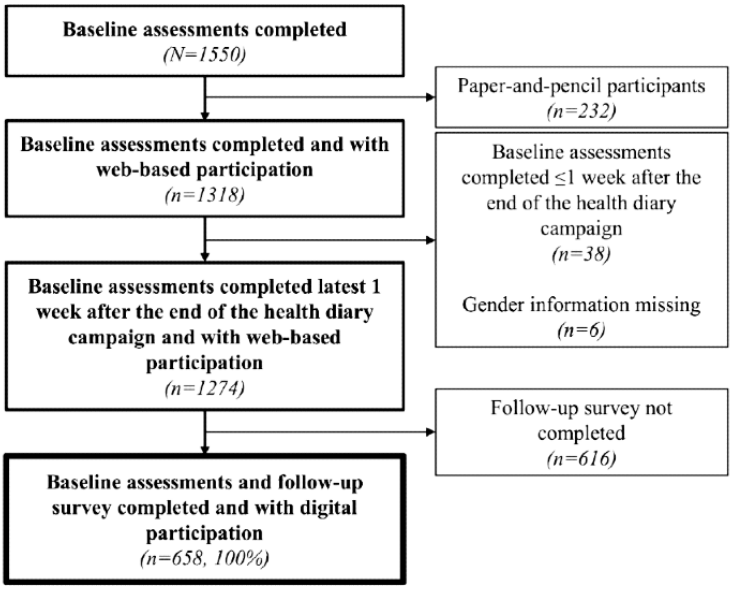
The campaign included different communication activities (Multimedia Appendix 3, red numbers). First, a general announcement was issued by email on January 28, 2019. One month later (February 28, 2019), a personal invitation was sent by email to all web-based participants, followed by a reminder on March 8, 2019 (ie, the day before the launch of the campaign). A final motivational reminder was sent through email on March 14, 2019. An announcement and daily campaign updates were posted on the SMSR website as well as on the public website and the Facebook page of the Swiss Multiple Sclerosis Society. The daily campaign updates included visual summaries of different health diary items (average EQ-VAS score, percentage of participants with a specific mood, the 3 most frequent symptoms, percentage of users of complementary therapies, and the 3 most frequent complementary therapies).

Study Population

Of the 1550 SMSR enrollees who had completed the baseline assessments, 1318 (85.03%) participated through the SMSR web platform (Figure 1). Paper-and-pencil participants were excluded a priori because the health diary was only available on the web. After the application of inclusion criteria and data quality checks, we included 96.66% (1274/1318) of the web-based enrollees with baseline assessments completed until 1 week after the campaign ended (March 26, 2019). This time frame was chosen because a small number of participants (n=21) joined the SMSR during the diary campaign but completed the baseline assessment only a few days after the campaign ended. Of these 1274 enrollees, 658 (51.65%) had completed the next semiannual survey in the spring of 2020—hereafter *follow-up* survey—that included all standardized assessments from the diary (health-related quality of life, symptoms, and medication use). Among these 658 enrollees, 134 (20.4%) had made a nonempty free-text entry in the electronic health diary collected between February 27, 2019 (1 day before the invitation to the campaign was emailed), and March 19, 2019 (2 days after the official end of the campaign), including retrospective entries (Multimedia Appendix 4). This analysis time frame was chosen because the first campaign announcement on February 28, 2019, had already triggered numerous entries (Multimedia Appendix 3). Entries for retrospective dates were included if they were created during the analysis time frame.

XSL•FO
RenderX

Figure 1. Flowchart of the study population, from the perspective of the completion of the baseline and follow-up surveys (refer to Multimedia Appendix 3 for more detailed information regarding the electronic health diary entries).



Measures

We analyzed 3 different types of data items that were collected both in the diary and the follow-up survey. The first data item consisted of textual diary entries describing notable daily events or current physical and mental well-being. Only German diary entries containing at least one word were considered for the analysis. The second analysis included the EQ-5D-5L index (derived from French reference value sets [20-22]) and the EQ-VAS. The third type of data pertained to current symptoms and the use of immunomodulatory medications [20,23].

In the diary and in the follow-up survey, participants were asked to indicate their current MS symptoms and the MS symptoms they had experienced within the last 12 months, respectively, from the following list: visual impairments, speech disorders, swallowing difficulties (dysphagia), weakness, signs of paralysis, fatigue, paresthesia (eg, deafness and tingling), dizziness, pain, gait disorder, vestibular disorders, bladder disorders (eg, bladder weakness), spasms (muscle cramps), convulsions and tics, tremor, intestinal disorders (eg, constipation), epileptic convulsions, sexual disorders, memory disorders, depression, concentration problems, problems with spatial orientation, affective lability or lack of control over emotions, or other unspecified symptoms.

Any current use of an immunomodulatory MS drug (diary and follow-up survey) and use of an immunomodulatory MS drug within the last 6 months (follow-up survey) were collected based on the following medication list: teriflunomide (Aubagio), interferon beta-1a (Avonex), interferon beta-1b (Betaferon), glatiramer acetate (Copaxone), cyclophosphamide (Endoxan), interferon beta-1b (Extavia), fingolimod (Gilenya), azathioprine (Imurek), alemtuzumab (Lemtrada), rituximab (MabThera), laquinimod (Nerventra), mitoxantrone (Novantron), ocrelizumab (Ocrevus), peginterferon beta-1a (Plegridy), interferon beta-1a (Rebif), cyclosporine (Sandimmun), tetracosactide (Synacthen), dimethyl fumarate (Tecfidera), natalizumab (Tysabri), and other unspecified medications.

Statistical Analyses

Baseline characteristics of diary campaign participants (that is, persons with at least one valid health diary entry during the analysis time frame) and nonparticipants were compared descriptively (aim 1). Continuous data were analyzed by medians and IQRs and categorical information by frequency counts and percentages. Using visual plots, we analyzed the frequency and patterns of diary use over time, intending to learn more about the impact of reminders on participants' behavior.

For aim 2, the content of diary text entries was analyzed using the Linguistic Inquiry and Word Count (LIWC) 2015 software (Pennebaker Conglomerates, Inc) [24,25] using a German dictionary [26]. On the basis of this dictionary, the LIWC program scores the texts according to 118 word attributes and categories by dividing the number of words belonging to a specific attribute or category by the number of words within a text entry to standardize for diary entry length [24]. The following seven categories related to health and personal experience were selected for further inspection: (1) health, (2) body, (3) family, (4) friends, (5) work, (6) positive emotions (eg, happy, good, and love), and (7) negative emotions (eg, sadness, anger, and fear). Using these 7 word categories, we calculated the percentage of entries containing at least one category-specific keyword across all diary entries. In addition, we analyzed an LIWC-integrated summary score for emotional tone by plotting its distribution across all diary entries in a histogram. The LIWC-generated *emotional tone* score represents the overall emotional tone of a text, ranging from 0 (negative) to 100 (positive). The score is computed based on the difference between dictionary-based positive and negative emotion scores, but details of this proprietary algorithm are not publicly available [24,27].

Furthermore, German diary text entries containing at least 10 words were analyzed descriptively by visualizing them as a word cloud of the 100 most frequent words in Python (version 3.7; Python Software Foundation) in the Spyder integrated development environment (version 4.1.5). The text entries were

XSL•FO
RenderX

prepared for the visualization in several preprocessing steps. Stop words (ie, words without a specific meaning such as *and* or *the*) were removed using an open-source German stop word list [28]; numerals and punctuation marks were removed as well. The remaining words were filtered to retain only nouns, adjectives, and verbs. Words with the same meaning were replaced manually by 1 unique word (eg, the word *work* replaced the following words: occupation, job, company, office, workplace, work hours, work colleagues, etc). The words were then lemmatized using the Python library *spaCy* (version 2.2.2) [29]. Lemmatization refers to the removal of inflectional endings to obtain the dictionary form of a word [30]. Vowels followed by an *e* (such as *oe*) were replaced by a vowel with an umlaut (in this specific case, *ö*) as commonly used in German. Capital letters were then replaced by lowercase letters. Finally, all lemmatized entries were translated into English using DeepL Pro (DeepL SE).

Moreover, we evaluated the individual-level similarity of health-related quality of life indices (EQ-5D-5L and EQ-VAS), symptoms, and medication use between the structured diary data and follow-up survey data (aim 3). Similarity was expressed by the Jaccard index, which reflects the proportion of persons with the concurrent mention of a specific attribute value (eg, a symptom) in the diary and the follow-up survey (overlap), divided by the number of persons with at least one mention of the same attribute value in either the diary or the follow-up survey (union) [31,32]. The Jaccard index was chosen over standard percentages because it is less affected by the rarity of an attribute value. In case a participant had several complete health diary entries, the health-related quality of life indices of the closest follow-up survey were used for the similarity analyses. The diary-based categorical symptom and medication use variables were compared with similar follow-up survey information on symptoms and medication use. The continuous EQ-5D-5L and EQ-VAS scores were dichotomized using median splits to calculate the Jaccard indices. The respective medians were computed from the follow-up survey of all included SMSR participants (n=658).

All descriptive analyses were conducted in R (version 4.0.3), using the RStudio integrated development environment (version 1.4.1103). Text preprocessing and word cloud visualization were performed using Python (version 3.7) in the Spyder integrated development environment (version 4.1.5). The dictionary-based text categorizations were performed using the LIWC software [24,25]. All text analyses were conducted using the original German text entries, and the results were subsequently translated into English for presentation using DeepL Pro.

Ethics Approval

The SMSR has been approved by the responsible ethics committee (Zurich Cantonal Ethics Committee; study number PB-2016-00894), and informed consent was obtained from all participants.

Results

Study Populations and Diary Use

After we had applied our inclusion and exclusion criteria, our study population comprised 658 persons with MS, of whom 134 (20.4%) had used the electronic health diary between February 27, 2019, and March 19, 2019 (Figure 1). During this period, 815 nonempty, unique diary entries written in German (n=632, 77.5%), French (n=135, 16.6%), and Italian (n=48, 5.9%) were collected (Multimedia Appendix 4).

The median time span between the last diary entry written by each diary campaign participant and the follow-up survey was 14 (IQR 13-15) months.

Participant Characteristics

Characteristics of the SMSR web-based study enrollees who participated in the campaign by completing at least one electronic health diary entry (n=134; henceforth referred to as *participants*) and those who did not (n=524; henceforth referred to as *nonparticipants*) are compared in Table 1. The median age of the participants (52.5, IQR 43.25-59.75 years) was higher than that of the nonparticipants (47, IQR 38-55 years). Of the 134 participants, 101 (75.4%) were women, whereas of the 524 nonparticipants, 356 (67.9%) were women. The median time since the diagnosis of MS was 9 (IQR 4-19 for participants and IQR 4-16 for nonparticipants) years for both groups. Persons with primary progressive MS (PPMS) and secondary progressive MS (SPMS) tended to be more represented among participants (PPMS: 18/134, 13.4%; SPMS: 32/134, 23.9%) than among nonparticipants (PPMS: 50/524, 9.5%; SPMS: 79/524, 15.1%). In addition, the participants' group contained proportionally fewer persons in the lowest Self-Reported Disability Status Scale stratum (84/134, 62.7%) than the nonparticipants' group (381/524, 72.7%).

The nonparticipants' group comprised a somewhat larger percentage of persons with higher professional education (including with an applied university or university degree; 250/524, 47.7%, vs 57/134, 42.5%, for participants). However, there were fewer people working >40% of weekly working hours among the participants (45/134, 33.6%, vs 252/524, 48.1%, for nonparticipants). Nonparticipants seemed to talk more openly about their MS with their relatives (485/524, 92.5%, vs 116/134, 86.6%, for participants).

Regarding symptoms experienced within the last 12 months, participants reported a higher percentage for the 10 most frequent symptoms. The largest differences in percentages were observed for the following symptoms: gait disorder (52/134, 38.8%, vs 138/524, 26.3%, for nonparticipants), spasms (45/134, 33.6%, vs 119/524, 22.7%, for nonparticipants), and bladder disorders (40/134, 29.9%, vs 109/524, 20.8%, for nonparticipants). Furthermore, a higher proportion of diary campaign participants reported no use of a disease-modifying drug within the last 6 months (53/134, 39.6%, vs 156/524, 29.8%, for nonparticipants). The full list of symptoms and disease-modifying medications is available in Multimedia Appendix 5.

XSL•FO
RenderX

Table 1. Characteristics of the study population and their corresponding *P* values (significance level: .05; N=658).[a]

Characteristics	Participants in health diary campaign (n=134)	Nonparticipants (n=524)	*P* value
Age (years), median (IQR)	52.5 (43.25-59.75)	47 (38-55)	<.001[b]
Sex, n (%)			.12[c]
Female	101 (75.4)	356 (67.9)	
Male	33 (24.6)	168 (32.1)	
Language, n (%)			<.001[c]
German	100 (74.6)	433 (82.6)	
French	25 (18.7)	73 (13.9)	
Italian	9 (6.7)	18 (3.5)	
MS[d] duration (years), median (IQR)	9 (4-19)	9 (4-16)	.17[b]
MS type, n (%)			<.001[c]
CIS[e]	2 (1.5)	10 (1.9)	
PPMS[f]	18 (13.4)	50 (9.5)	
RRMS[g]	79 (59)	352 (67.2)	
Transition	3 (2.2)	16 (3.1)	
SPMS[h]	32 (23.9)	79 (15.1)	
Missing information	N/A[i]	17 (3.2)	
Marital status, n (%)			<.001[c]
Unmarried	43 (32.1)	174 (33.2)	
Registered partnership or married	71 (53)	273 (52.1)	
Separated or divorced	17 (12.7)	59 (11.3)	
Widowed	1 (0.7)	8 (1.5)	
Missing information	2 (1.5)	10 (1.9)	
Education, n (%)			.59[c]
Partial or completed mandatory schooling	3 (2.2)	8 (1.5)	
Apprenticeship or qualification to study at university level (Matura diploma)	68 (50.8)	250 (47.7)	
Higher professional education, applied university, or university	57 (42.5)	250 (47.7)	
Other	1 (0.8)	7 (1.4)	
Missing information	5 (3.7)	9 (1.7)	
Work situation or work percentage, n (%)			.02[c]
Not working	63 (47)	184 (35.1)	
1% to 40%	22 (16.4)	69 (13.2)	
41% to 80%	27 (20.2)	138 (26.3)	
81% to full time	18 (13.4)	114 (21.8)	
Missing information	4 (3)	19 (3.6)	
Talk about MS with...[j]			.87[c]
...relatives	116 (86.6)	485 (92.6)	
...friends	109 (81.3)	412 (78.6)	
...boss	48 (35.8)	179 (34.2)	

XSL•FO
RenderX

Characteristics	Participants in health diary campaign (n=134)	Nonparticipants (n=524)	*P* value
...work colleagues	46 (34.3)	189 (36.1)	
...leisure-time partners	36 (26.9)	131 (25)	
...other	7 (5.2)	23 (4.4)	
...nobody	6 (4.5)	13 (2.5)	
SRDSS[k] score, n (%)			.02[c]
0 to 3.5	84 (62.7)	381 (72.7)	
4 to 6.5	37 (27.6)	90 (17.2)	
≥7	13 (9.7)	39 (7.4)	
Missing information	N/A	14 (2.7)	
Ten most frequent symptoms within the last 12 months[j,l], n (%)			.42[c]
None	50 (37.3)	227 (43.3)	
Fatigue	57 (42.5)	210 (40.1)	
Gait disorder	52 (38.8)	138 (26.3)	
Paresthesia (eg, numbness and tingling)	50 (37.3)	187 (35.7)	
Spasms (muscle cramps)	45 (33.6)	119 (22.7)	
Vestibular disorders	43 (32.1)	135 (25.8)	
Weakness	40 (29.9)	129 (24.6)	
Pain	40 (29.9)	143 (27.3)	
Bladder disorders (eg, bladder weakness)	40 (29.9)	109 (20.8)	
Concentration problems	35 (26.1)	127 (24.2)	
Intestinal disorders (eg, constipation)	29 (21.6)	83 (15.8)	
Other	64 (47.8)	193 (36.8)	
Ten most frequent disease-modifying medications within the last 6 months[j,l], n (%)			.20[c]
None	53 (39.6)	156 (29.8)	
Ocrelizumab (Ocrevus)	28 (20.9)	87 (16.6)	
Fingolimod (Gilenya)	15 (11.2)	90 (17.2)	
Dimethyl fumarate (Tecfidera)	10 (7.5)	54 (10.3)	
Interferon beta-1a (Rebif)	6 (4.5)	19 (3.6)	
Glatiramer acetate (Copaxone)	6 (4.5)	18 (3.4)	
Interferon beta-1b (Betaferon)	5 (3.7)	14 (2.7)	
Teriflunomide (Aubagio)	5 (3.7)	15 (2.9)	
Natalizumab (Tysabri)	3 (2.2)	40 (7.6)	
Rituximab (MabThera)	3 (2.2)	15 (2.9)	
Interferon beta-1a (Avonex)	2 (1.5)	10 (1.9)	

XSL•FO
RenderX

Characteristics	Participants in health diary campaign (n=134)	Nonparticipants (n=524)	*P* value
Other	1 (0.7)	10 (1.9)	

[a]Comparison between participants in the Swiss Multiple Sclerosis Registry electronic health diary campaign (defined by having at least one valid diary entry; column 2) and nonparticipants (column 3).

[b]Two-tailed *t* test conducted.

[c]Chi-square test conducted.

[d]MS: multiple sclerosis.

[e]CIS: clinically isolated syndrome.

[f]PPMS: primary progressive multiple sclerosis.

[g]RRMS: relapsing-remitting multiple sclerosis.

[h]SPMS: secondary progressive multiple sclerosis.

[i]N/A: not applicable.

[j]Multiple answers possible.

[k]SRDSS: Self-Reported Disability Status Scale.

[l]Full list of symptoms and disease-modifying medications available in Multimedia Appendix 5.

Health Diary Use Patterns

The use of the electronic health diary was monitored over 21 days between February 27, 2019, and March 19, 2019 (Multimedia Appendix 3). Over these 21 days, a total of 815 unique electronic health diary entries were collected from 134 diary campaign participants. Of note, more health diary entries were collected on days when email invitations or reminders were sent than on the adjacent days; for example, the final reminder sent on March 14, 2019 (bar marked by a red numeral 3 in Multimedia Appendix 3), displayed the largest number of entries (n=130) of the whole observation period. In addition, the electronic health diary was used more intensively during the official health diary campaign period (striped bars in Multimedia Appendix 3).

As shown in Table 2, 23.1% (31/134) of the participants wrote only 1 entry in the diary, and 11.2% (15/134) wrote 2. Moreover, 15.7% (21/134) of the participants had ≥10 entries (reflecting ≥1 entry every 2 days). The highest number of entries written by 1 person was 32 (including retrospective entries made during the analysis time frame). Multimedia Appendix 6 summarizes the completeness of health diary entries.

Table 2. Frequency counts and percentages of participants totalizing a certain number of diary health entries written during the campaign (N=134).

Number of entries in the health diary per participant	Participants, n (%)
1	31 (23.1)
2	15 (11.2)
3	6 (4.5)
4	5 (3.7)
5	6 (4.5)
6	6 (4.5)
7	9 (6.7)
8	16 (11.9)
9	19 (14.2)
10	6 (4.5)
11	5 (3.7)
12	2 (1.5)
13	1 (0.7)
15	2 (1.5)
17	1 (0.7)
18	1 (0.7)
19	1 (0.7)
20	1 (0.7)
32	1 (0.7)

XSL•FO
RenderX

Content Analysis of the Textual Diary Data

The 134 electronic health diary campaign participants provided a total of 815 electronic health diary entries (in German: n=632, 77.5%; in French: n=135, 16.6%; and in Italian: n=48, 5.9%). For the subsequent analysis, we only used the German health diary entries, provided by 100 participants. The emotional scores of the 632 diary entries were low overall, reflecting negative emotions such as very anxious, sad, or hostile tone [33]. Of the 632 diary entries, 163 (25.8%) scored between 0 and 10, and 420 (66.5%) scored between 20 and 30. Furthermore, positive emotions were expressed in only 7.6% (48/632) of the entries, whereas words related to negative emotions were used in 30.5% (193/632) of the entries (Table 3). Other frequent LIWC-based topic categories pertained to the body (191/632, 30.2%), health (94/632, 14.9%), and work (67/632, 10.6%).

Of the 632 German electronic health diary entries, 526 (83.2%; written by 93 participants) contained at least 10 words and were used to create a 100-word word cloud (Multimedia Appendix 7; refer to Multimedia Appendix 8 for the frequency of the 25 most common words). As shown in Multimedia Appendix 7 and Multimedia Appendix 8, the words “good” and “day” occur most often, followed by “work,” “up,” “go,” “sleep,” and “tired.” Among the 25 most frequent words, some refer to time, including “day,” “evening,” “morning,” “hour,” “time,” and “afternoon.” Other words pertained to activities or movement, such as “work,” “go,” “walk,” “make,” and “errand.” Several frequent words also related to body parts, physical well-being, and health issues; for example, “tired,” “leg,” “pain,” “therapy,” and “feel.” Although comparatively fewer words were associated with social contacts, some were very common in diary entries, such as “contact” or “family.” A word cloud and the frequency of the most common words for each MS type (ie, PPMS, relapsing-remitting multiple sclerosis, transition, and SPMS) are available in Multimedia Appendix 9.

Table 3. Frequency counts and percentages of the diary entries containing relevant Linguistic Inquiry and Word Count word categories (N=632).

Characteristics	Values, n (%)
Negative emotion	193 (30.5)
Body	191 (30.2)
Health	94 (14.9)
Work	67 (10.6)
Positive emotion	48 (7.6)
Friends	10 (1.6)
Family	8 (1.3)

Similarities Between Structured Electronic Health Diary Data and Questionnaire-Based Data

The responses to the electronic health diary and the follow-up survey exhibited several notable differences and similarities quantified by the Jaccard index (Table 4). The EQ-5D-5L (Jaccard indices: 0.73 and 0.63 below and above the median, respectively) and the EQ-VAS (Jaccard indices: 0.59 and 0.49, respectively) presented the highest Jaccard indices. Furthermore, the electronic health diary and the follow-up survey methods converged for the evaluation of gait disorder, signs of paralysis, and fatigue (Jaccard indices: 0.5, 0.5, and 0.49, respectively). By contrast, several symptoms, including pain, vestibular disorder, weakness, spasms, bladder disorders, and concentration problems displayed comparatively lower Jaccard indices (<0.4), despite being reported quite frequently (n>45; refer to the column *Union* in Table 4 and the *Ten most frequent symptoms within the last 12 months* section in Table 1) and at least once in the diary and the survey. The injectable immunomodulatory drugs interferon beta-1a (Avonex) and glatiramer acetate (Copaxone), as well as the oral drugs fingolimod (Gilenya) and teriflunomide (Aubagio), exhibited the highest similarities across both assessment methods (Jaccard indices between 0.5 and 1; refer to Multimedia Appendix 10 for the complete list of disease-modifying medications). Of note, despite currently being used frequently, ocrelizumab (Ocrevus) was not mentioned in the diary.

XSL•FO
RenderX

Table 4. Similarity between diary and follow-up survey information on health-related quality of life, symptoms, and medication use provided by 134 electronic health diary campaign participants.[a]

Characteristics	Overlap	Union	Jaccard index
EQ-5D-5L			
EQ-5D-5L: ≤0.71 (overall median)	52	71	0.73
EQ-5D-5L: >0.71	33	52	0.63
EQ-VAS[b]			
EQ-VAS: ≤80 (overall median)	34	58	0.59
EQ-VAS: >80	23	47	0.49
Symptoms			
Gait disorder	33	66	0.50
Signs of paralysis	12	24	0.50
Fatigue	39	80	0.49
Paresthesia (eg, numbness and tingling)	34	73	0.47
Tremor	8	18	0.44
Intestinal disorders (eg, constipation)	14	34	0.41
Weakness	22	58	0.38
Vestibular disorders	23	62	0.37
Pain	22	63	0.35
Affective lability and lack of control over emotions	5	15	0.33
Spasms (muscle cramps)	18	56	0.32
Memory disorders	9	28	0.32
Bladder disorders (eg, bladder weakness)	16	51	0.31
Convulsions and tics	5	17	0.29
Sexual disorders	5	19	0.26
Visual impairments	7	30	0.23
Problems with spatial orientation	7	30	0.23
Dizziness	7	32	0.22
Speech disorders	4	20	0.20
Swallowing difficulties (dysphagia)	3	15	0.20
Depression	3	16	0.19
Concentration problems	7	46	0.15
Other	0	10	0
Epileptic convulsions	0	1	0
Ten most frequent disease-modifying medications [c]			
Interferon beta-1a (Avonex)	2	2	1
Glatiramer acetate (Copaxone)	5	7	0.71
Fingolimod (Gilenya)	12	21	0.57
Teriflunomide (Aubagio)	4	8	0.50
Interferon beta-1b (Betaferon)	2	5	0.40
Dimethyl fumarate (Tecfidera)	1	32	0.03
Ocrelizumab (Ocrevus)	0	36	0
Other	0	33	0
Interferon beta-1a (Rebif)	0	6	0

XSL•FO
RenderX

Characteristics	Overlap	Union	Jaccard index
Natalizumab (Tysabri)	0	6	0
Rituximab (MabThera)	0	3	0

[a]The Jaccard index was used to measure the similarities (reporting overlap) among the different sources (0=no similarity and 1=maximum similarity). The column *Overlap* represents the number of participants with overlapping reports in the diary and in the follow-up survey. The column *Union* represents the total number of participants who reported a specific data item at least once in the diary or the follow-up survey collecting the symptoms experienced within the last 12 months and the medication used within the last 6 months. The column *Jaccard index* represents the quotient of the *Overlap* values and the *Union* values. The EQ-5D-5L index and the EuroQol visual analog scale index were dichotomized using the respective medians of all included Swiss Multiple Sclerosis Registry enrollees with completed follow-up survey (n=658).

[b]EQ-VAS: EuroQol visual analog scale.

[c]The full list of disease-modifying medications is available in Multimedia Appendix 10.

Discussion

Principal Findings

On the basis of a health diary campaign nested into the SMSR, our study sought to explore the factors associated with electronic health diary use as well as the added value of diary entries vis-à-vis survey-collected information. By comparing the characteristics of 134 health diary campaign participants with those of 524 nonparticipants, our study revealed that women, persons with a more advanced disease state, those with a lower work percentage, or those without higher education (eg, a university degree) were more likely to devote time to the electronic health diary.

Contrary to our initial expectations, electronic health diary participants were on average 5 years older than the nonparticipants, more likely to be women, less likely to have completed higher education, and more frequently had an advanced Self-Reported Disability Status Scale score (≥4). In addition, a larger proportion of participants reported experiencing diverse MS symptoms within the last 12 months, in particular from more strongly hindering symptoms such as fatigue, gait disorder, spasms, and vestibular disorders. Not only are these symptoms disabling, but they may also lead to a more confined life at home. These findings may seem counterintuitive at first glance because studies comparing characteristics between web-based and paper-and-pencil study participants (including a study from the SMSR) usually found younger age, higher education, and less disease severity to be associated with a higher probability of web-based participation [34,35]. However, this study focused entirely on web-based participants. Thus, the stratification by health diary use reflects a further segmentation of this population. In our analysis, persons with a higher symptom burden and not working full time, which are two factors associated with older age, were more likely to contribute to the electronic health diary campaign. This finding overlaps with an independent SMSR analysis showing that, based on a set of standardized questions, persons with a high affinity for using digital tools for MS management also tended to be older (but still middle-aged) and to have a higher disease burden (Nittas, unpublished data, March 2022). It can be speculated that an elevated disease burden seems to come with more time at disposal and a stronger desire to tell one's story as well as to understand one's disease.

When analyzing the content of health diary entries, participants' reports mainly addressed themes related to the body, health, and work, as confirmed by the LIWC and word cloud analytic approaches. Moreover, the frequent mention of words such as "sleep" or "tired" corresponds well with frequent reports of fatigue in the diary as well as in the SMSR follow-up survey (Table 1).

A substantial number of participant reports had a negative emotional coloration and were referring to, for example, the adverse effects of MS on their body and daily life activities. Indeed, keywords from the LIWC categories *negative emotion* and *body* appeared in the same proportion of diary entries and were thus possibly co-occurring. Diary participants seemed to rather express concerns—possibly about their disease affecting their body and their active life. Whereas a larger proportion of the nonparticipants talk about MS to their relatives, diary participants seemed to confide their worries in the diary [36].

Of further note, the comparison of structured information from the electronic health diary and the follow-up survey generally exhibited good overlaps for health-related quality of life aspects, relatively stable and frequent symptoms, and—to a lesser extent—for medications with a frequent intake schedule (eg, injectable or oral drugs). The health-related quality of life index EQ-5D-5L revealed the highest overlap, which is consistent with the general notion that this indicator is not very responsive to smaller health status changes. The more subjective EQ-VAS showed somewhat smaller overlaps, possibly because the EQ-VAS is more responsive to daily fluctuations [9]. The Jaccard index was also relatively high for stable or slow-changing symptoms such as gait disorder, signs of paralysis, and fatigue. By contrast, the similarity analysis of medication use mostly yielded low Jaccard indices, suggesting little overlap. Nonetheless, older, established injection therapies for relapsing-remitting MS, such as interferon beta-1a (Avonex) [37] and glatiramer acetate (Copaxone) [38], although not used by a large proportion of our study participants, displayed the highest Jaccard indices, demonstrating little change in medication. Furthermore, more modern and convenient treatments such as fingolimod (Gilenya) and teriflunomide (Aubagio), which are two oral therapies with daily intake for relapsing-remitting MS, also exhibited relatively high Jaccard indices. Nonetheless, other modern MS drugs administered in monthly (natalizumab [Tysabri]) infusions at clinics demonstrated low Jaccard indices because of the diary instructions that asked participants for MS medication in the previous day only. Furthermore, the frequently used infusion

XSL•FO
RenderX

drug ocrelizumab (Ocrevus) was only widely prescribed in Switzerland after the health diary campaign ended [39].

To the best of our knowledge, this is the first study to assess the similarities between data collected through a web-based survey and data collected by means of an electronic diary. The different person and disease characteristics associated with electronic health diary use can help refine the target population for implementing electronic health diary tools in research contexts. Furthermore, the findings suggest that health diaries can form a meaningful complement to regular surveys by providing additional insights into daily life topics and timely updates on health-related quality of life or symptom and medication status of persons with MS. Although ecological momentary assessment research has highlighted the possible risk of retrospection bias in retrospectively reporting symptoms averaged over longer times [40], subjective perceptions of illnesses vary strongly and meaningfully across people [41]. Indeed, evaluating subjective views with regard to, for example, symptoms of people with MS not only adds value to standardized assessments pertaining to personalized treatment planning but also helps to explain adherence to treatment regimens [42]. Moreover, free-text questions as implemented in our diary can complement regular standardized assessments by providing additional levels of detail. Furthermore, being given the opportunity to report on their experiences in their own words is often appreciated by study participants and, as suggested by the extensive expressive writing literature, may even be of therapeutical value in itself [43]. Our electronic health diary combines the known advantages of ecological momentary assessment with the advantage of not restricting answers to forced, given choices.

The study was also informative on a methodological level. Although the health diary has been part of the SMSR platform since its launch, voluntary use was very low, on average <1 entry per day per participant. By contrast, we noticed that the use of the electronic health diary was enhanced after motivational emails were sent to the participants. These experiences suggest that health diary studies may benefit from being embedded into a campaign with clear aims and a limited time frame, as revealed by previous studies [44,45].

Strengths and Limitations

Our study benefited from a large, diverse, well-documented study base for enrollment [18,19]. Moreover, this is the first health diary study in MS to use a blend of different analytical approaches (descriptive statistics and natural language processing) to glean insights into health diary use patterns and daily life aspects; for example, closer inspection of diary entries revealed novel aspects such as the stress imposed by application procedures for disability insurance or individual coping strategies for well-being by persons with MS (not shown). Therefore, diary studies harbor a significant untapped potential for hypothesis generation and inspiration for research topics.

However, some limitations should be noted. First, our study excluded persons who preferred paper-and-pencil surveys. This excluded group reflects a population with more advanced disease states and possibly lower digital literacy (Nittas, unpublished data, March 2022). However, similar to our study, earlier investigations also observed that at an early stage of the disease, persons with MS are less likely to engage in research studies (Nittas, unpublished data, March 2022), [46], hypothetically because they are in denial regarding their disease or do not entirely realize the scope of it yet. In addition, the distance in time between the health diary campaign and the follow-up survey completion was relatively long. Hence, the MS symptomatology, as well as treatment strategies, may have changed in the meantime (eg, ocrelizumab [Ocrevus] has started to be widely prescribed in Switzerland after the end of the health diary campaign). It is important to mention that the results could be replicated using different comparison data provided by the SMSR. Future research would benefit from further investigating the added value of such free-text health diaries by comparing entries with standardized instruments of physical and mental well-being assessed in parallel. Furthermore, among the 658 eligible web-based SMSR participants, only a fraction (n=134, 20.4%) took part in the diary study, thus limiting the generalizability of the results. Besides, the vast majority (113/134, 84.3%) of the participants made <1 entry every 2 days in the diary, and 34% (46/134) contributed only once or twice to the diary campaign. This lack of regularity in diary use was also a limitation to our study. The implementation of a diary with automated reminders [47] might be a promising way forward to increase participation [12,14]. Nevertheless, these limitations observed in our study may also hinder a broader application of electronic diaries in ecological momentary assessments.

Therefore, a better understanding of participant needs, motivational factors, and the effectiveness of incentives is urgently needed to enable a broader application and long-term use of health diaries in health research and disease management. In addition, in light of recent advances in fields such as natural language processing [48] and speech recognition [49] (eg, Hugging Face [50]) based on machine learning or wearable sensor technologies (eg, fitness trackers) [42,51,52], future studies should examine how health diaries could be optimally combined with novel technologies.

Conclusions

To summarize, our study suggests that health diaries can be a valuable complement to regular, structured questionnaires in the context of MS research. However, they should ideally be embedded into a campaign with motivational activities such as email reminders or regular data feedback. Our findings further suggest that a topical focus on daily life aspects, health-related quality of life, and stable symptoms elicited more similar responses to standardized assessments and were thus less informative than medication diaries. Hence, medication diaries for daily dispensed (oral or injectable) drugs may offer opportunities for drug intake compliance tracking and unwanted drug effect occurrence monitoring.

XSL•FO
RenderX

Acknowledgments

The authors are most grateful to all participants of the Swiss Multiple Sclerosis Registry (SMSR) who dedicated their time and thereby made a crucial contribution to this research. The authors would further like to thank the participant board of the SMSR for their contribution to the development and validation of a multiple sclerosis–related dictionary. Furthermore, the authors' gratitude goes to the Swiss Multiple Sclerosis Society for funding the SMSR and for its continuous support. The members of the SMSR are Bernd Anderseck, PC, Andrew Chan, Claudio Gobbi, Roger Häussler, CPK, Jürg Kesselring (President), Jens Kuhle (Chair of the Clinical and Laboratory Research Committee), Roland Kurmann, Christoph Lotter, Kurt Luyckx, Patricia Monin, Stefanie Müller, Krassen Nedeltchev, CP, Milo A Puhan, Irene Rapold, Anke Salmen, Klaas Enno Stephan, Zina-Mary Manjaly, Claude Vaney (Chair of Patient and Population Research Committee), VvW (Chair of the IT and Data Committee), and CZ. The SMSR is supported by the scientific advisory board of the Swiss Multiple Sclerosis Society.

Conflicts of Interest

Ente Ospedaliero Cantonale (CZ's employer) received compensation for CZ's speaking activities, consulting fees, or research grants from Almirall, Biogen Idec, Bristol Meyer Squibb, Genzyme, Lundbeck, Merck, Novartis, Teva Pharma, and Roche.

Multimedia Appendix 1
Electronic health diary of the Swiss Multiple Sclerosis Registry.
[DOCX File , 764 KB - mhealth_v10i10e38709_app1.docx]

Multimedia Appendix 2
Invitation letters to the electronic health diary campaign for the Swiss Multiple Sclerosis Registry participants.
[DOCX File , 38 KB - mhealth_v10i10e38709_app2.docx]

Multimedia Appendix 3
Number of entries made per day in the electronic health diary between February 27, 2019, and March 19, 2019 (N=815). The striped bars (March 9, 2019, to March 17, 2019) correspond to the official health diary campaign period. The red numbers indicate when the Swiss Multiple Sclerosis Society sent emails to its participants: (1) invitation email for the campaign sent out to all Swiss Multiple Sclerosis Registry participants and information released as a news item on the public web page of the Swiss Multiple Sclerosis Society, (2) reminder email sent out to all participants informing them that the campaign was about to start, and (3) motivational email that provided some statistical information about the health diary campaign and reminded participants to contribute.
[PNG File , 93 KB - mhealth_v10i10e38709_app3.png]

Multimedia Appendix 4
Flowchart of the Swiss Multiple Sclerosis Registry enrollees and the electronic health diary participants.
[DOCX File , 432 KB - mhealth_v10i10e38709_app4.docx]

Multimedia Appendix 5
Full list of the symptoms experienced within the last 12 months and the disease-modifying medications used within the last 6 months.
[DOCX File , 41 KB - mhealth_v10i10e38709_app5.docx]

Multimedia Appendix 6
Completeness of the diary entries.
[DOCX File , 38 KB - mhealth_v10i10e38709_app6.docx]

Multimedia Appendix 7
Word cloud of the 100 most frequent words appearing in the electronic health diary entries of a minimum length of 10 words (n=526). The font size reflects the frequency of the words' occurrence.
[PNG File , 322 KB - mhealth_v10i10e38709_app7.png]

Multimedia Appendix 8
Word frequency of the 25 most frequent words of the word cloud.
[DOCX File , 78 KB - mhealth_v10i10e38709_app8.docx]

XSL•FO
RenderX

Multimedia Appendix 9
Word cloud of 100 words and word frequency of the 25 most frequent words throughout the electronic health diary entries per multiple sclerosis type.
[DOCX File , 994 KB - mhealth_v10i10e38709_app9.docx]

Multimedia Appendix 10
Jaccard indices of the full disease-modifying medication list.
[DOCX File , 16 KB - mhealth_v10i10e38709_app10.docx]

References

1. Verbrugge LM. Health diaries. Med Care 1980 Jan;18(1):73-95. [doi: 10.1097/00005650-198001000-00006] [Medline: 6986517]
2. Al-Ubaydli M, Paton C. The doctor's PDA and smartphone handbook. Personal digital assistant. J R Soc Med 2005 Nov;98(11):494-495 [FREE Full text] [Medline: 16260797]
3. Pennebaker JW. Putting stress into words: health, linguistic, and therapeutic implications. Behav Res Ther 1993 Jul;31(6):539-548. [doi: 10.1016/0005-7967(93)90105-4] [Medline: 8347112]
4. Pennebaker JW, Barger SD, Tiebout J. Disclosure of traumas and health among Holocaust survivors. Psychosom Med 1989;51(5):577-589. [doi: 10.1097/00006842-198909000-00009] [Medline: 2798704]
5. Smyth JM. Written emotional expression: effect sizes, outcome types, and moderating variables. J Consult Clin Psychol 1998 Feb;66(1):174-184. [doi: 10.1037//0022-006x.66.1.174] [Medline: 9489272]
6. Frattaroli J. Experimental disclosure and its moderators: a meta-analysis. Psychol Bull 2006 Nov;132(6):823-865. [doi: 10.1037/0033-2909.132.6.823] [Medline: 17073523]
7. Sebastião E, McAuley E, Shigematsu R, Adamson BC, Bollaert RE, Motl RW. Home-based, square-stepping exercise program among older adults with multiple sclerosis: results of a feasibility randomized controlled study. Contemp Clin Trials 2018 Oct;73:136-144. [doi: 10.1016/j.cct.2018.09.008] [Medline: 30243811]
8. Freeman J, Hendrie W, Jarrett L, Hawton A, Barton A, Dennett R, et al. Assessment of a home-based standing frame programme in people with progressive multiple sclerosis (SUMS): a pragmatic, multi-centre, randomised, controlled trial and cost-effectiveness analysis. Lancet Neurol 2019 Aug;18(8):736-747 [FREE Full text] [doi: 10.1016/S1474-4422(19)30190-5] [Medline: 31301748]
9. Parkin D, Rice N, Jacoby A, Doughty J. Use of a visual analogue scale in a daily patient diary: modelling cross-sectional time-series data on health-related quality of life. Soc Sci Med 2004 Jul;59(2):351-360. [doi: 10.1016/j.socscimed.2003.10.015] [Medline: 15110425]
10. Bove R, Secor E, Healy BC, Musallam A, Vaughan T, Glanz BI, et al. Evaluation of an online platform for multiple sclerosis research: patient description, validation of severity scale, and exploration of BMI effects on disease course. PLoS One 2013;8(3):e59707 [FREE Full text] [doi: 10.1371/journal.pone.0059707] [Medline: 23527256]
11. Jongen PJ, Sinnige LG, van Geel BM, Verheul F, Verhagen WI, van der Kruijk RA, et al. The interactive web-based program MSmonitor for self-management and multidisciplinary care in multiple sclerosis: concept, content, and pilot results. Patient Prefer Adherence 2015;9:1741-1750 [FREE Full text] [doi: 10.2147/PPA.S93783] [Medline: 26715841]
12. Golan D, Sagiv S, Glass-Marmor L, Miller A. Mobile-phone-based e-diary derived patient reported outcomes: association with clinical disease activity, psychological status and quality of life of patients with multiple sclerosis. PLoS One 2021;16(5):e0250647 [FREE Full text] [doi: 10.1371/journal.pone.0250647] [Medline: 33951061]
13. Jongen PJ, Sinnige LG, van Geel BM, Verheul F, Verhagen WI, van der Kruijk RA, et al. The interactive web-based program MSmonitor for self-management and multidisciplinary care in multiple sclerosis: utilization and valuation by patients. Patient Prefer Adherence 2016;10:243-250 [FREE Full text] [doi: 10.2147/PPA.S93786] [Medline: 27042018]
14. Zettl UK, Bauer-Steinhusen U, Glaser T, Czekalla J, Hechenbichler K, Limmroth V, et al. Adherence to long-term interferon beta-1b injection therapy in patients with multiple sclerosis using an electronic diary. Adv Ther 2016 May;33(5):834-847. [doi: 10.1007/s12325-016-0325-6] [Medline: 27090116]
15. Richardson A. The health diary: an examination of its use as a data collection method. J Adv Nurs 1994 Apr;19(4):782-791. [doi: 10.1111/j.1365-2648.1994.tb01151.x] [Medline: 8021401]
16. Gajofatto A, Benedetti MD. Treatment strategies for multiple sclerosis: when to start, when to change, when to stop? World J Clin Cases 2015 Jul 16;3(7):545-555 [FREE Full text] [doi: 10.12998/wjcc.v3.i7.545] [Medline: 26244148]
17. Ciotti JR, Cross AH. Disease-modifying treatment in progressive multiple sclerosis. Curr Treat Options Neurol 2018 Apr 07;20(5):12. [doi: 10.1007/s11940-018-0496-3] [Medline: 29627873]
18. Puhan MA, Steinemann N, Kamm CP, Müller S, Kuhle J, Kurmann R, Swiss Multiple Sclerosis Registry Smsr. A digitally facilitated citizen-science driven approach accelerates participant recruitment and increases study population diversity. Swiss Med Wkly 2018;148:w14623 [FREE Full text] [doi: 10.4414/smw.2018.14623] [Medline: 29767828]
19. Steinemann N, Kuhle J, Calabrese P, Kesselring J, Disanto G, Merkler D, Swiss Multiple Sclerosis Registry. The Swiss Multiple Sclerosis Registry (SMSR): study protocol of a participatory, nationwide registry to promote epidemiological and

XSL•FO
RenderX

patient-centered MS research. BMC Neurol 2018 Aug 13;18(1):111 [FREE Full text] [doi: 10.1186/s12883-018-1118-0] [Medline: 30103695]
20. Barin L, Salmen A, Disanto G, Babačić H, Calabrese P, Chan A, Swiss Multiple Sclerosis Registry (SMSR). The disease burden of Multiple Sclerosis from the individual and population perspective: which symptoms matter most? Mult Scler Relat Disord 2018 Oct;25:112-121. [doi: 10.1016/j.msard.2018.07.013] [Medline: 30059895]
21. Matter-Walstra K, Klingbiel D, Szucs T, Pestalozzi BC, Schwenkglenks M. Using the EuroQol EQ-5D in Swiss cancer patients, which value set should be applied? Pharmacoeconomics 2014 Jun;32(6):591-599. [doi: 10.1007/s40273-014-0151-0] [Medline: 24671924]
22. Perneger TV, Combescure C, Courvoisier DS. General population reference values for the French version of the EuroQol EQ-5D health utility instrument. Value Health 2010 Aug;13(5):631-635 [FREE Full text] [doi: 10.1111/j.1524-4733.2010.00727.x] [Medline: 20412541]
23. McGinley MP, Goldschmidt CH, Rae-Grant AD. Diagnosis and treatment of multiple sclerosis: a review. JAMA 2021 Feb 23;325(8):765-779. [doi: 10.1001/jama.2020.26858] [Medline: 33620411]
24. Pennebaker J, Boyd R, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. University of Texas. 2015. URL: https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf [accessed 2022-09-19]
25. Welcome to LIWC-22. LIWC. URL: https://www.liwc.app/ [accessed 2022-08-26]
26. Meier T, Boyd R, Pennebaker J, Mehl M, Martin M, Wolf M, et al. "LIWC auf Deutsch": the development, psychometrics, and introduction of DE- LIWC2015. PsyArXiv 2019 Feb 14. [doi: 10.31234/osf.io/uq8zt]
27. Cohn MA, Mehl MR, Pennebaker JW. Linguistic markers of psychological change surrounding September 11, 2001. Psychol Sci 2004 Oct;15(10):687-693. [doi: 10.1111/j.0956-7976.2004.00741.x] [Medline: 15447640]
28. Diaz G. GitHub - stopwords-iso/stopwords-de: German stopwords collection. GitHub. 2020 Mar 6. URL: https://github.com/stopwords-iso/stopwords-de [accessed 2022-07-23]
29. Honnibal M, Montani I. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Spacy. 2017. URL: http://citebay.com/how-to-cite/spacy/ [accessed 2022-07-23]
30. Ahn I, Na W, Kwon O, Yang DH, Park G, Gwon H, et al. CardioNet: a manually curated database for artificial intelligence-based research on cardiovascular diseases. BMC Med Inform Decis Mak 2021 Jan 28;21(1):29 [FREE Full text] [doi: 10.1186/s12911-021-01392-2] [Medline: 33509180]
31. Jaccard P. The distribution of the flora in the alpine zone. New Phytol 1912 Feb;11(2):37-50. [doi: 10.1111/j.1469-8137.1912.tb05611.x]
32. Rhee S, Liu TF, Holmes SP, Shafer RW. HIV-1 subtype B protease and reverse transcriptase amino acid covariation. PLoS Comput Biol 2007 May;3(5):e87 [FREE Full text] [doi: 10.1371/journal.pcbi.0030087] [Medline: 17500586]
33. Pennebaker J, Booth R, Boyd R, Francis M. Linguistic inquiry and word count: LIWC2015. Pennebaker Conglomerates. 2015. URL: https://www.researchgate.net/publication/337731895_Linguistic_Inquiry_and_Word_Count_LIWC2015 [accessed 2022-09-19]
34. Potdar R, Thomas A, DiMeglio M, Mohiuddin K, Djibo DA, Laudanski K, et al. Access to internet, smartphone usage, and acceptability of mobile health technology among cancer patients. Support Care Cancer 2020 Nov;28(11):5455-5461. [doi: 10.1007/s00520-020-05393-1] [Medline: 32166381]
35. Fox G, Connolly R. Mobile health technology adoption across generations: narrowing the digital divide. Info Systems J 2018 Jan 29;28(6):995-1019. [doi: 10.1111/isj.12179]
36. Grose J, Freeman J, Skirton H. Value of a confidant relationship in psychosocial care of people with multiple sclerosis. Int J MS Care 2012;14(3):115-122 [FREE Full text] [doi: 10.7224/1537-2073-14.3.115] [Medline: 24453742]
37. Why Avonex. Avonex. URL: https://www.avonex.com/en_us/home/why-avonex/about.html [accessed 2022-07-24]
38. Wynn DR. Enduring clinical value of Copaxone® (Glatiramer Acetate) in multiple sclerosis after 20 years of use. Mult Scler Int 2019;2019:7151685 [FREE Full text] [doi: 10.1155/2019/7151685] [Medline: 30775037]
39. Roche's OCREVUS (ocrelizumab) approved in Switzerland for primary progressive and relapsing forms of multiple sclerosis. Roche. 2017 Sep 28. URL: https://www.roche.ch/en/media-switzerland/informationen/med-ch-2017-09-28.htm [accessed 2022-07-28]
40. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
41. Petrie KJ, Jago LA, Devcich DA. The role of illness perceptions in patients with medical conditions. Curr Opin Psychiatry 2007 Mar;20(2):163-167. [doi: 10.1097/YCO.0b013e328014a871] [Medline: 17278916]
42. Wendrich K, van Oirschot P, Martens MB, Heerings M, Jongen PJ, Krabbenborg L. Toward digital self-monitoring of multiple sclerosis: investigating first experiences, needs, and wishes of people with MS. Int J MS Care 2019;21(6):282-291 [FREE Full text] [doi: 10.7224/1537-2073.2018-083] [Medline: 31889935]
43. Pennebaker J, Chung CK. Expressive writing, emotional upheavals, and health. In: Foundations of Health Psychology. Oxford, UK: Oxford University Press; 2007.

XSL•FO
RenderX

44. Greenhalgh J, Ford H, Long AF, Hurst K. The MS Symptom and Impact Diary (MSSID): psychometric evaluation of a new instrument to measure the day to day impact of multiple sclerosis. J Neurol Neurosurg Psychiatry 2004 Apr;75(4):577-582 [FREE Full text] [doi: 10.1136/jnnp.2003.020529] [Medline: 15026500]
45. Hellard ME, Sinclair MI, Forbes AB, Fairley CK. A randomized, blinded, controlled trial investigating the gastrointestinal health effects of drinking water quality. Environ Health Perspect 2001 Aug;109(8):773-778 [FREE Full text] [doi: 10.1289/ehp.01109773] [Medline: 11564611]
46. Zhao Y, Ni Q, Zhou R. What factors influence the mobile health service adoption? A meta-analysis and the moderating role of age. Intl J Inform Manag 2018 Dec;43:342-350. [doi: 10.1016/j.ijinfomgt.2017.08.006]
47. Stone AA, Shiffman S, Schwartz JE, Broderick JE, Hufford MR. Patient compliance with paper and electronic diaries. Control Clin Trials 2003 Apr;24(2):182-199. [doi: 10.1016/s0197-2456(02)00320-3] [Medline: 12689739]
48. Zhang A, Xing L, Zou J, Wu JC. Shifting machine learning for healthcare from development to deployment and from models to data. Nat Biomed Eng 2022 Jul 04 (forthcoming). [doi: 10.1038/s41551-022-00898-y] [Medline: 35788685]
49. Koops S, Brederoo SG, de Boer JN, Nadema FG, Voppel AE, Sommer IE. Speech as a biomarker for depression. CNS Neurol Disord Drug Targets 2021 Dec 13 (forthcoming). [doi: 10.2174/1871527320666211213125847] [Medline: 34961469]
50. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020 Presented at: 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; Oct, 2020; Online. [doi: 10.18653/v1/2020.emnlp-demos.6]
51. Thorpe J, Forchhammer BH, Maier AM. Adapting mobile and wearable technology to provide support and monitoring in rehabilitation for dementia: feasibility case series. JMIR Form Res 2019 Oct 17;3(4):e12346 [FREE Full text] [doi: 10.2196/12346] [Medline: 31625951]
52. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and wellbeing. J Biomed Inform 2018 Jan;77:120-132 [FREE Full text] [doi: 10.1016/j.jbi.2017.12.008] [Medline: 29248628]

Abbreviations

EQ-VAS: EuroQol visual analog scale
LIWC: Linguistic Inquiry and Word Count
MS: multiple sclerosis
PPMS: primary progressive multiple sclerosis
SMSR: Swiss Multiple Sclerosis Registry
SPMS: secondary progressive multiple sclerosis

Edited by L Buis; submitted 13.04.22; peer-reviewed by PJ Jongen, I Ahn, M Clarysse; comments to author 25.05.22; revised version received 29.07.22; accepted 24.08.22; published 05.10.22.

Please cite as:
Sieber C, Chiavi D, Haag C, Kaufmann M, Horn AB, Dressel H, Zecca C, Calabrese P, Pot C, Kamm CP, von Wyl V, Swiss Multiple Sclerosis Registry
Electronic Health Diary Campaigns to Complement Longitudinal Assessments in Persons With Multiple Sclerosis: Nested Observational Study
JMIR Mhealth Uhealth 2022;10(10):e38709
URL: https://mhealth.jmir.org/2022/10/e38709
doi:10.2196/38709
PMID:36197713

©Chloé Sieber, Deborah Chiavi, Christina Haag, Marco Kaufmann, Andrea B Horn, Holger Dressel, Chiara Zecca, Pasquale Calabrese, Caroline Pot, Christian Philipp Kamm, Viktor von Wyl, Swiss Multiple Sclerosis Registry. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 05.10.2022. 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.

XSL•FO
RenderX

Original Paper

Retention, Fasting Patterns, and Weight Loss With an Intermittent Fasting App: Large-Scale, 52-Week Observational Study

Luisa Torres[1], PhD; Joy L Lee[2,3], MSc, PhD; Seho Park[2,3], MSc, PhD; R Christian Di Lorenzo[1,4], MSc; Jonathan P Branam[1], MSc; Shelagh A Fraser[1,5], MSc, MD; Benjamin A Salisbury[1,6], MSc, PhD

[1]LifeOmic, Indianapolis, IN, United States
[2]Indiana University School of Medicine, Indianapolis, IN, United States
[3]The Regenstrief Institute, Indianapolis, IN, United States
[4]Infinia ML, Durham, NC, United States
[5]Priority Physicians, Carmel, IN, United States
[6]Program in Biomedical Sciences, Wright State University, Dayton, OH, United States

Corresponding Author:
Benjamin A Salisbury, MSc, PhD
LifeOmic
351 W. 10th St
Indianapolis, IN, 46202
United States
Phone: 1 9377512962
Email: ben.salisbury@lifeomic.com

Abstract

Background: Intermittent fasting (IF) is an increasingly popular approach to dietary control that focuses on the timing of eating rather than the quantity and content of caloric intake. IF practitioners typically seek to improve their weight and other health factors. Millions of practitioners have turned to purpose-built mobile apps to help them track and adhere to their fasts and monitor changes in their weight and other biometrics.

Objective: This study aimed to quantify user retention, fasting patterns, and weight loss by users of 2 IF mobile apps. We also sought to describe and model starting BMI, amount of fasting, frequency of weight tracking, and other demographics as correlates of retention and weight change.

Methods: We assembled height, weight, fasting, and demographic data of adult users (ages 18-100 years) of the LIFE Fasting Tracker and LIFE Extend apps from 2018 to 2020. Retention for up to 52 weeks was quantified based on recorded fasts and correlated with user demographics. Users who provided height and at least 2 readings of weight and whose first fast and weight records were contemporaneous were included in the weight loss analysis. Fasting was quantified as extended fasting hours (EFH; hours beyond 12 in a fast) averaged per day (EFH per day). Retention was modeled using a Cox proportional hazards regression. Weight loss was analyzed using linear regression.

Results: A total of 792,692 users were followed for retention based on 26 million recorded fasts. Of these, 132,775 (16.7%) users were retained at 13 weeks, 54,881 (6.9%) at 26 weeks, and 16,478 (2.1%) at 52 weeks, allowing 4 consecutive weeks of inactivity. The survival analysis using Cox regression indicated that retention was positively associated with age and exercise and negatively associated with stress and smoking. Weight loss in the qualifying cohort (n=161,346) was strongly correlated with starting BMI and EFH per day, which displayed a positive interaction. Users with a BMI ≥40 kg/m^2 lost 13.9% of their starting weight by 52 weeks versus a slight weight gain on average for users with starting BMI <23 kg/m^2. EFH per day was an approximately linear predictor of weight loss. By week 26, users lost over 1% of their starting weight per EFH per day on average. The regression analysis using all variables was highly predictive of weight change at 26 weeks (R^2=0.334) with starting BMI and EFH per day as the most significant predictors.

Conclusions: IF with LIFE mobile apps appears to be a sustainable approach to weight reduction in the overweight and obese population. Healthy weight and underweight individuals do not lose much weight on average, even with extensive fasting. Users who are obese lose substantial weight over time, with more weight loss in those who fast more.

(JMIR Mhealth Uhealth 2022;10(10):e35896) doi:10.2196/35896

XSL•FO
RenderX

KEYWORDS

intermittent fasting; time-restricted eating; weight loss; obesity; mobile apps; diet trackers; retention

Introduction

Background

Worldwide, 13% of adults have obesity (BMI ≥30 kg/m^2) and 39% are overweight (BMI ≥25 kg/m^2) [1]. In the United States, obesity is >3 times higher at 42.5% of the adult population [2], and by 2030, the prevalence is expected to be close to 50% [3]. Mobile health apps that incorporate practices such as intermittent fasting (IF) may be a cost-effective approach to mitigating weight gain.

IF is a set of dietary patterns commonly pursued for weight loss that limits the timing of eating without restricting food content. Studies have shown that various IF methods are effective for weight loss in people who are overweight [4-9], including time-restricted eating, alternate day fasting, and a 5:2 diet [10-14]. However, these studies have been conducted in small populations (<200 completers), for short durations (a few weeks up to 6 months), and with narrowly defined IF protocols assigned to participants. In the real-world setting, IF patterns may not be as cleanly defined, especially over longer durations during which multiple fasting patterns may be explored.

Mobile apps for IF and weight tracking offer an opportunity to examine IF in a less-controlled setting and investigate its real-world efficacy for weight management. They are also a low-cost intervention for addressing obesity in the general population and may incentivize the adoption of healthy habits, including exercise and healthy eating [15,16]. Despite their potential benefits, the use of mobile health apps has been limited owing to low retention rates [17-19], and only a handful of available health apps have been subjected to rigorous study to establish their efficacy.

This Study

We evaluated retention, fasting patterns, and weight change among users of 2 free IF tracking apps, collectively known as LIFE apps: LIFE Fasting Tracker (LFT), which is focused on fasting, and LIFE Extend (LX), which additionally supports tracking of physical activity, mindfulness, sleep, and healthy plant intake. From 2018 to 2020, the 2 apps acquired a combined user base of 2.5 million downloads. User accounts and backend data storage for the 2 apps are shared, such that fasts could be started in one app and stopped in the other, and all the data are interchangeable. LFT was launched over a year earlier than LX, so only a small fraction of the data in this study was generated via LX.

We followed nearly 800,000 users for retention and real-world fasting behaviors. We further analyzed weight change patterns relative to app use and demographics in a subpopulation of over 160,000 users who used apps to track their weight over time.

We showed that practicing IF with a dedicated mobile app is an effective and sustainable approach to weight loss in individuals initially classified as overweight and obese. Many users consistently used the apps to record fasts every week for months. Users with obesity lost substantial weight over time, with more weight loss in avid fasters. Weight loss in users with obese or overweight BMI was sustained, on average, for up to a year with little rebound. Conversely, users with a healthy or underweight BMI lost little or no weight, even with extensive fasting over 52 weeks.

Methods

Mobile Apps and Users

We assembled all fasting and weight data for users who began using either of the LIFE apps between the launch of the LFT in May 2018 and December 2020. Analyses of fasting, retention, and weight are all relative to when the user began recording fasts in the apps, minimizing seasonal and other calendar effects. Other voluntary data collected were sex, age, race, height, diet, exercise frequency, stress level, smoking status, and primary health concern. For inclusion in our analyses, users had to have provided sex and date of birth and recorded at least one fast, the first of which had to have been started on or after their 18th birthday.

For the weight change analyses, we further required the user to have provided height and an initial weight recorded within 7 days of the first fast. Heights and weights had to have been within validation ranges of 145-203 cm and 25-249 kg, respectively. Height and weight can be entered in either metric or imperial units, with subsequent conversion to metric units for storage and analysis. We identified 902 users whose weight change at weeks 1 to 52 was >5 SD from the average across all users for that week. Without knowing which value or values were presumably misentered, we simply excluded those users entirely from the weight analysis, made feasible by the study's large sample size. Weights were subject to a 24-hour *burn-in* period, using the last weight recorded during that time as the baseline value. This burn-in accommodated users who may have entered an initial weight in the app based on their recollection and entered an update after checking it on a scale or who corrected their entry after checking units.

Fasting

We assembled all fasting records for the full set of nearly 800,000 users. Although the apps allowed shorter and longer fasts to be tracked, we eliminated fasts under 8 hours and truncated fasts to a maximum length of 240 hours. To reduce the effects of forgotten fasts that were ended and saved in the apps long after eating had resumed, we eliminated any fast that was 120 hours or longer but where a fasting goal of under 24 hours had been specified by the user. This yielded 25,983,817 fasts for our analyses.

We aggregated fasting statistics for each user for weeks 1 to 104 but primarily investigated weeks 1 to 52. Information regarding week 53 to 104 was used, when available and applicable, to determine retention. For each week, we totaled the number of started fasts, the average fast length, and the sum of hours beyond the first 12 in a fast, which we call extended

XSL•FO
RenderX

fasting hours (EFH), and EFH per day (ie, EFH per day=sum of EFH for all fasts started in a week/7). The 12-hour time point in a fast is when the body is expected to have depleted the energy from recently consumed food and may begin a metabolic switch to deriving energy from the breakdown of fat [20]. This shift is also referred to as entering ketosis and is thought to help drive weight loss and other health benefits [20]. EFH per day also presents a unified way to analyze fasting effects for people with different total fasting time and frequency but similar time in ketosis. For example, a user who performs daily 16-hour fasts will have the same 4 EFH per day as a user who performs two 26-hour fasts per week, even though their total recorded fasting time is quite different (112 vs 52 hours per week).

We also calculated the cumulative means of these measures for all weeks, up to and including the given week.

Retention

We assessed user retention based solely on records of completed fasts and not on other user behavior such as log-ins or use of other app features. Starting with the date of each user's first fast, we assessed their fasting activity for each week. The most restrictive definition of retention is when a user is only considered retained so long as they record a fast in each consecutive week. We refer to this definition as retention with a 0-week grace period. In contrast, the most lenient definition of retention is where the user is considered retained the entire time between their first and last recorded fast, regardless of how much activity they have in between. We refer to this as retention with an unlimited grace period. This definition is also sometimes called rolling retention [21].

We explored retention by varying the number of weeks in the grace period. We looked at 0, 2, 4, 8, 13, 26, and unlimited-week grace periods. After considering this spectrum of retention metrics, we decided to apply the 4-week grace period retention definition for all subsequent analyses. For example, if the user recorded no fasts in weeks 10 to 13 but did fast in week 14, the user was still considered retained in weeks 10 to 14, but if they resumed fasting in week 15 or later, their retention would have ended with week 9. Note that our univariate estimates of retention are conservative because many users start near the end of our data collection period, thus not having the opportunity to be counted as active in the app during the full 52 weeks (plus the grace period) that they might otherwise have counted toward. In the multivariate analysis, we used right censoring to account for this issue.

Weight Change and BMI

Users were included in the weight change analyses for all weeks for which they satisfied the 4-week grace period retention definition and in which they had a recorded weight. To account for the wide range of starting weights, weight change was analyzed as percent change from the user's starting weight. The effect of obesity was also considered in some analyses by stratification on starting BMI. We categorized BMI using the Centers for Disease Control and Prevention definitions [22], with the further division of the healthy BMI category into healthy low and healthy high defined as the ranges 18.5 to 22 and 23 to 24, respectively. Healthy low and healthy high categories had an approximately equal representation in our baseline user data.

For the weekly weight aggregates, we calculated mean weight and the number of weights recorded during the week. The baseline weights were excluded from the week 1 aggregates.

Analysis

We performed all data analysis using Python 3.9 libraries in a JupyterLab [23] notebook environment installed within LifeOmic's Precision Health Cloud, the secure, Health Insurance Portability and Accountability Act–compliant backend of the LIFE apps. A security review process was used to ensure that no identifiable data were released from the precision health cloud. In addition to providing descriptive summary statistics, including means, SDs, and percentages, we used multivariate modeling approaches. Retention was modelled using Cox proportional hazards regression, as implemented in the Lifelines package (version 0.27.1) [24]. Right censoring was applied to users who joined late in the study and did not have the opportunity to be retained for 52 weeks. Weight change was modeled using ordinary least squares regression, as implemented in Statsmodels (version 0.12.2) [25]. Graphs were generated in Seaborn (version 0.11.1) [26], which was also used to generate the CIs displayed, except for the hazard ratios and regression figures, which were generated in Plotly (version 5.0.0) [27]. Data handling was managed using Pandas (version 1.3.1) [28].

Ethical Considerations

This study was exempt from institutional review board approval per Indiana University's research guidelines [29]. The study consisted of retrospective secondary analysis of deidentified data. The use of these data for research and aggregate reporting is covered in the privacy policy of the LIFE apps [30].

Results

LIFE Apps Users

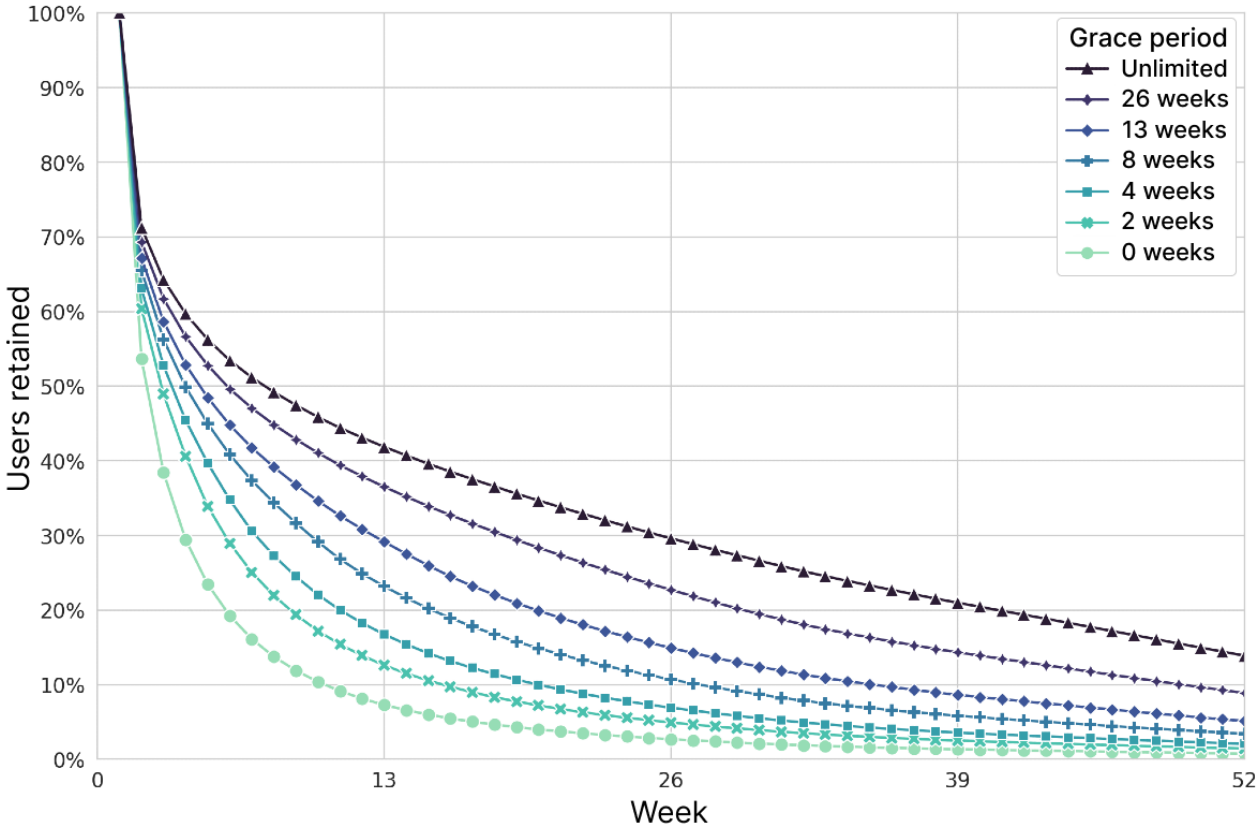
A total of 792,692 users satisfied the inclusion requirements for the fasting and retention analysis. The detailed demographic and biometric data for this population are presented in Multimedia Appendix 1. Their mean age was 36.7 (SD 10.9, range 18-100) years, and 81.3% of users were female. Users were located in nearly 200 different countries, with the majority being in the United States. Of these, 161,346 users met the height and weight measurement requirements and recorded at least one post–burn-in weight. This subpopulation was demographically similar to the entire population.

Retention

Figure 1 displays the retention patterns for the LIFE apps over the course of 52 weeks, calculated using 7 different fasting activity grace periods. There was an immediate drop of 28.7% of users (227,867/792,692) who never recorded a fast beyond week 1. Under the unlimited grace period, where up to 102 weeks of no fasting records were permitted, 41.9%, 29.6%, 21%, and 13.9% of users were retained at 13, 26, 39, and 52 weeks, respectively. At the other extreme, 0-week grace period retention (also known as full retention) captured a much smaller fraction of users (7.3%, 2.7%, 1.4%, and 0.8%, respectively).

XSL•FO
RenderX

Figure 1. User retention, calculated by different grace periods of inactivity. In total, 792,692 users were tracked starting from their first recorded fast. Grace periods extended out to 104 weeks for the unlimited definition.



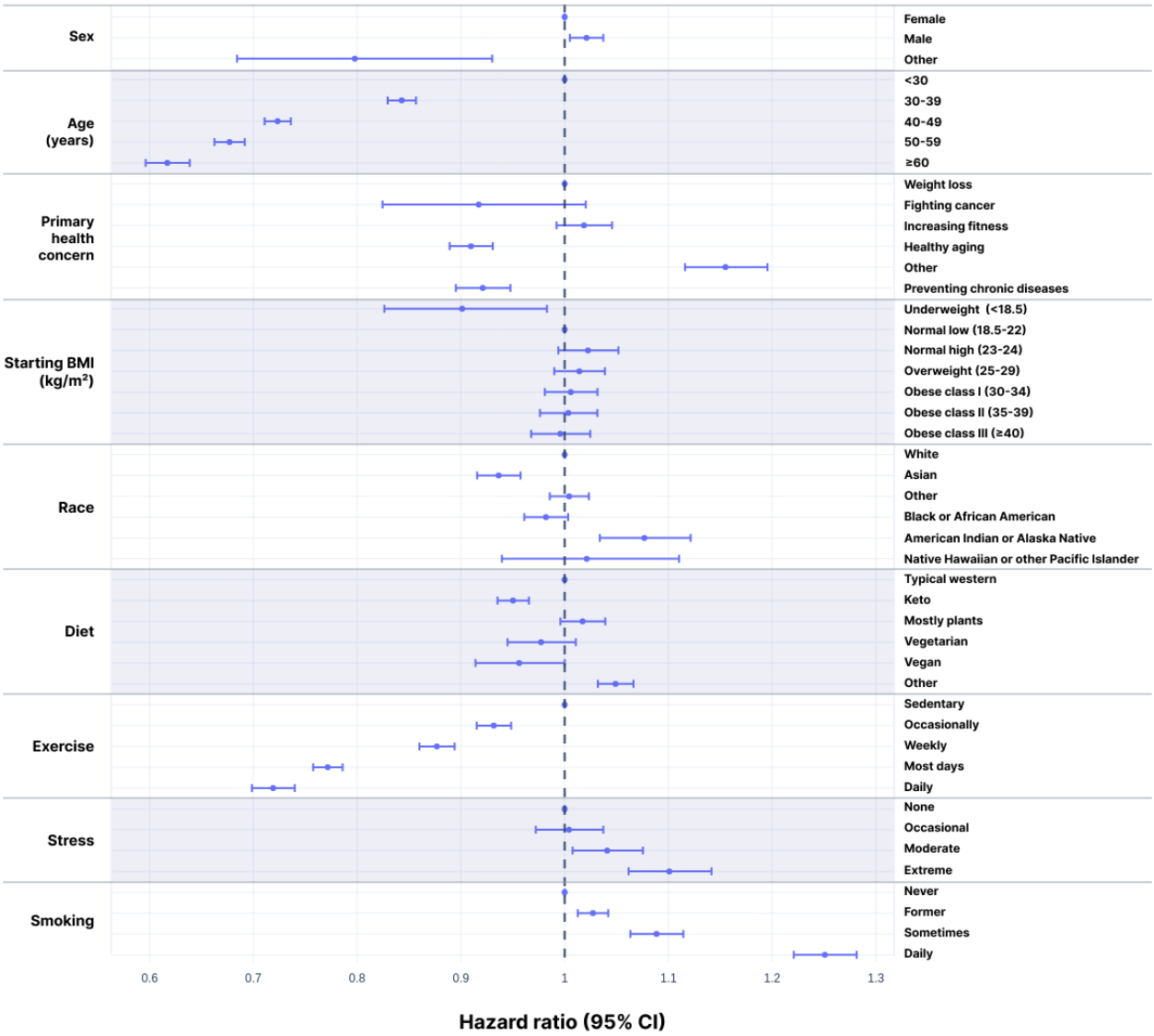
Intermediate grace periods corresponded naturally to intermediate retention rates (Figure 1). For all remaining analyses in this study, we opted to use the 4-week grace period definition of retention because it allowed us to study the evident variability of use while precluding highly prolonged inactivity. These users recorded a fast approximately every month at a minimum. Retention rates under this definition were 16.7%, 6.9%, 3.6%, and 2.1% at 13, 26, 39, and 52 weeks, respectively. While users may have slowly increased their fasting frequency, taken a break, or ramped down at the end, exploring such behavioral dynamics falls outside the scope of this study.

Demographics

Retention using the 4-week grace period differed substantially by several demographic criteria (Multimedia Appendix 1). The Cox proportional hazards regression model built over the first 52 weeks confirmed that several factors were significant, even after controlling for other factors (Multimedia Appendix 1; Figure 2). While many demographic and behavioral factors were found to correlate with retention, 4 trends were particularly notable in the Cox model. Older users had higher retention—a hazard ratio of 0.617 (95% CI 0.596-0.639) for users aged ≥60 years means they are estimated to be about 38% less likely to drop each week than users <30 years. Similarly, increasing levels of exercise (as reported at baseline) reflected much greater retention, with daily exercisers dropping about 28% less often than users with a sedentary lifestyle. Conversely, stress and smoking conferred lower retention rates—10% and 25% higher drop rates respectively for users with extreme stress or daily smoking habits relative to users who have no stress and never smoked. While losing weight was the most common primary health concern, those users' retention was substantially lower than for users whose primary concerns were healthy aging and preventing chronic disease. Sex and starting BMI appeared to have only small effects on retention.

XSL•FO
RenderX

Figure 2. Hazard ratios with 95% CIs for failure to retain. Based on the Cox proportional hazards model over the 52-week study. HR=1 corresponds to the reference values: female, age <30 years, primary health concern as weight loss, starting BMI in the normal low category, white, typical western diet, sedentary, and never smoker. HR<1 reflects higher retention rates.



Fasting Practices and Patterns

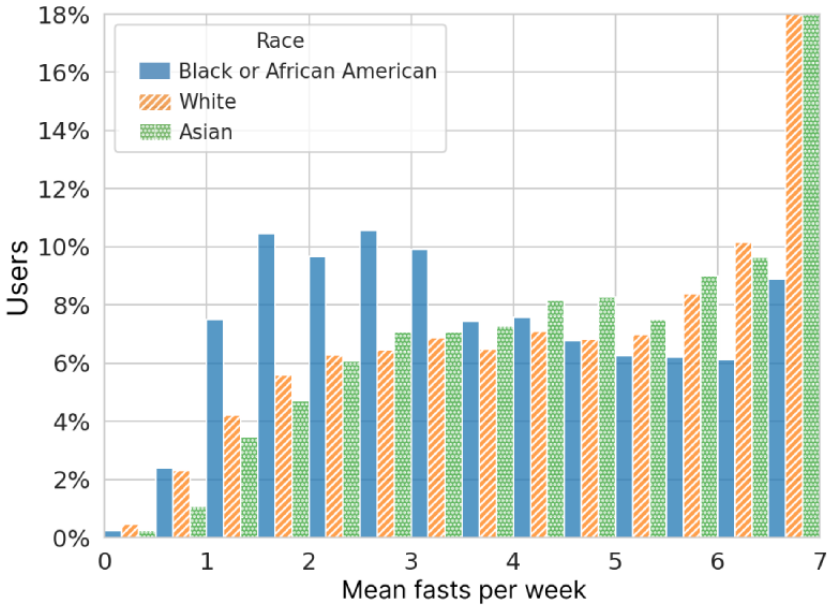
Weekly Fasting Frequency

Even when retained, user fasting behavior is likely to change over time. We examined fasting patterns based mainly on the first 26 weeks among users retained that long. The 26-week period is long enough to see what long-term use of the fasting apps is like, while affording a larger sample size than looking only at users who were retained at 52 weeks. It also avoids overweighting the first few weeks of use when we had the largest sample but while users were still establishing their fasting routines.

The most common days to start a fast were Sunday, Monday, and Tuesday, whereas Friday and Saturday were the least popular. We also examined the distribution of fasts per user per week over the first 26 weeks for 54,811 users retained at 26 weeks using the 4-week grace period. The mean frequency was 4.25 (SD 1.91) fasts per week. Fasting frequency was approximately bimodal, with a broad peak centered on 3 fasts per week and a sharp peak at 7. Slightly more than one-quarter (13,981/54,881, 25.5%) of the users fasted 6 to 7 times per week. In Figure 3, weekly fasting frequency is shown separately for the 3 most common self-reported race values. The differences suggest large cultural influences on a user's choice of fasting routine. Older users were also much more likely to fast 6 to 7 times per week than younger users (4033/11,768, 34.3%, vs 1521/9572, 15.9%) for users ≥50 years versus those <30 years.

XSL•FO
RenderX

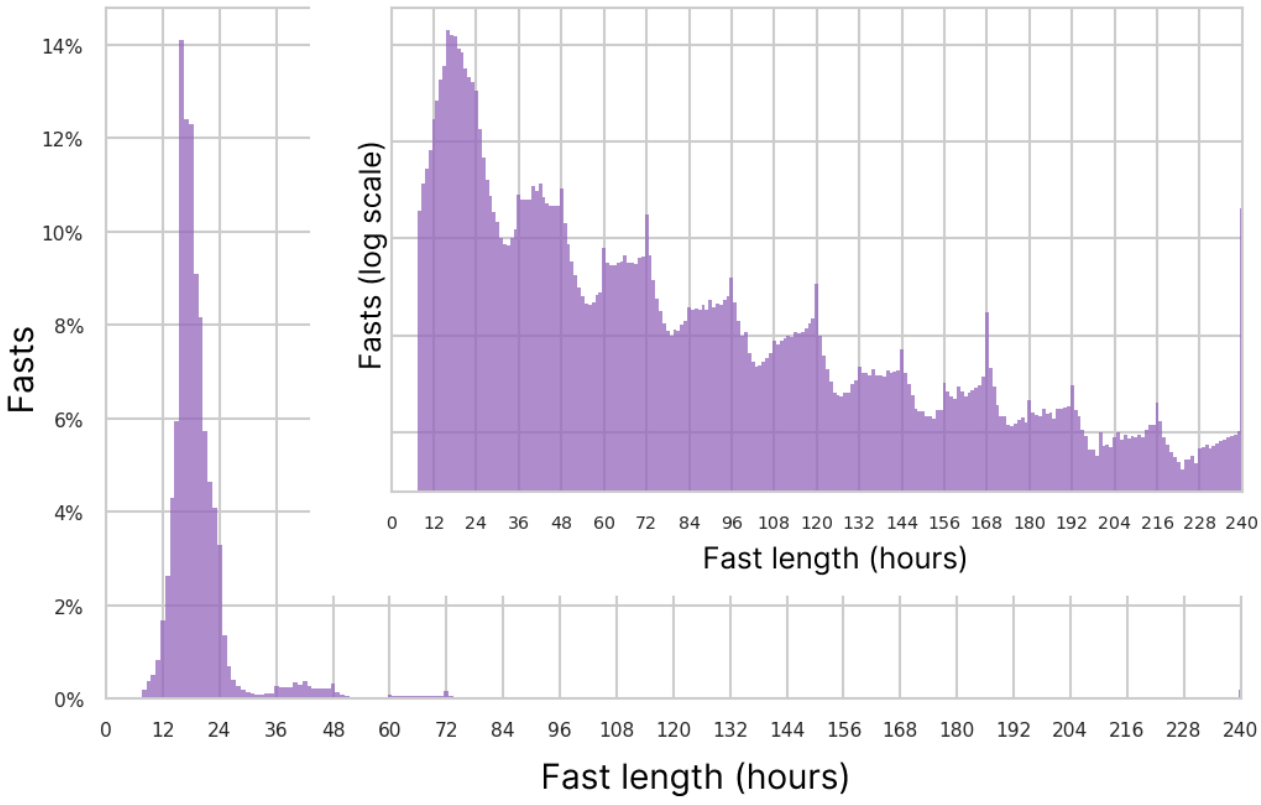
Figure 3. Fasting frequency statistics for users retained at 26 weeks, averaged over the first 26 weeks of use and grouped by self-reported race. Bins are half-fast width, left-inclusive, and include 7 in the highest bin.



Fasting Lengths

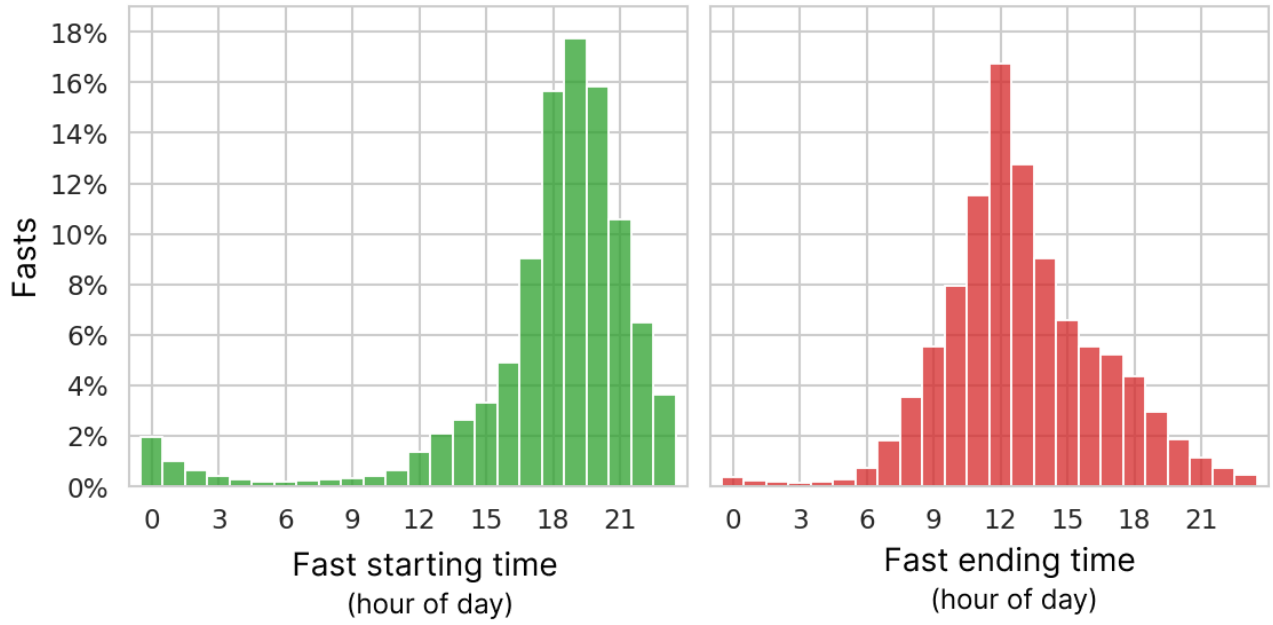
The most common fasting length of the 26 million fasts analyzed over the entire length of the study was 16 hours. The mean and median lengths were 21.0 and 18.0 hours, respectively, while the lower and upper quartiles were 16.1 and 20.9 hours. Figure 4 shows the complete distribution of fasting lengths. The modal starting and ending hours were 7 PM and noon, respectively (Figure 5). A total of 93.5% (24,289,517/25,983,817) of fasts were ≤32 hours, typically spanning a single night. A pattern of multiday fasts is evident when plotted on the log scale in Figure 4 (inset), with smaller peaks for each additional day and clear spikes at precise multiples of 24 hours.

Figure 4. Histogram of fast lengths and a log scale histogram inset.



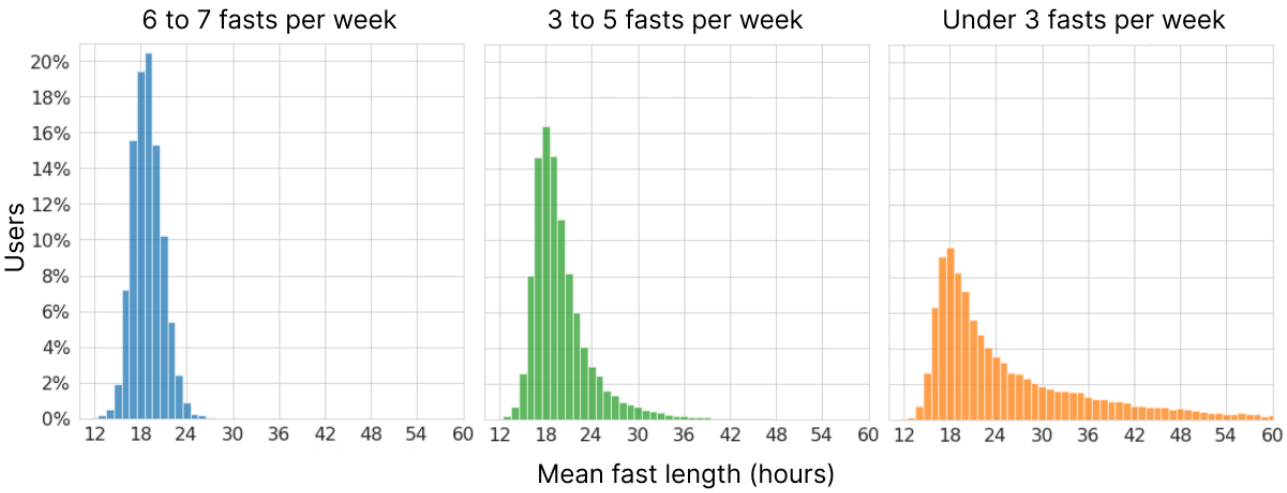
XSL•FO
RenderX

Figure 5. Distribution of starting and ending hour of fasts in local time.



We also examined the average fast length by user for the 54,881 users who were retained at 26 weeks under the 4-week grace period definition. Figure 6 shows those 26-week averages broken down by user fasting frequency. Overall, 8.2% (4506/54,881) of users had a mean fast length of >32 hours, indicating a pattern of multiday fasts. As expected, among users who fasted <3 times per week, a much larger fraction (4055/17,057, 23.7%) was in the multiday zone of >32-hour average fasts, although the modal average was 18 hours. The average fast lengths for users who fasted 6 to 7 times per week also varied greatly, peaking at 19 hours.

Figure 6. Distribution of average fast lengths per user across the first 26 weeks for users still retained at 26 weeks, broken down by weekly fasting frequency.



Combining fasting length and frequency, the cumulative mean EFH per day was 5.0 at 26 weeks, which would correspond to a daily fasting routine of 17 hours.

Weight Change

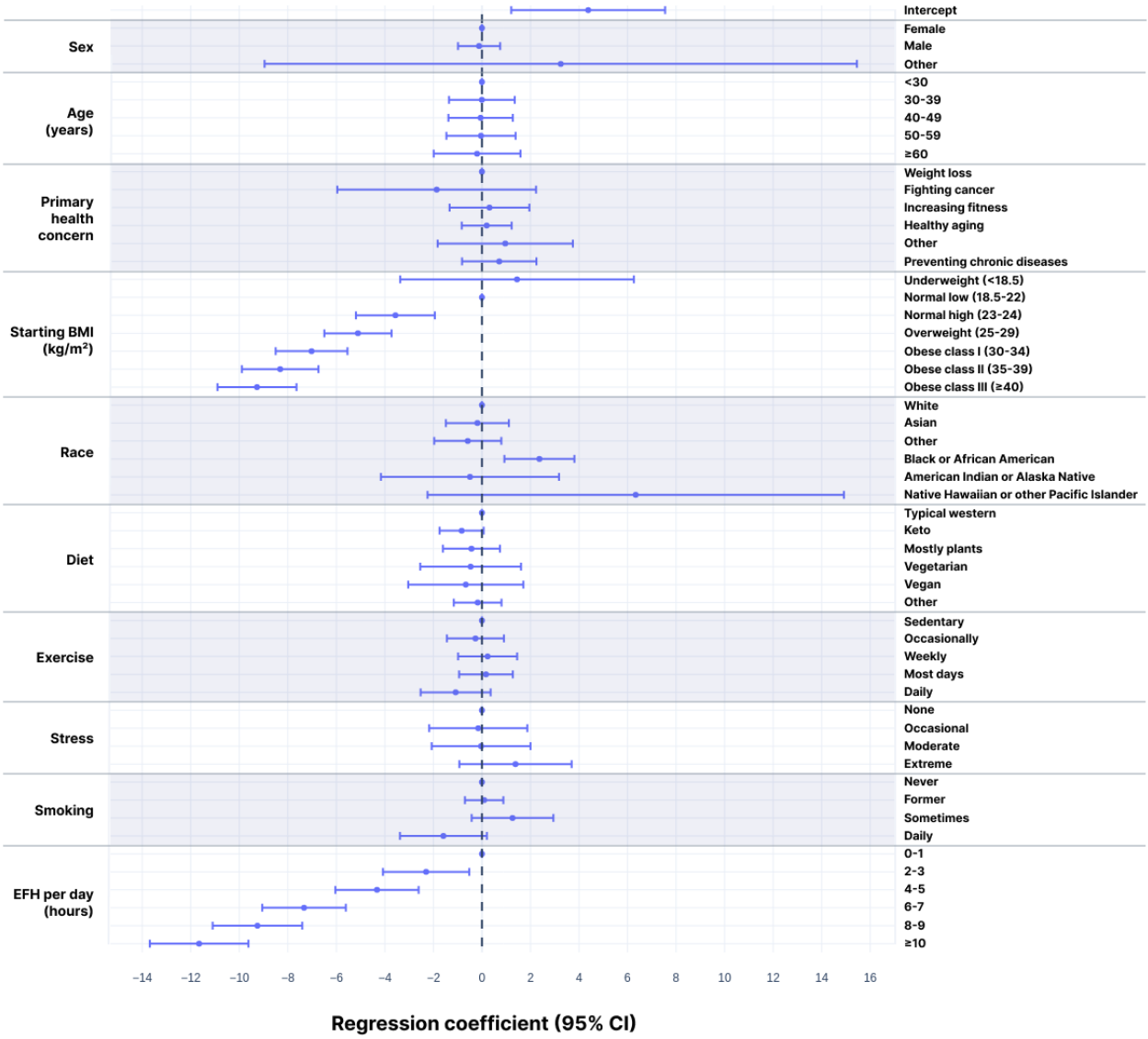
Demographics

We analyzed weight change for the 161,346 users who met the 4-week grace period retention criteria and recorded multiple weights in the fasting apps. From the univariate perspective, weight change as an outcome varied by several factors, including age, primary health concern, starting BMI, and EFH per day (Multimedia Appendix 2).

To address the correlation and confounding among variables, we conducted an ordinary least squares regression analysis of weight change at the 26-week time point. At 26 weeks, there were 1252 users with a recorded weight and values for all input variables. The result was that the only factors with $P<.05$ were starting BMI, EFH per day, and Black or African American race (Multimedia Appendix 2; Figure 7); $R^2=0.334$. Results were similar for the models built at weeks 13, 39, and 52 (data not shown).

XSL•FO
RenderX

Figure 7. Regression coefficients with 95% CIs for weight change at 26 weeks. The model was built using ordinary least squares linear regression for the 1252 users who had answers for all variables and a weight recorded in week 26 (R^2=0.334). Coefficients are shown as zero for the reference states: female, age <30 years, primary health concern as weight loss, starting BMI in the normal low category, white, typical western diet, sedentary, and never smoker. The regression coefficients reflect the difference in percent weight change at 26 weeks relative to the reference state for that category.



We further graphically explored the 52-week patterns of weight change relative to EFH per day and starting BMI, which emerged as the main variables explaining variability in weight change. Figure 8 depicts weight change for users who are not categorized as underweight binned weekly based on their cumulative average EFH per day. While users fasting less than 2 EFH per day lost only about 2% of their starting weight by 26 weeks, users with more extensive fasting lost more than 1% of their starting weight for each additional hour of EFH per day. Within each EFH per day bin, weight change appeared to eventually plateau, with weight loss continuing longer at higher levels of fasting. Weight loss continued for 39 weeks for users with ≥8 EFH per day before plateauing. A graph of weight change stratified by starting BMI is shown in Figure 9.

XSL•FO
RenderX

Figure 8. Weight change over time, stratified by users' cumulative average extended fasting hours (EFH) per day. Excludes users with underweight starting BMI. Mean values are plotted with 95% CIs represented by shading.

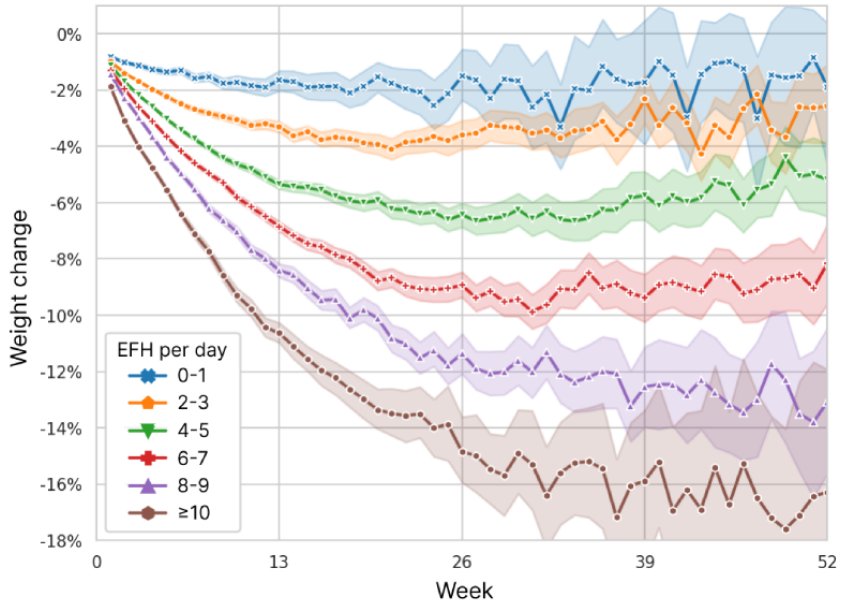
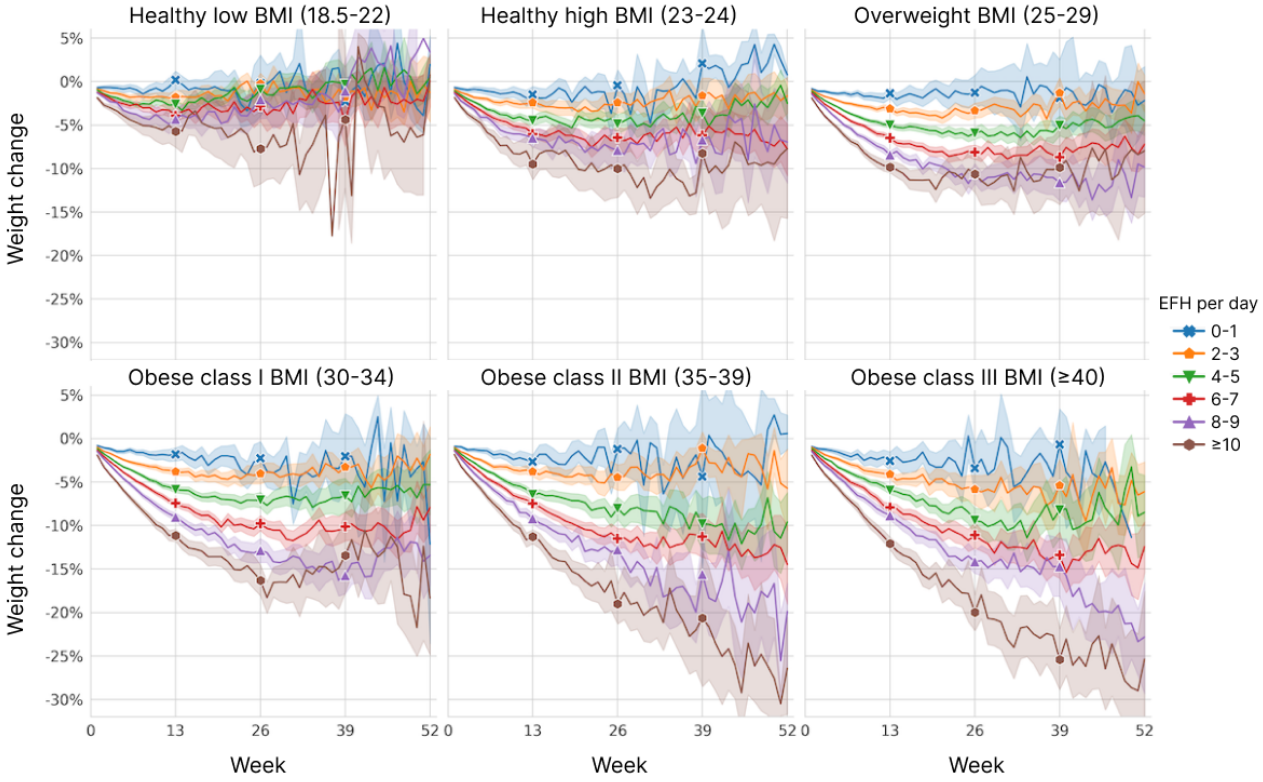


Figure 9. Weight change over time, stratified by user's starting BMI category. Mean values are plotted with 95% CIs represented by shading.

We examined the combined effects of starting BMI and fasting quantity by plotting the EFH per day strata separately for each starting BMI category (Figure 10). Within each category, the effect of increasing EFH per day appears to be approximately linear, as seen previously in Figure 8, but the scale at which extended fasting impacts weight loss increases with higher BMI. Similarly, it is clear that the starting BMI is still predictive of weight loss, even after accounting for the amount of fasting. The evident interaction between these 2 factors was confirmed by rebuilding the 26-week regression model with the addition of an interaction term for continuous measures of starting BMI and EFH per day. In that analysis, the *P* value for the interaction term was <.001, whereas the *P* values for the EFH per day bins increased to >.05. R^2 increased slightly to 0.356.

XSL•FO
RenderX

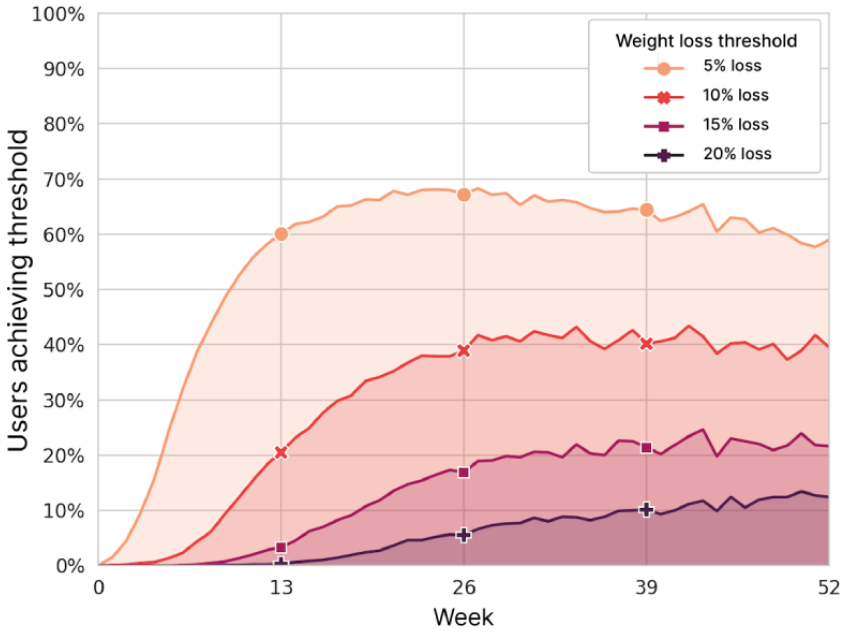
Figure 10. Weight change over time, stratified by user's starting BMI level and cumulative average extended fasting hours (EFH) per day. Mean values are plotted with 95% CIs represented by shading.



Weight Loss Thresholds

We also examined the number of users who achieved certain thresholds of weight loss. Figure 11 shows the proportion of users with starting BMI ≥25 kg/m^2 (ie, overweight or obese) who reached weight loss of 5%, 10%, 15%, and 20% over time. Success in reaching the 5% weight loss threshold was mostly achieved in the first 13 weeks and plateaued or peaked at 26 weeks. By 26 weeks, 67.2% (1475/2194) had lost at least 5% of their starting weight, and 38.9% (854/2194) had lost at least 10% of body weight. Reaching higher weight loss thresholds generally took much longer to achieve, with gradually larger fractions of users reaching them in 52 weeks.

Figure 11. Percentage of users with obese or overweight starting BMI (≥25 kg/m^2) who achieved 5%, 10%, 15%, and 20% weight loss by week.



XSL•FO
RenderX

Discussion

Context

This study is the largest examination of IF conducted to date and is orders of magnitude larger than any previous effort [5,7,12,31-33]. Owing to the use of mobile apps to record fasting events and weight, we were able to document real-world behavior and results, including both retention and weight change. Unlike most prior studies, we included people with healthy weight or underweight rather than just people categorized as overweight or obese, and our population covered extensive demographic variability, including an age range of 18 to 100 years.

Retention

The spectrum of retention metrics (Figure 1) shows that there are many users who consistently used the apps to record fasts every week for months (0-week grace period). Other users took breaks lasting weeks or months, but came back to the apps later—13.9% of users recorded a fast during weeks 52 to 104 (ie, retention at week 52 with an unlimited grace period), which is an underestimate because most users downloaded the app less than 2 years before the end of the study. Whether users are engaged in IF during reporting gaps is unknown, but these variable use patterns are likely typical for mobile health apps, as well as health behavior in general [34,35].

Retention statistics for mobile apps are not commonly available for proprietary apps. An analysis by AppsFlyer found that day 30 retention (fraction of original users active on day 30) for health and fitness apps in the United States in 2020 averaged <6% [36]. The week-12 retention statistics (fraction of original users active during week 12) for all apps in 2020 were 3.6% for Android and 5.1% for iOS. This 12-week retention definition is more generous than our 0-week grace period definition, yet the retention we observed at that time point (Figure 1) was greater, perhaps owing to the simple core utility of the fasting app for tracking the timing of fasts.

Age was the best predictor of retention in our study, which is consistent with other analyses of retention predictors for lifestyle interventions [37]. Older users may become more fixed in new habits, appreciate the consistency of using apps regularly, and be less likely to try out multiple competing apps. They are also more likely to have serious health concerns, such as healthy aging, which may increase their motivation to adhere to new interventions. Users whose stated primary health concern was weight loss were younger and had the lowest retention rate.

Several other variables were notable in their relationship to retention in both univariate and multivariate analyses (Multimedia Appendix 1). Consistent with previous reports [37,38], 26-week retention was more than twice as high for daily exercisers versus sedentary users. This was also concordant with previous findings of higher dietary compliance among those who exercise regularly [39]. Stress has been shown to predict poor adherence to weight loss programs [37], and users of the LIFE apps reporting extreme stress had the lowest retention. Smoking conferred lower retention rates compared with nonsmoking, which is also consistent with previous reports [40,41]. Interestingly, retention differences by sex, diet, and starting BMI were among the smallest.

Fasting Patterns and Weight Change

The real-world spectrum of fasting behavior documented in our study shows variable and flexible adherence to IF regimens, making specific idealized fasting protocols hard to discern in the data. We did clearly see a group of 25.5% of users who practice a daily, or nearly daily, fasting routine (≥6 days per week) averaged over the first 26 weeks, whereas the rest skip multiple days per week. Among users who fast, on average, fewer than 3 times per week, the majority fast under 24 hours, suggesting that they are more sporadic in their fasting or less vigorous in tracking in the apps. A total of 7.5% of users had average fasts over 32 hours, likely corresponding to the extended paradigms of IF such as 5:2 and alternate day fasting.

Owing to the multidimensional gradations of fasting patterns, we proposed the concept of EFH per day, as a metric to quantify fasting across all users. EFH per day combines fasting frequency and fasting length into a single measure and serves to unify the various fasting regimens for analysis. EFH per day was predictive of weight loss in a nearly linear fashion (Multimedia Appendix 2; Figures 8 and 10), supporting it as a relevant framework for quantifying fasting. We also showed that the magnitude of the fasting effect varied by starting BMI, with greater weight loss in individuals with higher levels of obesity practicing the same level of IF. Our findings have clear implications for people who wish to lose weight by using IF. The daily 16:8 IF routine that is commonly promoted is a minimum for those who wish to lose more than a few percent of their weight.

To explain these results, we hypothesize that the correlation between EFH per day and weight loss and the interaction with starting BMI can be attributed primarily to differences in caloric restriction. In previous studies, those who practiced alternate day fasting, the 5:2 diet, or time-restricted eating reduced their daily calorie intake by 10% to 30% [42]. With shorter eating windows, users with low starting BMI may be able to consume sufficient calories to maintain their weight, while users with higher BMI cannot, resulting in disproportionate weight loss. Various IF regimens have been shown to be as effective for weight loss as intentional caloric restriction [6,8,43,44], although IF might be easier to adopt and follow in the long term [13,45].

The weight loss effects of longer fasts may additionally be driven by the metabolic switch from glucose to ketones derived from fat tissue and free fatty acids [46]. This switch has previously been associated with weight loss [20] and occurs between 12 hours and 24 hours into a fast, depending on previous carbohydrate intake and energy expenditure [46,47]. A primary benefit of the switch from use of glucose to free fatty acids and ketones is the mobilization of body fat stores while preserving muscle mass, thereby improving body composition. Associated improvements in insulin sensitivity, visceral fat mass, and systemic inflammation may persist at the end of each fast when the system reverts to glucose metabolism, in part because of preserved muscle mass [19]. These effects are amplified in longer fasts, as insulin sensitivity reaches a nadir at 54 hours.

XSL•FO
RenderX

Analysis of weight loss threshold achievement facilitates the comparison of IF with other interventions. By 26 weeks, 67.2% (1475/2194) of users with overweight and obese starting BMI lost at least 5% of their starting weight. This is comparable to the results achieved by users of the Diabetes Prevention Program through the Noom platform, 64% of whom lost over 5% of their body weight in 24 weeks [48]. In our study, 38.9% (854/2194) of users with BMI ≥25 kg/m^2 lost at least 10% of their body weight in 26 weeks, while this threshold was achieved by only 23% of Noom users in the same amount of time [49]. In another comparison, only 25% of the participants enrolled in the Livongo Diabetes Prevention Program lost more than 10% of their body weight by 54 weeks, with most users achieving 5% weight loss [50].

While it is common for people who lose weight to be subject to weight regain [51], overweight users sustained most of their weight loss at 52 weeks, with users in obese class II and III (BMI ≥35 kg/m^2) trending toward even more weight loss at 52 weeks. In contrast, users whose starting BMI classified them as underweight had no weight change on average at 13 weeks and proceeded to gain weight if they continued with the apps out to 52 weeks. This is an important finding because of the frequently raised concern that IF may promote eating disorders [52]. Our findings suggest that IF is generally a safe practice even for users at the low end of BMI because even those users who fasted extensively tended to lose little weight (Figure 10).

Age was positively associated with greater weight loss, consistent with previous findings [53]. The effect we observed is explained by more fasting by older users rather than as a consequence of metabolic or dietary differences, according to our multivariate model. This is consistent with a 2013 study that showed that, compared with younger participants, older adults lost significantly more weight after a diet and exercise intervention and were more successful at maintaining weight loss even after 3 years of an internet-based maintenance protocol [54]. The lack of weight loss difference between women and men is also consistent with previous findings [55].

Conclusions

As of 2016, close to 50% of adults in the United States had tried to lose weight within the preceding 12 months according to Centers for Disease Control and Prevention data [56]. Moreover, as of 2018, over 40% of adults in the United States were considered obese [57]. In our study of mobile app users in the real-world setting, we found that IF is an effective strategy for weight loss for many people. Studies in people with obesity demonstrate that losing 10% of your body weight is enough to improve blood pressure and normalize cholesterol blood levels, while losing just 5% is enough to improve glycemic control, which is central to the prevention and management of diabetes [58]. In just 13 weeks, among LIFE apps users with a starting classification of overweight or obese, 60% of them lost 5% or more of their starting weight and 21% lost ≥10%, reflecting the potential clinical value that is achievable. These rates of success were higher in our users than in users of paid apps with active coaching, such as Noom and Livongo.

Self-reported data in this mobile setting offered many intriguing correlates of retention and weight loss. Meaningful factors included diet, exercise, stress, and smoking, all of which lend themselves well to mobile tracking, including integrated wearables, for reliable measurement and analysis.

Future Work

To better understand the mechanisms and residual variability of weight loss by app users, we would like to study the caloric input and expenditure of users directly. This can be achieved by asking users to record their daily dietary intake and exercise. Tracking exercise is amenable to passive tracking with wearable technology, and many users of the LIFE apps (specifically LX) already have regular data ingestion established with the most common fitness trackers. Such an additional study could help resolve the somewhat surprising finding that the diet and exercise habits self-reported at the beginning of the study did not correlate with weight change.

Although weight management is a clinically important objective, other clinically relevant outcomes could be measured and correlated with fasting behavior. These include assessments of mental and physical health, disease incidence, insulin resistance, medical costs, and professional and educational absenteeism. Advocates of IF point to studies in animal models and humans that suggest many of these benefits [11,14,59-61]. Facilitated by mobile and digital technology, we may be able to evaluate real-world evidence for these promises and tease apart their etiology.

Finally, studies show that social support improves health and well-being, and that people who have strong support networks are more likely to lose weight than those who do not [62,63]. The LIFE apps have a social “Circles” feature, where users can communicate with other users within the app. An analysis of circle participation is a subject for future work, but preliminary results suggest that retention is higher for users who are socially active in the app.

Limitations

The primary limitation of this study was that most data were self-reported, except for some weight values that were entered by smart scales. This limitation is compensated for by the large sample size of the study.

The observed weight change averages may be potentially confounded by users who stopped recording weights or even stopped using the app because of a lack of progress. Conversely, users who achieved success might have been less motivated to continue recording fasts and weights. Similarly, users may have been more likely to weigh themselves and record their weight in the app if they had lost weight, which could then exaggerate the weight loss estimates in this study. These effects may be challenging to untangle, but the trends and correlates of weight change identified should be robust.

Owing to the limited observational nature of this study, users who fasted longer may have adopted other diet-related practices more than users who fasted less without our knowledge. Similarly, we did not have information about users’ previous experience with IF. Given that the largest retention losses and

XSL•FO
RenderX

greatest rates of weight change occurred in the earliest weeks of app use, previous fasting experience or even recent weight changes could skew the reported progress. In an IF-naive population, we expect weight loss to be relatively larger.

Another limitation of the study is that, due to being observational, it lacked explicit controls. However, this limitation was offset by the wide range of fasting behaviors among users. We used this natural variability as a form of self-directed intervention, which allowed us to contrast and quantify the effects of different levels of fasting on a much broader scale than would be feasible for a randomized controlled trial.

Acknowledgments

The authors would like to thank Steven Bray for his close reading of the manuscript. The authors would like to appreciate the detailed and constructive feedback from the reviewers. RCD was affiliated with LifeOmic at the time of the study and is currently affiliated with Infinia ML.

Authors' Contributions

BAS performed most of the data collation and generation of figures and tables with substantial assistance from RCD and JPB. All major decisions were discussed and agreed upon by all authors. All authors contributed to writing the manuscript. Statistical approaches were guided by SP and executed by BAS. LT conducted most of the literature review.

Conflicts of Interest

BAS, LT, RCD, JPB, and SF are or were employees of LifeOmic, the makers of the apps studied in this paper, and they have ownership rights in the company. No one at LifeOmic other than the authors had any editorial oversight in performing this study or writing this paper.

Multimedia Appendix 1

Baseline demographics and retention at 13-week intervals using the 4-week grace period retention definition. Hazard ratios and corresponding *P* values are based on the 52-week Cox proportional hazards regression model applied to retention. Hazard ratios <1.0 reflect greater rates of retention.

[PDF File (Adobe PDF File), 115 KB - mhealth_v10i10e35896_app1.pdf]

Multimedia Appendix 2

Weight change at weeks 13, 26, 39, and 52 using the 4-week grace period retention definition relative to user demographics. Starting BMI (mean, SD) for each demographic category is included for context. Sample sizes reflect the maximum eligible users at each time point. Regression coefficients and *P* values refer to an ordinary least squares model of weight change at 26 weeks (n=1252). The intercept in the model was 4.378 (*P*=.007). Baseline BMI values for the EFH per day groupings are based on the 26-week cohort used in the regression analysis.

[PDF File (Adobe PDF File), 116 KB - mhealth_v10i10e35896_app2.pdf]

References

1. Ritchie H, Roser M. Obesity. Our World in Data. 2017 Aug 11. URL: https://ourworldindata.org/obesity [accessed 2021-12-08]
2. FastStats. Centers for Disease Control and Prevention. 2021. URL: https://www.cdc.gov/nchs/fastats/obesity-overweight.htm [accessed 2021-12-08]
3. Ward ZJ, Bleich SN, Cradock AL, Barrett JL, Giles CM, Flax C, et al. Projected U.S. state-level prevalence of adult obesity and severe obesity. N Engl J Med 2019 Dec 19;381(25):2440-2450. [doi: 10.1056/NEJMsa1909301] [Medline: 31851800]
4. Kalam F, Gabel K, Cienfuegos S, Wiseman E, Ezpeleta M, Steward M, et al. Alternate day fasting combined with a low-carbohydrate diet for weight loss, weight maintenance, and metabolic disease risk reduction. Obes Sci Pract 2019 Dec;5(6):531-539 [FREE Full text] [doi: 10.1002/osp4.367] [Medline: 31890243]
5. Cienfuegos S, Gabel K, Kalam F, Ezpeleta M, Wiseman E, Pavlou V, et al. Effects of 4- and 6-h time-restricted feeding on weight and cardiometabolic health: a randomized controlled trial in adults with obesity. Cell Metab 2020 Sep 01;32(3):366-78.e3 [FREE Full text] [doi: 10.1016/j.cmet.2020.06.018] [Medline: 32673591]
6. Harvie MN, Pegington M, Mattson MP, Frystyk J, Dillon B, Evans G, et al. The effects of intermittent or continuous energy restriction on weight loss and metabolic disease risk markers: a randomized trial in young overweight women. Int J Obes (Lond) 2011 May;35(5):714-727 [FREE Full text] [doi: 10.1038/ijo.2010.171] [Medline: 20921964]
7. Gabel K, Hoddy KK, Haggerty N, Song J, Kroeger CM, Trepanowski JF, et al. Effects of 8-hour time restricted feeding on body weight and metabolic disease risk factors in obese adults: a pilot study. Nutr Healthy Aging 2018 Jun 15;4(4):345-353 [FREE Full text] [doi: 10.3233/NHA-170036] [Medline: 29951594]

XSL•FO
RenderX

8. Gill S, Panda S. A smartphone app reveals erratic diurnal eating patterns in humans that can be modulated for health benefits. Cell Metab 2015 Nov 03;22(5):789-798 [FREE Full text] [doi: 10.1016/j.cmet.2015.09.005] [Medline: 26411343]
9. Conley M, Le Fevre L, Haywood C, Proietto J. Is two days of intermittent energy restriction per week a feasible weight loss approach in obese males? A randomised pilot study. Nutr Diet 2018 Feb;75(1):65-72. [doi: 10.1111/1747-0080.12372] [Medline: 28791787]
10. Trepanowski JF, Kroeger CM, Barnosky A, Klempel MC, Bhutani S, Hoddy KK, et al. Effect of alternate-day fasting on weight loss, weight maintenance, and cardioprotection among metabolically healthy obese adults: a randomized clinical trial. JAMA Intern Med 2017 Jul 01;177(7):930-938 [FREE Full text] [doi: 10.1001/jamainternmed.2017.0936] [Medline: 28459931]
11. Sutton EF, Beyl R, Early KS, Cefalu WT, Ravussin E, Peterson CM. Early time-restricted feeding improves insulin sensitivity, blood pressure, and oxidative stress even without weight loss in men with prediabetes. Cell Metab 2018 Jun 05;27(6):1212-21.e3 [FREE Full text] [doi: 10.1016/j.cmet.2018.04.010] [Medline: 29754952]
12. Wilkinson MJ, Manoogian EN, Zadourian A, Lo H, Fakhouri S, Shoghi A, et al. Ten-hour time-restricted eating reduces weight, blood pressure, and atherogenic lipids in patients with metabolic syndrome. Cell Metab 2020 Jan 07;31(1):92-104.e5 [FREE Full text] [doi: 10.1016/j.cmet.2019.11.004] [Medline: 31813824]
13. Hajek P, Przulj D, Pesola F, McRobbie H, Peerbux S, Phillips-Waller A, et al. A randomised controlled trial of the 5:2 diet. PLoS One 2021 Nov 17;16(11):e0258853 [FREE Full text] [doi: 10.1371/journal.pone.0258853] [Medline: 34788298]
14. Patikorn C, Roubal K, Veettil SK, Chandran V, Pham T, Lee YY, et al. Intermittent fasting and obesity-related health outcomes: an umbrella review of meta-analyses of randomized clinical trials. JAMA Netw Open 2021 Dec 01;4(12):e2139558 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.39558] [Medline: 34919135]
15. Dounavi K, Tsoumani O. Mobile health applications in weight management: a systematic literature review. Am J Prev Med 2019 Jun;56(6):894-903 [FREE Full text] [doi: 10.1016/j.amepre.2018.12.005] [Medline: 31003801]
16. Carroll JK, Moorhead A, Bond R, LeBlanc WG, Petrella RJ, Fiscella K. Who uses mobile phone health apps and does use matter? A secondary data analytics approach. J Med Internet Res 2017 Apr 19;19(4):e125 [FREE Full text] [doi: 10.2196/jmir.5604] [Medline: 28428170]
17. Krebs P, Duncan D. Health app use among US mobile phone owners: a national survey. JMIR Mhealth Uhealth 2015 Nov 04;3(4):e101 [FREE Full text] [doi: 10.2196/mhealth.4924] [Medline: 26537656]
18. Vaghefi I, Tulu B. The continued use of mobile health apps: insights from a longitudinal study. JMIR Mhealth Uhealth 2019 Aug 29;7(8):e12983 [FREE Full text] [doi: 10.2196/12983] [Medline: 31469081]
19. Ghelani DP, Moran LJ, Johnson C, Mousa A, Naderpoor N. Mobile apps for weight management: a review of the latest evidence to inform practice. Front Endocrinol (Lausanne) 2020 Jun 24;11:412 [FREE Full text] [doi: 10.3389/fendo.2020.00412] [Medline: 32670197]
20. Anton SD, Moehl K, Donahoo WT, Marosi K, Lee SA, Mainous AG, et al. Flipping the metabolic switch: understanding and applying the health benefits of fasting. Obesity (Silver Spring) 2018 Feb;26(2):254-268 [FREE Full text] [doi: 10.1002/oby.22065] [Medline: 29086496]
21. Owen JE, Jaworski BK, Kuhn E, Makin-Byrd KN, Ramsey KM, Hoffman JE. mHealth in the wild: using novel data to examine the reach, use, and impact of PTSD coach. JMIR Ment Health 2015 Mar 25;2(1):e7 [FREE Full text] [doi: 10.2196/mental.3935] [Medline: 26543913]
22. Centers for Disease Control and Prevention. 2021. URL: https://www.cdc.gov/obesity/adult/defining.html [accessed 2021-12-08]
23. Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, Jupyter Development Team. Jupyter Notebooks – a publishing format for reproducible computational workflows. In: Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing. 2016 Presented at: ELPUB '16; June 7-9, 2016; Göttingen, Germany p. 87-90. [doi: 10.3233/978-1-61499-649-1-87]
24. Davidson-Pilon C. lifelines, survival analysis in Python. Zenodo. 2022. URL: https://zenodo.org/record/6746356 [accessed 2022-07-13]
25. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with Python. In: Proceedings of the 9th Python in Science Conference. 2010 Presented at: SciPy '10; June 28-July 3, 2016; Austin, TX, USA. [doi: 10.25080/majora-92bf1922-011]
26. Waskom ML. seaborn: statistical data visualization. J Open Source Softw 2021 Apr 6;6(60):3021. [doi: 10.21105/joss.03021]
27. Collaborative data science. Plotly Technologies. Montréal, Canada; 2015. URL: https://plot.ly [accessed 2022-07-13]
28. Reback J, Jbrockmendel, McKinney W, Van den Bossche J, Roeschke M, Augspurger T, Gfyoung, Sinhrks, JHM Darbyshire, et al. pandas-dev/pandas: Pandas 1.4.3. Zenodo. 2022. URL: https://zenodo.org/record/6702671 [accessed 2022-07-13]
29. Levels of review: Human Subjects and Institutional Review Boards. Indiana University. URL: https://research.iu.edu/compliance/human-subjects/review-levels/index.html [accessed 2022-08-23]
30. LIFE Mobile Apps App Privacy Policy. LifeOmic Holdings. 2020 May 6. URL: https://lifeapps.io/life-mobile-apps-privacy-policy/ [accessed 2022-08-23]

XSL•FO
RenderX

31. Przulj D, Ladmore D, Smith KM, Phillips-Waller A, Hajek P. Time restricted eating as a weight loss intervention in adults with obesity. PLoS One 2021 Jan 28;16(1):e0246186 [FREE Full text] [doi: 10.1371/journal.pone.0246186] [Medline: 33508009]
32. Manoogian EN, Chow LS, Taub PR, Laferrère B, Panda S. Time-restricted eating for the prevention and management of metabolic diseases. Endocr Rev 2022 Mar 09;43(2):405-436. [doi: 10.1210/endrev/bnab027] [Medline: 34550357]
33. Lowe DA, Wu N, Rohdin-Bibby L, Moore AH, Kelly N, Liu YE, et al. Effects of time-restricted eating on weight loss and other metabolic parameters in women and men with overweight and obesity: the TREAT randomized clinical trial. JAMA Intern Med 2020 Nov 01;180(11):1491-1499 [FREE Full text] [doi: 10.1001/jamainternmed.2020.4153] [Medline: 32986097]
34. Pellegrini CA, Conroy DE, Phillips SM, Pfammatter AF, McFadden HG, Spring B. Daily and seasonal influences on dietary self-monitoring using a smartphone application. J Nutr Educ Behav 2018 Jan;50(1):56-61.e1 [FREE Full text] [doi: 10.1016/j.jneb.2016.12.004] [Medline: 29325663]
35. Solbrig L, Jones R, Kavanagh D, May J, Parkin T, Andrade J. People trying to lose weight dislike calorie counting apps and want motivational support to help them achieve their goals. Internet Interv 2017 Mar;7:23-31 [FREE Full text] [doi: 10.1016/j.invent.2016.12.003] [Medline: 28286739]
36. Attention: retention! 2020 app retention benchmarks. AppsFlyer. URL: https://www.appsflyer.com/infograms/app-retention-benchmarks-2020/ [accessed 2022-07-13]
37. Leung AW, Chan RS, Sea MM, Woo J. An overview of factors associated with adherence to lifestyle modification programs for weight management in adults. Int J Environ Res Public Health 2017 Aug 16;14(8):922 [FREE Full text] [doi: 10.3390/ijerph14080922] [Medline: 28813030]
38. Susin N, de Melo Boff R, Ludwig MW, Feoli AM, da Silva AG, Macagnan FE, et al. Predictors of adherence in a prevention program for patients with metabolic syndrome. J Health Psychol 2016 Oct;21(10):2156-2167. [doi: 10.1177/1359105315572451] [Medline: 25805660]
39. Racette SB, Schoeller DA, Kushner RF, Neil KM. Exercise enhances dietary compliance during moderate energy restriction in obese women. Am J Clin Nutr 1995 Aug;62(2):345-349. [doi: 10.1093/ajcn/62.2.345] [Medline: 7625341]
40. Bradshaw AJ, Horwath CC, Katzer L, Gray A. Non-dieting group interventions for overweight and obese women: what predicts non-completion and does completion improve outcomes? Public Health Nutr 2010 Oct;13(10):1622-1628. [doi: 10.1017/S1368980009992977] [Medline: 20025832]
41. Greenberg I, Stampfer MJ, Schwarzfuchs D, Shai I, DIRECT Group. Adherence and success in long-term weight loss diets: the dietary intervention randomized controlled trial (DIRECT). J Am Coll Nutr 2009 Apr;28(2):159-168. [doi: 10.1080/07315724.2009.10719767] [Medline: 19828901]
42. Varady KA, Cienfuegos S, Ezpeleta M, Gabel K. Clinical application of intermittent fasting for weight loss: progress and future directions. Nat Rev Endocrinol 2022 May;18(5):309-321. [doi: 10.1038/s41574-022-00638-x] [Medline: 35194176]
43. Schübel R, Nattenmüller J, Sookthai D, Nonnenmacher T, Graf ME, Riedl L, et al. Effects of intermittent and continuous calorie restriction on body weight and metabolism over 50 wk: a randomized controlled trial. Am J Clin Nutr 2018 Nov 01;108(5):933-945 [FREE Full text] [doi: 10.1093/ajcn/nqy196] [Medline: 30475957]
44. Gabel K, Kroeger CM, Trepanowski JF, Hoddy KK, Cienfuegos S, Kalam F, et al. Differential effects of alternate-day fasting versus daily calorie restriction on insulin resistance. Obesity (Silver Spring) 2019 Sep;27(9):1443-1450 [FREE Full text] [doi: 10.1002/oby.22564] [Medline: 31328895]
45. Rynders CA, Thomas EA, Zaman A, Pan Z, Catenacci VA, Melanson EL. Effectiveness of intermittent fasting and time-restricted feeding compared to continuous energy restriction for weight loss. Nutrients 2019 Oct 14;11(10):2442 [FREE Full text] [doi: 10.3390/nu11102442] [Medline: 31614992]
46. Mattson MP, Moehl K, Ghena N, Schmaedick M, Cheng A. Intermittent metabolic switching, neuroplasticity and brain health. Nat Rev Neurosci 2018 Feb;19(2):63-80 [FREE Full text] [doi: 10.1038/nrn.2017.156] [Medline: 29321682]
47. Longo VD, Di Tano M, Mattson MP, Guidi N. Intermittent and periodic fasting, longevity and disease. Nat Aging 2021 Jan;1(1):47-59 [FREE Full text] [doi: 10.1038/s43587-020-00013-3] [Medline: 35310455]
48. Michaelides A, Raby C, Wood M, Farr K, Toro-Ramos T. Weight loss efficacy of a novel mobile Diabetes Prevention Program delivery platform with human coaching. BMJ Open Diabetes Res Care 2016 Sep 5;4(1):e000264 [FREE Full text] [doi: 10.1136/bmjdrc-2016-000264] [Medline: 27651911]
49. Chin SO, Keum C, Woo J, Park J, Choi HJ, Woo JT, et al. Successful weight reduction and maintenance by using a smartphone application in those with overweight and obesity. Sci Rep 2016 Nov 07;6:34563 [FREE Full text] [doi: 10.1038/srep34563] [Medline: 27819345]
50. Painter SL, Lu W, Schneider J, James R, Shah B. Drivers of weight loss in a CDC-recognized digital diabetes prevention program. BMJ Open Diabetes Res Care 2020 Jul;8(1):e001132 [FREE Full text] [doi: 10.1136/bmjdrc-2019-001132] [Medline: 32624481]
51. Soeliman FA, Azadbakht L. Weight loss maintenance: a review on dietary related strategies. J Res Med Sci 2014 Mar;19(3):268-275 [FREE Full text] [Medline: 24949037]
52. Cuccolo K, Kramer R, Petros T, Thoennes M. Intermittent fasting implementation and association with eating disorder symptomatology. Eat Disord (forthcoming) 2021 Jun 30:1-21. [doi: 10.1080/10640266.2021.1922145] [Medline: 34191688]

XSL•FO
RenderX

53. DeLuca L, Toro-Ramos T, Michaelides A, Seng E, Swencionis C. Relationship between age and weight loss in noom: quasi-experimental study. JMIR Diabetes 2020 Jun 04;5(2):e18363 [FREE Full text] [doi: 10.2196/18363] [Medline: 32497017]
54. Svetkey LP, Clark JM, Funk K, Corsino L, Batch BC, Hollis JF, et al. Greater weight loss with increasing age in the weight loss maintenance trial. Obesity (Silver Spring) 2014 Jan;22(1):39-44 [FREE Full text] [doi: 10.1002/oby.20506] [Medline: 23640912]
55. Lin S, Lima Oliveira M, Gabel K, Kalam F, Cienfuegos S, Ezpeleta M, et al. Does the weight loss efficacy of alternate day fasting differ according to sex and menopausal status? Nutr Metab Cardiovasc Dis 2021 Feb 08;31(2):641-649 [FREE Full text] [doi: 10.1016/j.numecd.2020.10.018] [Medline: 33358713]
56. Martin CB, Herrick KA, Sarafrazi N, Ogden CL. NCHS Data Brief No. 313 – Attempts to Lose Weight Among Adults in the United States, 2013–2016. Centers for Disease Control and Prevention. 2018 Jul. URL: https://www.cdc.gov/nchs/products/databriefs/db313.htm [accessed 2021-12-08]
57. Hales CM, Carroll MD, Fryar CD, Ogden CL. NCHS Data Brief No. 360 – Prevalence of Obesity and Severe Obesity Among Adults: United States, 2017–2018. Centers for Disease Control and Prevention. 2020 Feb. URL: https://www.cdc.gov/nchs/products/databriefs/db360.htm [accessed 2021-12-08]
58. Orzano AJ, Scott JG. Diagnosis and treatment of obesity in adults: an applied evidence-based review. J Am Board Fam Pract 2004;17(5):359-369 [FREE Full text] [doi: 10.3122/jabfm.17.5.359] [Medline: 15355950]
59. Currenti W, Godos J, Castellano S, Caruso G, Ferri R, Caraci F, et al. Time-restricted feeding is associated with mental health in elderly Italian adults. Chronobiol Int 2021 Oct;38(10):1507-1516. [doi: 10.1080/07420528.2021.1932998] [Medline: 34100325]
60. Lee C, Raffaghello L, Brandhorst S, Safdie FM, Bianchi G, Martin-Montalvo A, et al. Fasting cycles retard growth of tumors and sensitize a range of cancer cell types to chemotherapy. Sci Transl Med 2012 Mar 07;4(124):124ra27 [FREE Full text] [doi: 10.1126/scitranslmed.3003293] [Medline: 22323820]
61. Marinac CR, Nelson SH, Breen CI, Hartman SJ, Natarajan L, Pierce JP, et al. Prolonged nightly fasting and breast cancer prognosis. JAMA Oncol 2016 Aug 01;2(8):1049-1055 [FREE Full text] [doi: 10.1001/jamaoncol.2016.0164] [Medline: 27032109]
62. Jane M, Hagger M, Foster J, Ho S, Pal S. Social media for health promotion and weight management: a critical debate. BMC Public Health 2018 Jul 28;18(1):932 [FREE Full text] [doi: 10.1186/s12889-018-5837-3] [Medline: 30055592]
63. He C, Wu S, Zhao Y, Li Z, Zhang Y, Le J, et al. Social media-promoted weight loss among an occupational population: cohort study using a WeChat mobile phone app-based campaign. J Med Internet Res 2017 Oct 23;19(10):e357 [FREE Full text] [doi: 10.2196/jmir.7861] [Medline: 29061555]

Abbreviations

EFH: extended fasting hours
IF: intermittent fasting
LFT: LIFE Fasting Tracker
LX: LIFE Extend

Edited by L Buis; submitted 22.12.21; peer-reviewed by J May, N Maglaveras; comments to author 19.02.22; revised version received 14.07.22; accepted 26.07.22; published 04.10.22.

Please cite as:
Torres L, Lee JL, Park S, Di Lorenzo RC, Branam JP, Fraser SA, Salisbury BA
Retention, Fasting Patterns, and Weight Loss With an Intermittent Fasting App: Large-Scale, 52-Week Observational Study
JMIR Mhealth Uhealth 2022;10(10):e35896
URL: https://mhealth.jmir.org/2022/10/e35896
doi:10.2196/35896
PMID:36194463

©Luisa Torres, Joy L Lee, Seho Park, R Christian Di Lorenzo, Jonathan P Branam, Shelagh A Fraser, Benjamin A Salisbury. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 04.10.2022. 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.

XSL•FO
RenderX

Original Paper

Validity of Chatbot Use for Mental Health Assessment: Experimental Study

Anita Schick[1], Dipl-Psych, PhD; Jasper Feine[2], MSc, PhD; Stefan Morana[3], PhD; Alexander Maedche[2], PhD; Ulrich Reininghaus[1], Dipl-Psych, MSc, PhD

[1]Department of Public Mental Health, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany
[2]Institute of Information Systems and Marketing, Karlsruhe Institute of Technology, Karlsruhe, Germany
[3]Junior Professorship for Digital Transformation and Information Systems, Saarland University, Saarbruecken, Germany

Corresponding Author:
Anita Schick, Dipl-Psych, PhD
Department of Public Mental Health, Central Institute of Mental Health
Medical Faculty Mannheim
Heidelberg University
Square J5
Mannheim, 68219
Germany
Phone: 49 62117031941
Email: anita.schick@zi-mannheim.de

Abstract

Background: Mental disorders in adolescence and young adulthood are major public health concerns. Digital tools such as text-based conversational agents (ie, chatbots) are a promising technology for facilitating mental health assessment. However, the human-like interaction style of chatbots may induce potential biases, such as socially desirable responding (SDR), and may require further effort to complete assessments.

Objective: This study aimed to investigate the convergent and discriminant validity of chatbots for mental health assessments, the effect of assessment mode on SDR, and the effort required by participants for assessments using chatbots compared with established modes.

Methods: In a counterbalanced within-subject design, we assessed 2 different constructs—psychological distress (Kessler Psychological Distress Scale and Brief Symptom Inventory-18) and problematic alcohol use (Alcohol Use Disorders Identification Test-3)—in 3 modes (chatbot, paper-and-pencil, and web-based), and examined convergent and discriminant validity. In addition, we investigated the effect of mode on SDR, controlling for perceived sensitivity of items and individuals' tendency to respond in a socially desirable way, and we also assessed the perceived social presence of modes. Including a between-subject condition, we further investigated whether SDR is increased in chatbot assessments when applied in a self-report setting versus when human interaction may be expected. Finally, the effort (ie, complexity, difficulty, burden, and time) required to complete the assessments was investigated.

Results: A total of 146 young adults (mean age 24, SD 6.42 years; n=67, 45.9% female) were recruited from a research panel for laboratory experiments. The results revealed high positive correlations (all P<.001) of measures of the same construct across different modes, indicating the convergent validity of chatbot assessments. Furthermore, there were no correlations between the distinct constructs, indicating discriminant validity. Moreover, there were no differences in SDR between modes and whether human interaction was expected, although the perceived social presence of the chatbot mode was higher than that of the established modes (P<.001). Finally, greater effort (all P<.05) and more time were needed to complete chatbot assessments than for completing the established modes (P<.001).

Conclusions: Our findings suggest that chatbots may yield valid results. Furthermore, an understanding of chatbot design trade-offs in terms of potential strengths (ie, increased social presence) and limitations (ie, increased effort) when assessing mental health were established.

(JMIR Mhealth Uhealth 2022;10(10):e28082) doi:10.2196/28082

XSL•FO
RenderX

KEYWORDS

chatbot; distress; monitoring; mobile health; social desirability; social presence

Introduction

Background

Mental disorders are a leading cause of disease burden in high-income countries and first emerge in adolescence and young adulthood [1]. Thus, mental health in young people is a major public health concern [2]. However, psychological help remains difficult to access [3]. To address this problem, digital technologies provide a scalable alternative for accessing low-threshold psychological assessments, digital diagnostics, and interventions [4]. In particular, digital technologies can support the early detection of symptoms, diagnostics, and treatment as they may improve access to mental health services for difficult-to-reach populations without requiring on-site visits using desktop PCs, tablets, or mobile devices [5].

Text-based conversational agents (ie, chatbots) are a promising digital technology in this context [6-12]. Chatbots interact with users via natural language [13], keeping individuals engaged in the task at hand, thereby increasing adherence [10,14]. Chatbots as software-based systems enabling asynchronous interactions have received increasing attention during the COVID-19 pandemic to provide information about infection numbers, rules, and restrictions [15], thereby improving health literacy and reducing the burden on the health care system. In addition, chatbots have been investigated in several studies and applied to assess or monitor mental health [16], deliver information for improving mental health literacy [9,14,15,17], and assist and compound therapy sessions as guided or blended care [18-22]. Irrespective of the popularity of chatbots, reviews of their application in the context of (mental) health emphasize the quasi-experimental nature of studies and the need to empirically evaluate their impact [7,16,23-26]. Specifically, for wider application, the extent to which a new mode for assessing a construct (eg, chatbots assessing psychological distress) converges with established assessment modes of the same construct (ie, the convergent validity) needs to be demonstrated. In addition, discriminant validity (ie, the extent to which a construct can be distinguished from another, unrelated construct) needs to be examined. However, to date, no study has specifically examined the validity of chatbot use in assessing mental health.

This is particularly relevant, as there is evidence that individuals preconsciously attribute human characteristics to chatbots because of increased perceived social presence [27-30]. Social presence can be defined as “the degree of salience of the other person in a mediated communication and the consequent salience of their interpersonal interactions” [31]. Thus, individuals may feel a sense of personal, sociable, and sensitive human contact during a computer-mediated interaction. Although an increase in perceived social presence in face-to-face interviews has been found to increase response biases [32-35], self-reported assessments associated with reduced social presence have demonstrated reliability and validity compared with, for example, face-to-face assessments [36-40]. However, the natural language interaction style of chatbots may yield response biases such as socially desirable responding (SDR) [32,41,42], where participants disclose less socially sensitive information, which might be of special interest when applying for mental health assessment.

Previous evidence indicates that SDR may increase when individuals expect their responses to be immediately reviewed and evaluated by a researcher [33,43,44]. If chatbots are perceived as human actors [42,45], this may lead individuals to believe that their responses are immediately reviewed and evaluated. This may bias the results compared with web-based assessments that are not presented with a natural language interface and would limit the application of chatbots in remote settings, in which information is not immediately shared with a clinician. Consequently, it is necessary to investigate whether SDR is increased in settings where individuals do or do not expect their responses to be immediately reviewed when assessed by chatbots.

Finally, there is evidence that chatbots may not necessarily reduce participants’ efforts to complete the assessments [46,47]. Although the completion of assessments delivered via established assessment modes is simple (eg, by ticking a box or clicking a button), chatbots require more complex natural language interactions. This may increase the cognitive resources and duration required for assessments using chatbots [46,47]. Thus, it is necessary to investigate whether individuals using a chatbot perceive assessments as more effortful (ie, as being more complex, difficult, and associated with more burden), as well as whether they require more time to complete assessments than when using established modes.

Objectives

This study aimed to investigate (1) the convergent and discriminant validity of assessments using chatbots, (2) the effect of assessments using chatbots on SDR, and (3) the effort of assessments using chatbots compared with established paper-and-pencil and web-based assessment modes. Specifically, we proposed the following hypotheses: chatbots applied to assess mental health (ie, psychological distress and problematic alcohol use) in healthy young adults will show high convergent validity with established assessment modes and high discriminant validity (hypothesis 1); increase SDR compared with established assessment modes (hypothesis 2a); increase SDR compared with established modes, especially in settings where individuals do not expect their responses to be immediately reviewed by the research team (hypothesis 2b); and be perceived as more effortful (ie, complex, difficult, and associated with more burden) and will require more time to complete than established assessment modes (hypothesis 3).

Methods

Experimental Design

A laboratory experiment applying a randomized mixed design with 3 within-subject conditions and 2 between-subject

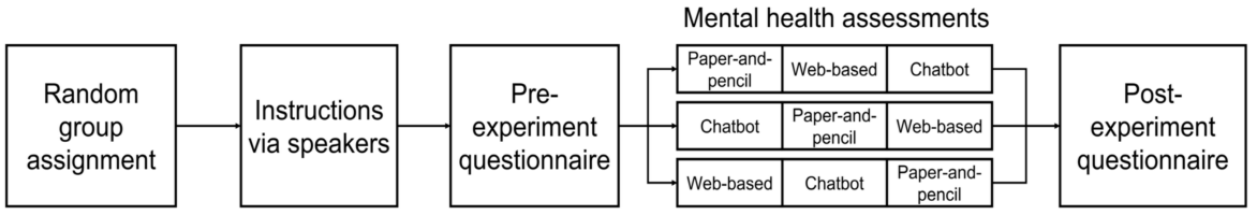
XSL•FO
RenderX

conditions was conducted. The within-subject manipulation comprised three assessment modes: (1) paper-and-pencil mode, (2) desktop computer using a typical web-based screening mode (web-based), and (3) assessment on a desktop computer screen using a chatbot (chatbot). For the between-subject manipulation, we randomly assigned participants to two conditions: participants in condition A (low-stake condition) were informed that their responses were not immediately reviewed by the research team, and participants in condition B (high-stake condition) were informed that their responses were immediately reviewed and may require a follow-up interaction with the research team.

Procedure and Manipulation

The experimental procedure is illustrated in Figure 1. First, participants were assigned to 1 of the 2 conditions. We conducted 6 experimental sessions on 2 consecutive days, with 3 sessions assigned to condition A (low-stake condition) and 3 sessions assigned to condition B (high-stake condition). After signing the informed consent form, participants were seated in front of a desktop computer screen in single air-conditioned and soundproof test chambers. Second, participants listened to a prerecorded voice message explaining the experimental procedure and the instructions. Participants in condition B were informed of their individual participation numbers. The number was displayed on the computer screen throughout the experiment: in the web-based mode, LimeSurvey [48] displayed the participant number at the top of the screen; in the paper-and-pencil mode, participants had to write their participant number on the questionnaire; and in the chatbot mode, participants were addressed with their participant number (ie, “Hello participant 324352”) displayed in the chat window below their responses.

Figure 1. Experimental procedure.

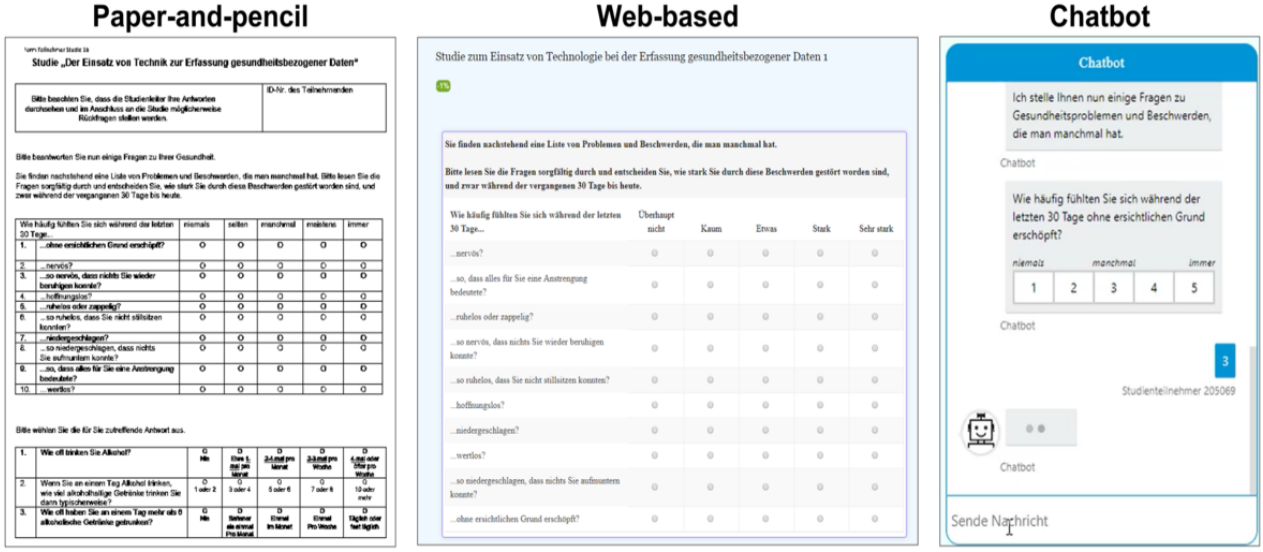


Next, the computer screen was automatically turned on, and the experiment began with a pre-experiment questionnaire using LimeSurvey [48]. Subsequently, mental health was assessed using the 3 different modes in a counterbalanced order (Figure 2). The web-based mode used the default LimeSurvey question format. The paper-and-pencil mode comprised a printout of the digital version, which was placed in an envelope in each chamber. After completing the paper-and-pencil mode, the participants were asked to place the questionnaire in the envelope and seal the envelope with adhesive tape. The chatbot mode was developed using the Microsoft Bot Framework [49] and was integrated into LimeSurvey. The chatbot presented the items one after another and offered 2 ways of responding, either by natural language or by selecting a value (implemented as a button). The chatbot incorporated the following social cues to further increase perceived social presence [28,30]: an anthropomorphic icon [50], the capability to engage in small talk [51], a dynamically calculated response delay based on the length of the response [30], and a typing indicator (3 moving dots indicating that a message is being prepared) [52]. Microsoft’s personality chat small talk package was used to enable a small talk interaction. This knowledge base was implemented in Microsoft’s QnA Maker and was connected to the chatbot. When the QnA model identified a high match with an incoming user message, the chatbot answered with an appropriate small talk phrase. However, the chatbot’s capabilities were restricted, and no sophisticated conversations were possible. For example, the small talk included greetings such as “Hi/Hello/Good Morning!” and “How are you?”; however, the small talk did not account for the context. After answering with a small talk phrase, the chatbot always repeated the prior question. In addition, we did not record the log files of the chats. On the continuum of machine-like to human-like appearance, we chose an intermediate design to avoid the induction of negative affect toward the chatbot, which has been postulated for the increased human-likeness of robots according to the uncanny valley theory by Mori [53]. In addition, we chose the name indicator *Chatbot*, as robotic names have been reported to be positively perceived [6].

Finally, the participants answered a postexperiment questionnaire using LimeSurvey. They were then debriefed and received their compensation.

XSL•FO
RenderX

Figure 2. Investigated assessment modes (displayed in German).



Measures

In the pre-experiment questionnaire, we assessed demographic variables (eg, sex, age, and education), followed by questions on participants' prior experience with using specific technologies (ie, internet and chatbots) with regard to health questions. Next, their experience with paper-and-pencil and web-based surveys, as well as with chatbots, was assessed on a scale ranging from 1 (no experience) to 5 (very much experience).

Balanced Inventory of Desirable Responding

On the one hand, we applied the short form of the Balanced Inventory of Desirable Responding (BIDR) scale, which comprises two subscales: self-deceptive enhancement and impression management [54,55] to capture SDR. The 18 items were rated on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). We calculated the total score for each scale and the BIDR total score, which ranged from 1 to 126.

On the other hand, we operationalized SDR as a response shift; that is, a change in participant's mental health scores between repeated assessments in different modes.

Mental Health Measures

Mental health was assessed using the following measures in all 3 modes.

Kessler Psychological Distress Scale

Psychological distress in the past month was measured using the Kessler Psychological Distress Scale (K10) [56]. This 10-item self-report questionnaire is rated on a Likert scale ranging from 1 (never) to 5 (always). The K10 total score was calculated. Strong psychometric properties of the K10 have been reported [56].

Brief Symptom Inventory

We used the short form of the Brief Symptom Inventory (BSI-18) [57,58] to assess psychological distress in the past 7 days. Participants indicated whether they had experienced 18 symptoms, comprising 3 dimensions: somatization, depression, and anxiety. The items were rated on a scale from 1 (not at all) to 5 (very much). We calculated the total score indicating general distress (BSI–General Severity Index) [58].

Alcohol Use Disorders Identification Test-3

We assessed alcohol use by applying the Alcohol Use Disorders Identification Test (AUDIT)–3 questionnaire [59,60], which has been shown to perform similarly well as the AUDIT-10 in detecting risky drinking behavior [60]. The items were presented on a 5-point scale with different labels asking about the amount of alcohol consumption. The total AUDIT-3 score was calculated.

The time at the beginning and end of data collection in each mode was recorded. In the postexperiment questionnaire, participants had to rank the 3 modes regarding complexity, difficulty, and burden. Subsequently, we asked participants to rate others' discomfort when answering each item of the mental health measures, thereby deriving a measure of subjective sensitivity in line with Bradburn et al [61].

Attention and Manipulation Checks

In the attention check, participants had to select a specific item on a Likert scale to verify that they carefully followed the instructions ("Please select the answer very often"). To test the within-subject manipulation, we investigated differences in the perceived social presence of each mode using the 4 items by Gefen and Straub [62], which were rated on a 7-point Likert scale. The internal consistency of the perceived social presence of the 3 modes was high (Cronbach α>.89).

Furthermore, participants had to indicate in the postexperiment questionnaire whether their answers were immediately reviewed, in line with Fisher [44] (between-subject manipulation check).

Power Analysis and Recruitment

An a priori analysis in G*Power software (Heinrich-Heine-Universität Düsseldorf) [63] estimated a total sample size of 116 (α=.05; f=0.15; Cohen d=0.95). For recruitment, we invited individuals registered with the university's research panel, comprising mainly students from

XSL•FO
RenderX

the Karlsruhe Institute of Technology. The experiment lasted 45 minutes on average and participants were compensated for their participation with €8 (US $8.06) after the experiment.

Statistical Analysis

SPSS Statistics (version 25; IBM Corp) and STATA (version 16.0; StataCorp) were used to analyze the data. Participant characteristics were summarized using means and SDs for continuous variables and frequencies and percentages for dichotomous variables. To investigate differences between groups, we calculated the ANOVAs for individuals' tendency to respond as socially desirable (BIDR) and the perceived sensitivity of each measure (K10, BSI-18, and AUDIT-3). Furthermore, differences between prior experience with, as well as the perceived social presence of, modes were investigated by calculating repeated-measures ANOVAs (rmANOVAs). As data on prior experience (χ^2_2=46.4; P<.001) and perceived social presence (χ^2_2=49.5; P<.001) violated the assumptions of sphericity, Huynh-Feldt corrections were applied.

The internal consistency of the mental health measures for each mode was evaluated using Cronbach α. Next, the test-retest reliabilities of the chatbot-based, paper-and-pencil–based, and desktop-based assessment modes were evaluated by calculating intraclass correlation coefficients (ICCs) ranging from 0 (no agreement) to 1 (perfect agreement).

To test hypothesis 1 on the discriminant and convergent validity of assessment modes, we calculated Pearson correlations and applied Bonferroni correction to account for multiple testing. In line with the multitrait-multimethod approach by Campbell and Fiske [64], we tested 3 independent assessment modes with 2 different constructs—psychological distress (K10 and BSI-18) and problematic alcohol use (AUDIT-3)—to derive discriminant and convergent validity. Validity is indicated by a correlation coefficient of ≥0.50 [63].

To test hypothesis 2a, we calculated repeated-measures analyses of covariance (rmANCOVAs) with the within-subject factor mode (paper-and-pencil, web-based, and chatbot) and the following covariates: (1) perceived sensitivity of the items and (2) individuals' tendency to respond socially desirable (BIDR). Sex was also included as a control variable in all the analyses. Lavene test revealed the homogeneity of variances for all 3 measures. As the AUDIT-3 data violated the assumptions of sphericity (χ^2_2=13.2; P=.001), the Huynh-Feldt correction was applied in the rmANCOVA.

To test hypothesis 2b, rmANCOVAs with the within-subject factor mode (paper-and-pencil, web-based, and chatbot) and condition (A and B) as additional covariates were calculated. Lavene test revealed the homogeneity of variances for all modes. Again, the AUDIT-3 data violated the assumption of sphericity (χ^2_2=13.4; P=.001), and the Huynh-Feldt correction was applied.

To test hypothesis 3 on the effort of assessment, we analyzed the ranked-ordered data on complexity, difficulty, and burden by calculating Friedman tests and Dunn-Bonferroni post hoc signed-rank tests for pairwise comparisons. Differences in the duration to complete the assessments were investigated by calculating rmANOVAs with the within-subject factor mode (paper-and-pencil, web-based, and chatbot). As the data violated the assumptions of sphericity (χ^2_2=9.1; P=.01), the Huynh-Feldt correction was applied.

Ethics Approval

The experiment took place at the Karlsruhe Decision and Design Lab, adhering to its procedural and ethical guidelines. No ethics approval was applied for as participants were recruited from the registered participant panel of healthy students. Individuals voluntarily participated after being fully informed about the study procedures and signing the informed consent form. No identifying data were collected.

Results

Sample Characteristics

We invited all individuals registered in the university's research panel to participate in the experiment. A total of 155 individuals participated in the study, of whom 9 (5.8%) participants were excluded as they failed the attention check, indicating that they may not have followed the instructions of the experiment or had not read the individual items carefully. Consequently, 146 participants were included in the analysis, of whom 72 (49.3%) were in condition A and 74 (50.7%) were in condition B.

The sample characteristics and control variables are presented in Table 1. Overall, we investigated a sample of young students from which most participants had a high school or bachelor's degree. In addition, two-thirds of the participants (100/146, 68.5%) indicated that they had used the internet to access information on mental health before. However, only 4.1% (6/146) of participants replied having interacted with a chatbot in a health-related context before. Prior experience with assessment modes differed across the 3 modes, as revealed by the rmANOVA ($F_{1.58, 229.39}$=225.23; P<.001). Post hoc analyses with a Bonferroni adjustment further showed that the experience with chatbots (mean 1.73, SD 1.02) was lower than the experience with paper-and-pencil surveys (mean 3.45, SD 0.85), as well as the experience with web-based surveys (mean 3.52, SD 0.82, all P<.001). Experience with paper-and-pencil surveys did not significantly differ from that with web-based surveys (P=.78). Individuals' tendency to respond socially desirable, as measured using the BIDR, did not differ between conditions ($F_{1,144}$=0.131; P=.72) and was centered on the mean (W_{146}=0.98; P=.09). The perceived sensitivity of the items of the 3 mental health measures did not differ between the 2 conditions (all P>.47) but differed between the 3 measures ($F_{1.41, 88.22}$ =105.64; P<.001). Post hoc analyses with Bonferroni adjustment indicated that AUDIT-3 items (mean 3.39, SD 1.07) were rated as more sensitive than K10 items (mean 2.59, SD 0.66; P<.001), as well as BSI-18 items (mean 2.33, SD 2.33, P<.001). Furthermore, the K10 items (mean 2.59, SD 0.66) were perceived to be more sensitive than the BSI-18 items (mean 2.33, SD 0.58; P<.001).

XSL•FO
RenderX

Table 1. Sample characteristics (N=146).

Variable	Full sample	Low-stake condition (n=72)	High-stake condition (n=74)
Age (years), mean (SD)	24.2 (6.42)	23.44 (6.06)	24.93 (6.71)
Female, n (%)	67 (45.9)	30 (41.7)	37 (50)
Education, n (%)			
Middle school	3 (2.1)	2 (2.8)	1 (1.4)
High school	89 (60.9)	43 (59.7)	46 (62.2)
Bachelor's	46 (31.5)	25 (34.7)	21 (28.4)
Master's	8 (5.5)	2 (2.8)	6 (8.1)
Technology experience[a], n (%)			
Internet	100 (68.5)	51 (70.8)	49 (66.2)
Chatbot	6 (4.1)	2 (2.8)	4 (5.4)
Survey experience, mean (SD)			
Paper-and-pencil	3.45 (0.85)	3.53 (0.87)	3.36 (0.82)
Web-based	3.52 (0.82)	3.57 (0.77)	3.47 (0.88)
Chatbot	1.73 (1.02)	1.64 (0.86)	1.82 (1.15)
Social desirability, mean (SD)			
BIDR[b] total	83.60 (9.38)	83.32 (9.15)	83.89 (9.67)
BIDR-SDE[c]	41.55 (5.00)	41.65 (4.62)	41.46 (5.39)
BIDR-IM[d]	42.05 (6.93)	41.68 (7.06)	42.43 (6.82)
Sensitivity of measures, mean (SD)			
K10[e]	2.59 (0.66)	2.61 (0.71)	2.57 (0.62)
BSI-18[f]	2.33 (0.58)	2.34 (0.58)	2.33 (0.57)
AUDIT-3[g]	3.39 (1.07)	3.45 (1.07)	3.32 (1.08)

[a]Number of participants who previously used technology in a health-related context.

[b]BIDR: Balanced Inventory of Desirable Responding.

[c]BIDR-SDE: Balanced Inventory of Desirable Responding–Self-deceptive enhancement.

[d]BIDR-IM: Balanced Inventory of Desirable Responding–Impression management.

[e]K10: Kessler Psychological Distress Scale.

[f]BSI-18: Brief Symptom Inventory-18.

[g]AUDIT-3: Alcohol Use Disorders Identification Test-3.

Manipulation Checks

With regard to the within-subject manipulation, the results of the rmANOVA revealed a significant effect of mode on perceived social presence ($F_{1.56,226.67}$=61.96; P<.001), with social presence rated highest in the chatbot mode (mean 2.74, SD=1.51) compared with the web-based mode (mean 1.48, SD 0.88; P<.001) and paper-and-pencil mode (mean 1.79, SD 1.21; P<.001).

Responses to the between-subject manipulation check showed that 93.2% (136/146) of participants provided a correct answer—2.7% (4/146) of individuals with wrong answers were in condition A and 4.1% (6/146) were in condition B—and were aware of their condition. Consequently, we concluded that both within-subject and between-subject manipulations were successful.

Reliability of Chatbots for Mental Health Assessments

Table 2 displays the mean, SD, Cronbach α, and ICC for the mental health measures in each mode by condition. The ICCs of the paper-based, desktop-based, and chatbot modes were high and ranged between 0.96 and 1.00, indicating excellent agreement across modes and a high test-retest reliability. Cronbach α did not strongly vary between modes and ranged between 0.74 and 0.92, indicating an acceptable to excellent internal consistency of the measures.

XSL•FO
RenderX

Table 2. Internal consistency and test-retest reliability of mental health assessments.

Measure and mode	Full sample		Low-stake condition		High-stake condition		ICC[a]
	Values, mean (SD)	Cronbach α	Values, mean (SD)	Cronbach α	Values, mean (SD)	Cronbach α	
K10[b]							0.96
Paper-based	19.36 (6.53)	.89	19.44 (5.66)	.84	19.28 (7.31)	.92	
Web-based	19.77 (6.67)	.88	19.47 (5.63)	.82	20.05 (7.57)	.91	
Chatbot-based	19.7 (6.45)	.86	19.43 (5.81)	.82	19.95 (7.04)	.89	
BSI-18[c]							0.99
Paper-based	11.54 (8.45)	.86	11.35 (6.72)	.78	11.73 (9.9)	.9	
Web-based	11.56 (8.89)	.87	11.29 (7.48)	.82	11.81 (10.12)	90	
Chatbot-based	11.09 (8.4)	.86	10.71 (7.09)	.8	11.46 (9.54)	.89	
AUDIT-3[d]							1.00
Paper-based	3.42 (2.45)	.80	3.50 (2.60)	.85	3.34 (2.30)	.74	
Web-based	3.40 (2.44)	.81	3.49 (2.62)	.86	3.32 (2.28)	.75	
Chatbot-based	3.43 (2.49)	.82	3.49 (2.64)	.86	3.38 (2.36)	.76	

[a]ICC: intraclass correlation coefficient.

[b]K10: Kessler Psychological Distress Scale.

[c]BSI-18: Brief Symptom Inventory-18.

[d]AUDIT-3: Alcohol Use Disorders Identification Test-3.

Validity of Assessments Using Chatbots (Hypothesis 1)

As depicted in Table 3, there were strong positive correlations between the measures of psychological distress (K10 and BSI-18) assessed by the different modes, with correlation coefficients ranging from 0.83 to 0.96, indicating convergent validity. Furthermore, there were strong positive correlations between the AUDIT-3 scores assessed using the different modes. There were no significant correlations among AUDIT-3, K10, and BSI-18 after Bonferroni correction, indicating discriminant validity between the different constructs.

XSL•FO
RenderX

Table 3. Pearson correlation of questionnaires and modes. Higher numbers reflect a stronger association between variables.

Mode	K10[a]			BSI-18[b]			AUDIT-3[c]		
	Paper-based *r* (*P* value[d])	Web-based *r* (*P* value)	Chatbot-based *r* (*P* value)	Paper-based *r* (*P* value)	Web-based *r* (*P* value)	Chatbot-based *r* (*P* value)	Paper-based *r* (*P* value)	Web-based *r* (*P* value)	Chatbot-based *r* (*P* value)
K10									
Paper-based	1	0.89 (<.001)	0.88 (<.001)	0.89 (<.001)	0.83 (<.001)	0.85 (<.001)	−0.1 (.21)	−0.12 (.14)	−0.13 (.12)
Web-based	0.89 (<.001)	1	0.87 (<.001)	0.88 (<.001)	0.89 (<.001)	0.86 (<.001)	−0.18 (.04)	−0.19 (.02)	−0.20 (.02)
Chatbot-based	0.88 (<.001)	0.87 (<.001)	1	0.85 (<.001)	0.84 (<.001)	0.85 (<.001)	−0.09 (.27)	−0.11 (.17)	−0.12 (.16)
BSI-18									
Paper-based	0.89 (<.001)	0.88 (<.001)	0.85 (<.001)	1	0.96 (<.001)	0.96 (<.001)	−0.1 (.22)	−0.12 (.15)	−0.14 (.10)
Web-based	0.83 (<.001)	0.89 (<.001)	0.84 (<.001)	0.96 (<.001)	1	0.96 (<.001)	−0.14 (.09)	−0.16 (.06)	−0.18 (.04)
Chatbot-based	0.85 (<.001)	0.86 (<.001)	0.85 (<.001)	0.96 (<.001)	0.96 (<.001)	1	−0.15 (.07)	−0.16 (.05)	−0.17 (.04)
AUDIT-3									
Paper-based	−0.1 (.21)	−0.18 (.04)	−0.09 (.27)	−0.1 (.22)	−0.14 (.09)	−0.15 (.07)	1	0.99 (<.001)	0.99 (<.001)
Web-based	−0.12 (.14)	−0.19 (.02)	−0.11 (.17)	−0.12 (.15)	−0.16 (.06)	−0.16 (.05)	0.99 (<.001)	1	0.99 (<.001)
Chatbot-based	−0.13 (.12)	−0.20 (.02)	−0.12 (.16)	−0.14 (.10)	−0.18 (.04)	−0.17 (.04)	0.99 (<.001)	0.99 (<.001)	1

[a]K10: Kessler Psychological Distress Scale.

[b]BSI-18: Brief Symptom Inventory-18.

[c]AUDIT-3: Alcohol Use Disorders Identification Test-3.

[d]Unadjusted *P* value; the Bonferroni corrected significance level was computed by dividing the unadjusted *P* value by the total number of tests; that is, *P*=.05/45=.0011.

SDR to Chatbots in Mental Health Assessments (Hypotheses 2a and 2b)

Addressing hypothesis 2a, the rmANCOVA on the effect of mode on mental health assessment revealed no main effect of mode on K10 ($F_{2,284}$=0.35; *P*=.71). Moreover, there was no interaction between mode and social desirability ($F_{2,284}$=0.80; *P*=.45) or perceived sensitivity of the items ($F_{2,284}$=0.43; *P*=.65); however, there was a significant interaction with sex ($F_{2,284}$=3.21; *P*=.04). The second mental distress measure, the BSI-18, showed similar results. The rmANCOVA revealed no significant main effect of mode on general distress ($F_{2,248}$=0.90; *P*=.41). Again, there was no interaction between mode and social desirability ($F_{2,284}$=1.7; *P*=.19), sensitivity ($F_{2,284}$=0.23; *P*=.80), or sex ($F_{2,284}$=2.66; *P*=.07). Similarly, the rmANCOVA on AUDIT-3 scores revealed no significant main effect of mode ($F_{1.90,269.57}$=0.00; *P*=1.00), as well as no interaction of mode with social desirability ($F_{1.90,269.57}$=0.01; *P*=.99), perceived sensitivity of items ($F_{1.90,269.57}$=0.24; *P*=.77), or sex ($F_{1.90,269.57}$=0.33; *P*=.71).

The effect of the condition on mental health assessment (hypothesis 2b) was investigated using a second set of rmANCOVAs. The results revealed no significant interaction effect between mode and condition on psychological distress assessed by K10 ($F_{2,282}$=0.91; *P*=.41), general distress assessed using the BSI ($F_{2,282}$=0.29; *P*=.75), or alcohol use assessed by AUDIT-3 ($F_{1.91,\ 269.14}$=0.55; *P*=.57).

Difficulty of Assessments Using Chatbots (Hypothesis 3)

Table 4 shows the mean rating of complexity, difficulty, and burden. A Friedman test revealed a significant difference between the difficulty associated with the modes (χ^2_2=13.5; *P*=.001). Dunn-Bonferroni post hoc tests showed that the assessment by a chatbot was rated as significantly more difficult than using the paper-and-pencil mode (*z*=3.63; *P*=.001). Furthermore, there was a statistically significant difference in perceived complexity depending on the mode (χ^2_2=10.15; *P*=.006). Again, Dunn-Bonferroni post hoc tests showed that the chatbot assessment was ranked as more complex than the paper-and-pencil assessment (*z*=3.16; *P*=.005). In terms of burden, a Friedman test indicated that there was a statistically

XSL•FO
RenderX

significant difference (χ^2_2=12.4; *P*=.002), and Dunn-Bonferroni post hoc tests further revealed that the web-based assessment required significantly less effort than the chatbot (*z*=2.64; *P*=.03) and the paper-and-pencil assessment (*z*=−3.34; *P*=.003). The analysis of duration revealed a significant effect of mode ($F_{1.91, 276.68}$=186.60; *P*<.001). Post hoc analyses with Bonferroni adjustment revealed that the pairwise differences between all modes were significant (*P*<.001). The longest duration was logged to complete the chatbot assessment and the shortest duration was required to complete the web-based assessment.

Table 4. Effort of assessment modes.

Effort variable and mode	Rank, mean (SD)
Complexity	
Paper-and-pencil	1.80 (0.84)
Web-based	2.03 (0.66)
Chatbot	2.17 (0.89)
Difficulty	
Paper-and-pencil	1.81 (0.78)
Web-based	1.96 (0.7)
Chatbot	2.23 (0.9)
Burden	
Paper-and-pencil	2.16 (0.79)
Web-based	1.77 (0.73)
Chatbot	2.08 (0.87)
Duration (seconds)	
Paper-and-pencil	184.62 (79.28)
Web-based	128.78 (56.07)
Chatbot	265.1 (65.82)

Discussion

Principal Findings

This study examined the validity, effect on SDR, and effort required for the completion of chatbot-based assessments of mental health. The results revealed that all assessments of mental health (K10, BSI, and AUDIT) in each mode showed acceptable to excellent internal consistency and high test-retest reliability. High positive correlations between the measures of the same construct across different assessment modes indicated the convergent validity of the chatbot mode, and the absence of correlations between distinct constructs indicated discriminant validity (hypothesis 1). Although assessment modes were not affected by social desirability (hypothesis 2a), chatbot assessment was higher for perceived social presence. There was no evidence of an interaction between condition and mode, indicating that social desirability did not increase because of expectations around immediate follow-up contact with a researcher in the chatbot assessment mode (hypothesis 2b). Finally, in terms of participants' effort (hypothesis 3), the assessment using a chatbot was found to be more complex, difficult, and associated with more burden than the established modes, resulting in a longer duration to complete.

Limitations

The present findings must be considered in light of several limitations. First, the selection of a student sample may have resulted in the low external validity of the laboratory experiment. According to previous mental health assessments in the general population, our sample showed only moderate distress [65]. There is evidence that individuals disclose more information on sensitive topics such as health risk behavior in clinical settings [66]. Future research should further investigate the application of chatbots in clinical samples, as the present findings on social desirability or perceived social presence of chatbots do not readily generalize to clinical populations.

Second, we reduced the effect of between-person differences by selecting a within-person design, which had several limitations. Each participant completed questionnaires in all 3 modes, with an average break between modes of approximately 1 minute. During the break, participants rated their social presence and read the instructions in the next experimental section. The break may have been too short to minimize memory effects. In addition, all measures used Likert scales, which may have increased memory effects because of their simplicity. To address this limitation, we completely counterbalanced the order of the 3 modes in the experimental procedure. Furthermore, in a sensitivity analysis using data from only the first mode presented to the participants, we did not find any differences, which further supports the reported results (Multimedia Appendix 1, Table S1). However, other factors such as the need for consistent responses may have overcome social desirability. Again, a longer break between assessments or a between-subject design could be applied in future experiments.

XSL•FO
RenderX

Third, the lack of an effect of mode on change in mental health scores may have been a result of the experimental design or chatbot design. As mentioned previously, we did not assess social pressure; however, individuals showed stronger SDR in high-stakes assessment situations. Thus, the assessment of social pressure is recommended for future studies. Furthermore, in this experiment, the chatbot followed a procedural dialog flow using Likert scales and, in addition to basic small talk capabilities using several social cues [30], was unable to answer questions about topics other than the assessments. Although we demonstrated a higher perceived social presence of the chatbot, this may not have been sufficient to resemble the communication flow of a human interviewer. In addition, the perceived social presence of the chatbot may have led to increased expectations of participants in terms of the chatbot's interactivity and natural language capabilities [28]. Thus, the chatbot may have raised expectations that may not have been met [67]. Consequently, future research should investigate different chatbot designs that support less restricted non–goal-oriented natural language interactions. In this regard, further experiments should evaluate the influence of social and empathic responses on mental health assessments.

Fourth, this study investigated the convergent and discriminant validity of measures and modes to assess the constructs of psychological distress and alcohol use. We aimed to reduce the participant burden by selecting only 3 measures of mental health. However, other even less related constructs could have been investigated to facilitate the evaluation of discriminant validity. This issue should be addressed in future research.

Finally, the longer duration of completing the assessment using a chatbot may have resulted from participants potentially entering their responses by typing or using the menu option. In this study, we did not assess the method of entering data that was used. In future research, either one response option should be favored or the 2 response options may be compared by applying a microrandomized design.

Comparison With Prior Work

The use of chatbots for mental health assessment is an emerging field, and robust investigations of their positive and potential negative effects are required [16]. Given that recent studies have shown the feasibility of the application of chatbots in general, particularly in relation to monitoring [15], offering information on, as well as delivering interventions for, improving mental health [62,63], there is a need for methodological research on the use of chatbots in this context [7,16,23-26]. This appears to be particularly important in cases where chatbots may be seen as social actors (ie, human interviewers) evoking social desirability. Therefore, it needs to be shown that using chatbots for assessing mental health does not result in biased outcomes.

The application of chatbots has been previously shown to affect the collected data and either reduce [68-70] or increase [42] the SDR compared with assessments by human interviewers. Other studies have found that chatbot assessments may result in comparable results with established modes [8,46,71]. However, some studies have found this effect only in adult samples [72] or depending on the chatbot's visual and linguistic design [42,73]. In this context, chatbots with high conversational abilities or a more human-like embodiment have been shown to elicit more SDR to socially sensitive questions than established modes [42,73]. However, this was not the case when a chatbot with fewer human-like conversational abilities was presented [42,73], which is consistent with findings of this study. Thus, an assessment using a chatbot with the presented design and procedural dialog flow does not seem to induce additional SDR. Despite this finding, it may be of interest to develop chatbots with high conversational abilities as these may enhance adherence and increase compliance, for example, in digital interventions [8,11,21,24]. This is particularly important for delivering interventions and building stable human-chatbot interactions [51]. Therefore, further research on chatbots is required, for example, in which different conversational interaction strategies may be applied. A promising approach may be to enable reciprocal self-disclosure, in which the chatbot reveals sensitive information, as this has been shown to result in a reciprocal effect on promoting individuals' self-disclosure [70], as well as perceived intimacy and enjoyment [74]. Another promising approach may be the application of contingent interaction strategies, as individuals disclose more information on a website if contingent questions depending on previous interactions are displayed [75]. Moreover, voice-based conversational agents may improve response quality to sensitive questions [76]. However, more research on the design of voice-based conversational agents for mental health assessment is required [77]. In addition, unconstrained natural language input to conversational agents poses safety risks that must be evaluated thoroughly. As recently shown by Bickmore et al [78], voice-based assistants failed more than half of the time when presented with medical inquiries. Therefore, further evaluation of human-computer interactions and education about the capabilities of conversational agents is required.

In contrast to previous findings on assessments using chatbots reporting higher data quality or more engagement [8,9,11,47,69], we showed that chatbot assessments were more difficult, complex, and associated with more burden to complete than assessments using established modes. In addition, more time was required to complete the assessments. The latter has been previously shown [47] and may result from the increased cognitive demand of a communication flow, where an individual must decode and aggregate the impression-bearing and relational functions conveyed in computer-mediated communication [79]. In addition, increased effort may result from individual preferences or prior experiences with chatbots in other contexts. It has been shown that populations with high health literacy rates prefer established modes because of their efficiency and ability to proceed at their own pace [46]. This may be particularly relevant in a sample of young students. Furthermore, this finding is in line with the communication literature arguing that simple tasks may be conducted more efficiently through learner media [80]. Thus, simple tasks such as selecting Likert scale items in mental health questionnaires may be more efficiently conducted through the use of established modes such as paper-and-pencil or web-based assessments [81]. This may imply that the best application area of chatbots in mental health may not be symptom monitoring or screening but rather providing information or delivering an intervention in unstructured natural language interactions. Recent evidence

XSL•FO
RenderX

supports the use of chatbot-based interventions as they have been found to perform equally well as standard treatment methods (eg, face-to-face and telephone counseling) [7].

This work provides further evidence on the use of chatbots to assess mental health on site in clinics but also in asynchronous remote medical interactions (eg, at home) [17,70,82]. As the assessment modes between conditions did not differ, the results show that the application of a chatbot results in valid responses, regardless of whether the data are immediately reviewed and evaluated by a human actor [70,83]. Therefore, chatbots have the potential to reduce the workload in clinical settings by providing valid remote assessments, which is especially necessary for situations in which the medical system is at its limits. As stated by Miner et al [15], chatbots may be a digital solution that may help provide information, monitor symptoms, and even reduce psychosocial consequences during the COVID-19 pandemic. Recently, several chatbots for monitoring COVID-19 symptoms have been published, as reviewed by Golinelli et al [84]. In contrast to other mental health apps, chatbots have the advantage of providing communication that may additionally help to reduce loneliness during means of physical distancing [85,86]. For example, it has been shown that users may develop a strong social relationship with a chatbot when it expresses empathetic support [21,51,85,87-90]. Moreover, promising real-world examples of empathetic mental health chatbots have shown their effectiveness in practice, such as the mobile app chatbots Wysa [85], Woebot [6], and Replika [91]; however, they have also raised ethical concerns [10]. Thus, the application of chatbots in mental health research and practice may depend on the specific application (symptom monitoring vs guided intervention) and its potential advantages (ie, increased social presence) and disadvantages (ie, increased effort) while respecting users' privacy and safety.

Conclusions

These findings provide evidence of the validity of chatbots as digital technology for mental health assessment. In particular, when paper-and-pencil assessments are not applicable (eg, remote assessments in eHealth settings) or when it may be beneficial to increase perceived social presence (eg, to establish a long-term user-chatbot relationship), chatbots are promising alternatives for valid assessment of mental health without leading to socially desirable responses. However, as participants' efforts have increased, future research on appropriate chatbot designs and interaction flow is necessary to fully leverage their advantages in compounding digital care.

Acknowledgments

The authors would like to thank all the participants. This work was funded by a ForDigital grant from the Ministry of Science, Research, and Arts of the State of Baden-Württemberg, Germany. UR was supported by a Heisenberg professorship (number 389624707) funded by the German Research Foundation. The authors would like to thank the reviewers for their valuable comments on this manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Sensitivity analyses.
[DOCX File , 18 KB - mhealth_v10i10e28082_app1.docx]

References

1. Gore FM, Bloem PJ, Patton GC, Ferguson J, Joseph V, Coffey C, et al. Global burden of disease in young people aged 10-24 years: a systematic analysis. Lancet 2011 Jun 18;377(9783):2093-2102. [doi: 10.1016/S0140-6736(11)60512-6] [Medline: 21652063]
2. Patel V, Flisher AJ, Hetrick S, McGorry P. Mental health of young people: a global public-health challenge. Lancet 2007 Apr 14;369(9569):1302-1313. [doi: 10.1016/S0140-6736(07)60368-7] [Medline: 17434406]
3. D'Alfonso S, Santesteban-Echarri O, Rice S, Wadley G, Lederman R, Miles C, et al. Artificial intelligence-assisted online social therapy for youth mental health. Front Psychol 2017 Jun 2;8:796 [FREE Full text] [doi: 10.3389/fpsyg.2017.00796] [Medline: 28626431]
4. Zhang R. The stress-buffering effect of self-disclosure on Facebook: an examination of stressful life events, social support, and mental health among college students. Comput Human Behav 2017 Oct;75:527-537. [doi: 10.1016/j.chb.2017.05.043]
5. Uhlhaas P, Torous J. Digital tools for youth mental health. NPJ Digit Med 2019 Oct 18;2:104 [FREE Full text] [doi: 10.1038/s41746-019-0181-2] [Medline: 31646184]
6. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 06;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
7. Hoermann S, McCabe KL, Milne DN, Calvo RA. Application of synchronous text-based dialogue systems in mental health interventions: systematic review. J Med Internet Res 2017 Jul 21;19(8):e267 [FREE Full text] [doi: 10.2196/jmir.7023] [Medline: 28784594]

XSL•FO
RenderX

8. Celino I, Re Calegari G. Submitting surveys via a conversational interface: an evaluation of user acceptance and approach effectiveness. Int J Hum Comput Stud 2020 Jul;139:102410. [doi: 10.1016/j.ijhcs.2020.102410]
9. Bickmore TW, Utami D, Matsuyama R, Paasche-Orlow MK. Improving access to online health information with conversational agents: a randomized controlled experiment. J Med Internet Res 2016 Jan 04;18(1):e1 [FREE Full text] [doi: 10.2196/jmir.5239] [Medline: 26728964]
10. Kretzschmar K, Tyroll H, Pavarini G, Manzini A, Singh I, NeurOx Young People's Advisory Group. Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (Chatbots) in mental health support. Biomed Inform Insights 2019 Mar 5;11:1178222619829083 [FREE Full text] [doi: 10.1177/1178222619829083] [Medline: 30858710]
11. Ly KH, Ly AM, Andersson G. A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. Internet Interv 2017 Oct 10;10:39-46 [FREE Full text] [doi: 10.1016/j.invent.2017.10.002] [Medline: 30135751]
12. Greer S, Ramo D, Chang YJ, Fu M, Moskowitz J, Haritatos J. Use of the Chatbot "Vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: randomized controlled feasibility trial. JMIR Mhealth Uhealth 2019 Oct 31;7(10):e15018 [FREE Full text] [doi: 10.2196/15018] [Medline: 31674920]
13. McTear M, Callejas Z, Griol D. The Conversational Interface: Talking to Smart Devices. Cham, Switzerland: Springer International Publishing; 2016. URL: https://link.springer.com/book/10.1007/978-3-319-32967-3
14. Chaix B, Bibault JE, Pienkowski A, Delamon G, Guillemassé A, Nectoux P, et al. When chatbots meet patients: one-year prospective study of conversations between patients with breast cancer and a chatbot. JMIR Cancer 2019 May 02;5(1):e12856 [FREE Full text] [doi: 10.2196/12856] [Medline: 31045505]
15. Miner AS, Laranjo L, Kocaballi AB. Chatbots in the fight against the COVID-19 pandemic. NPJ Digit Med 2020 May 4;3:65 [FREE Full text] [doi: 10.1038/s41746-020-0280-0] [Medline: 32377576]
16. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J Med Internet Res 2020 Jul 13;22(7):e16021 [FREE Full text] [doi: 10.2196/16021] [Medline: 32673216]
17. Bibault JE, Chaix B, Guillemassé A, Cousin S, Escande A, Perrin M, et al. A chatbot versus physicians to provide information for patients with breast cancer: blind, randomized controlled noninferiority trial. J Med Internet Res 2019 Nov 27;21(11):e15787 [FREE Full text] [doi: 10.2196/15787] [Medline: 31774408]
18. Martínez-Miranda J. Embodied conversational agents for the detection and prevention of suicidal behaviour: current applications and open challenges. J Med Syst 2017 Sep;41(9):135. [doi: 10.1007/s10916-017-0784-6] [Medline: 28755270]
19. Pagliari C, Burton C, McKinstry B, Szentatotai A, David D, Serrano Blanco A, et al. Psychosocial implications of avatar use in supporting therapy for depression. Stud Health Technol Inform 2012;181:329-333. [Medline: 22954882]
20. Bickmore T, Gruber A. Relational agents in clinical psychiatry. Harv Rev Psychiatry 2010;18(2):119-130. [doi: 10.3109/10673221003707538] [Medline: 20235777]
21. Scholten MR, Kelders SM, Van Gemert-Pijnen JE. Self-guided Web-based interventions: scoping review on user needs and the potential of embodied conversational agents to address them. J Med Internet Res 2017 Nov 16;19(11):e383 [FREE Full text] [doi: 10.2196/jmir.7351] [Medline: 29146567]
22. Oh J, Jang S, Kim H, Kim JJ. Efficacy of mobile app-based interactive cognitive behavioral therapy using a chatbot for panic disorder. Int J Med Inform 2020 Aug;140:104171. [doi: 10.1016/j.ijmedinf.2020.104171] [Medline: 32446158]
23. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 01;25(9):1248-1258 [FREE Full text] [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
24. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 2019 Jul;64(7):456-464 [FREE Full text] [doi: 10.1177/0706743719828977] [Medline: 30897957]
25. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. JMIR Ment Health 2019 Oct 18;6(10):e14166 [FREE Full text] [doi: 10.2196/14166] [Medline: 31628789]
26. Bendig E, Erb B, Schulze-Thuesing L, Baumeister H. The next generation: chatbots in clinical psychology and psychotherapy to foster mental health – a scoping review. Verhaltenstherapie 2019 Aug 20:1-13. [doi: 10.1159/000501812]
27. Araujo T. Living up to the chatbot hype: the influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. Comput Human Behav 2018 Aug;85:183-189. [doi: 10.1016/j.chb.2018.03.051]
28. Go E, Sundar SS. Humanizing chatbots: the effects of visual, identity and conversational cues on humanness perceptions. Comput Human Behav 2019 Aug;97:304-316. [doi: 10.1016/j.chb.2019.01.020]
29. Nass C, Moon Y. Machines and mindlessness: social responses to computers. J Social Isssues 2000 Jan;56(1):81-103. [doi: 10.1111/0022-4537.00153]
30. Feine J, Gnewuch U, Morana S, Maedche A. A taxonomy of social cues for conversational agents. Int J Hum Comput Stud 2019 Dec;132:138-161. [doi: 10.1016/j.ijhcs.2019.07.009]
31. Short J, Williams E, Christie B. The Social Psychology of Telecommunications. London, UK: Wiley; 1976.

XSL•FO
RenderX

32. Tourangeau R, Smith TW. Asking sensitive questions: the impact of data collection mode, question format, and question context. Public Opin Q 1996;60(2):275-304. [doi: 10.1086/297751]
33. Tourangeau R, Yan T. Sensitive questions in surveys. Psychol Bull 2007 Sep;133(5):859-883. [doi: 10.1037/0033-2909.133.5.859] [Medline: 17723033]
34. Gnambs T, Kaspar K. Disclosure of sensitive behaviors across self-administered survey modes: a meta-analysis. Behav Res Methods 2015 Dec;47(4):1237-1259. [doi: 10.3758/s13428-014-0533-4] [Medline: 25410404]
35. Nederhof AJ. Methods of coping with social desirability bias: a review. Eur J Soc Psychol 1985 Jul;15(3):263-280. [doi: 10.1002/ejsp.2420150303]
36. Nguyen DP, Klein B, Meyer D, Austin DW, Abbott JA. The diagnostic validity and reliability of an internet-based clinical assessment program for mental disorders. J Med Internet Res 2015 Sep 21;17(9):e218 [FREE Full text] [doi: 10.2196/jmir.4195] [Medline: 26392066]
37. Ramo DE, Liu H, Prochaska JJ. Reliability and validity of young adults' anonymous online reports of marijuana use and thoughts about use. Psychol Addict Behav 2012 Dec;26(4):801-811 [FREE Full text] [doi: 10.1037/a0026201] [Medline: 22082344]
38. Jin H, Wu S. Text messaging as a screening tool for depression and related conditions in underserved, predominantly minority safety net primary care patients: validity study. J Med Internet Res 2020 Mar 26;22(3):e17282 [FREE Full text] [doi: 10.2196/17282] [Medline: 32213473]
39. Meuldijk D, Giltay EJ, Carlier IV, van Vliet IM, van Hemert AM, Zitman FG. A validation study of the Web Screening Questionnaire (WSQ) compared with the Mini-International Neuropsychiatric Interview-Plus (MINI-Plus). JMIR Ment Health 2017 Aug 29;4(3):e35 [FREE Full text] [doi: 10.2196/mental.5453] [Medline: 28851674]
40. Vallejo MA, Jordán CM, Díaz MI, Comeche MI, Ortega J. Psychological assessment via the Internet: a reliability and validity study of online (vs paper-and-pencil) versions of the General Health Questionnaire-28 (GHQ-28) and the Symptoms Check-List-90-Revised (SCL-90-R). J Med Internet Res 2007 Jan 31;9(1):e2 [FREE Full text] [doi: 10.2196/jmir.9.1.e2] [Medline: 17478411]
41. Tourangeau R, Rips LJ, Rasinski K. The Psychology of Survey Response. Cambridge, UK: Cambridge University Press; 2000.
42. Schuetzler RM, Giboney JS, Grimes GM, Nunamaker JF. The influence of conversational agent embodiment and conversational relevance on socially desirable responding. Decis Support Syst 2018 Oct;114:94-102. [doi: 10.1016/j.dss.2018.08.011]
43. Phillips DL, Clancy KJ. Some effects of "social desirability" in survey studies. Am J Sociol 1972 Mar;77(5):921-940. [doi: 10.1086/225231]
44. Fisher RJ. Social desirability bias and the validity of indirect questioning. J Consum Res 1993 Sep;20(2):303-315. [doi: 10.1086/209351]
45. Nass C, Steuer J, Tauber ER. Computers are social actors. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 1994 Presented at: CHI '94; April 24-28, 1994; Boston, MA, USA p. 72-78. [doi: 10.1145/191666.191703]
46. Kocielnik R, Agapie E, Argyle A, Hsieh DT, Yadav K, Taira B, et al. HarborBot: a chatbot for social needs screening. AMIA Annu Symp Proc 2019 Mar 4;2019:552-561 [FREE Full text] [Medline: 32308849]
47. Xiao Z, Zhou MX, Liao QV, Mark G, Chi C, Chen W, et al. Tell me about yourself: using an AI-powered chatbot to conduct conversational surveys with open-ended questions. ACM Trans Comput Hum Interact 2020 Jun 30;27(3):1-37. [doi: 10.1145/3381804]
48. LimeSurvey Development Team. LimeSurvey - The free and open source survey software tool!. LimeSurvey Development Team. 2012. URL: http://www.limesurvey.org [accessed 2019-08-07]
49. Bot Framework SDK. Microsoft. 2020. URL: https://github.com/microsoft/botframework-sdk [accessed 2019-08-07]
50. de Visser EJ, Monfort SS, McKendrick R, Smith MA, McKnight PE, Krueger F, et al. Almost human: anthropomorphism increases trust resilience in cognitive agents. J Exp Psychol Appl 2016 Sep;22(3):331-349. [doi: 10.1037/xap0000092] [Medline: 27505048]
51. Bickmore TW, Picard RW. Establishing and maintaining long-term human-computer relationships. ACM Trans Comput Hum Interact 2005 Jun;12(2):293-327. [doi: 10.1145/1067860.1067867]
52. Gnewuch U, Morana S, Adam MT, Maedche A. 'The Chatbot is…' - the role of typing indicators in human-chatbot interaction. In: Proceedings of the SIGHCI - The 17th Annual Pre-ICIS Workshop on HCI Research in MIS. 2018 Presented at: HCI '18; December 13, 2018; San Francisco, CA, USA URL: https://aisel.aisnet.org/sighci2018/14/
53. Mori M. The uncanny valley. Energy 1970;7(4):33-35.
54. Musch J, Brockhaus R, Bröder A. Ein Inventar zur Erfassung von zwei Faktoren sozialer Erwünschtheit. Diagnostica 2002 Jul;48(3):121-129. [doi: 10.1026//0012-1924.48.3.121]
55. Paulhus DL. The Balanced Inventory of Desirable Responding. Toronto, Canada: Multi-Health Systems; 1998.
56. Kessler RC, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SL, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002 Aug;32(6):959-976. [doi: 10.1017/s0033291702006074] [Medline: 12214795]

XSL•FO
RenderX

57. Derogatis LR. BSI 18, Brief Symptom Inventory 18 : Administration, Scoring and Procedures Manual. Minneapolis, MN, USA: NCS Pearson; 2001.
58. Spitzer C, Hammer S, Löwe B, Grabe HJ, Barnow S, Rose M, et al. Die Kurzform des Brief Symptom Inventory (BSI -18): erste Befunde zu den psychometrischen Kennwerten der deutschen Version. Fortschr Neurol Psychiatr 2011 Sep;79(9):517-523. [doi: 10.1055/s-0031-1281602] [Medline: 21870312]
59. Saunders JB, Aasland OG, Babor TF, de la Fuente JR, Grant M. Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on early detection of persons with harmful alcohol consumption--II. Addiction 1993 Jun;88(6):791-804. [doi: 10.1111/j.1360-0443.1993.tb02093.x] [Medline: 8329970]
60. Gual A, Segura L, Contel M, Heather N, Colom J. Audit-3 and audit-4: effectiveness of two short forms of the alcohol use disorders identification test. Alcohol Alcohol 2002;37(6):591-596. [doi: 10.1093/alcalc/37.6.591] [Medline: 12414553]
61. Bradburn NM. Improving Interview Method and Questionnaire Design: Response Effects to Threatening Questions in Survey Research. San Francisco, CA, USA: Jossey-Bass; 1979.
62. Gefen D, Straub DW. Gender differences in the perception and use of e-mail: an extension to the technology acceptance model. MIS Q 1997 Dec;21(4):389-400. [doi: 10.2307/249720]
63. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191. [doi: 10.3758/bf03193146] [Medline: 17695343]
64. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull 1959 Mar;56(2):81-105. [doi: 10.1037/h0046016]
65. Slade T, Grove R, Burgess P. Kessler Psychological Distress Scale: normative data from the 2007 Australian National Survey of Mental Health and Wellbeing. Aust N Z J Psychiatry 2011 Apr;45(4):308-316. [doi: 10.3109/00048674.2010.543653] [Medline: 21332432]
66. Gerbert B, Bronstone A, Pantilat S, McPhee S, Allerton M, Moe J. When asked, patients tell: disclosure of sensitive health-risk behaviors. Med Care 1999 Jan;37(1):104-111. [doi: 10.1097/00005650-199901000-00014] [Medline: 10413398]
67. Luger E, Sellen A. "Like having a really bad PA": the Gulf between user expectation and experience of conversational agents. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. 2016 Presented at: CHI '16; May 7-12, 2016; San Jose, CA, USA p. 5286-5297. [doi: 10.1145/2858036.2858288]
68. Gratch J, Lucas G, King A, Morency LP. It's only a computer: the impact of human-agent interaction in clinical interviews. In: Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems. 2014 Presented at: AAMAS '14; May 5-9, 2014; Paris, France p. 85-92.
69. Kim S, Lee J, Gweon G. Comparing data from chatbot and web surveys: effects of platform and conversational style on survey response quality. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19; May 4-9, 2019; Glasgow, UK p. 86. [doi: 10.1145/3290605.3300316]
70. Lee YC, Yamashita N, Huang Y. Designing a chatbot as a mediator for promoting deep self-disclosure to a real mental health professional. Proc ACM Hum Comput Interact 2020 May 28;4(CSCW1):1-27. [doi: 10.1145/3392836]
71. Kang SH, Fort Morie J. Users' socially desirable responding with computer interviewers. In: CHI '13 Extended Abstracts on Human Factors in Computing Systems. 2013 Presented at: CHI EA '13; April 27-May 2, 2013; Paris, France p. 229-234. [doi: 10.1145/2468356.2468398]
72. Jungmann SM, Klan T, Kuhn S, Jungmann F. Accuracy of a chatbot (Ada) in the diagnosis of mental disorders: comparative case study with lay and expert users. JMIR Form Res 2019 Oct 29;3(4):e13863 [FREE Full text] [doi: 10.2196/13863] [Medline: 31663858]
73. Sah YJ, Peng W. Effects of visual and linguistic anthropomorphic cues on social perception, self-awareness, and information disclosure in a health website. Comput Human Behav 2015 Apr;45(C):392-401. [doi: 10.1016/j.chb.2014.12.055]
74. Lee YC, Yamashita N, Huang Y, Fu W. "I hear you, I feel you": encouraging deep self-disclosure through a chatbot. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI '20; April 25-30, 2020; Honolulu, HI, USA p. 1-12. [doi: 10.1145/3313831.3376175]
75. Adam M, Klumpe J. Onboarding with a chat – the effects of message interactivity and platform self-disclosure on user disclosure propensity. In: Proceedings of the 27th European Conference on Information Systems. 2019 Presented at: ECIS '19; June 8-14, 2019; Stockholm, Sweden URL: https://aisel.aisnet.org/ecis2019_rp/68/
76. Yokotani K, Takagi G, Wakashima K. Advantages of virtual agents over clinical psychologists during comprehensive mental health interviews using a mixed methods design. Comput Human Behav 2018 Aug;85:135-145. [doi: 10.1016/j.chb.2018.03.045]
77. Bérubé C, Schachner T, Keller R, Fleisch E, V Wangenheim F, Barata F, et al. Voice-based conversational agents for the prevention and management of chronic and mental health conditions: systematic literature review. J Med Internet Res 2021 Mar 29;23(3):e25933 [FREE Full text] [doi: 10.2196/25933] [Medline: 33658174]
78. Bickmore TW, Trinh H, Olafsson S, O'Leary TK, Asadi R, Rickles NM, et al. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. J Med Internet Res 2018 Sep 04;20(9):e11510 [FREE Full text] [doi: 10.2196/11510] [Medline: 30181110]

XSL•FO
RenderX

79. Walther JB. Nonverbal dynamics in computer-mediated communication or :(and the net :('s with you, :) and you :) alone. In: Manusov V, Patterson ML, editors. The SAGE Handbook of Nonverbal Communication. Thousand Oaks, CA, USA: Sage Publications; 2006.
80. Daft RL, Lengel RH. Information richness: a new approach to managerial behavior and organizational design. Res Organ Behav 1984;6:191-233. [doi: 10.21236/ada128980]
81. Colasante E, Benedetti E, Fortunato L, Scalese M, Potente R, Cutilli A, et al. Paper-and-pencil versus computerized administration mode: comparison of data quality and risk behavior prevalence estimates in the European school Survey Project on Alcohol and other Drugs (ESPAD). PLoS One 2019 Nov 20;14(11):e0225140 [FREE Full text] [doi: 10.1371/journal.pone.0225140] [Medline: 31747446]
82. Zand A, Sharma A, Stokes Z, Reynolds C, Montilla A, Sauk J, et al. An exploration into the use of a chatbot for patients with inflammatory bowel diseases: retrospective cohort study. J Med Internet Res 2020 May 26;22(5):e15589 [FREE Full text] [doi: 10.2196/15589] [Medline: 32452808]
83. Ho A, Hancock J, Miner AS. Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. J Commun 2018 Aug;68(4):712-733 [FREE Full text] [doi: 10.1093/joc/jqy026] [Medline: 30100620]
84. Golinelli D, Boetto E, Carullo G, Nuzzolese AG, Landini MP, Fantini MP. Adoption of digital technologies in health care during the COVID-19 pandemic: systematic review of early scientific literature. J Med Internet Res 2020 Nov 06;22(11):e22280 [FREE Full text] [doi: 10.2196/22280] [Medline: 33079693]
85. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 2018 Nov 23;6(11):e12106 [FREE Full text] [doi: 10.2196/12106] [Medline: 30470676]
86. Lee KT, Noh MJ, Koo DM. Lonely people are no longer lonely on social networking sites: the mediating role of self-disclosure and social support. Cyberpsychol Behav Soc Netw 2013 Jun;16(6):413-418. [doi: 10.1089/cyber.2012.0553] [Medline: 23621716]
87. Morris RR, Kouddous K, Kshirsagar R, Schueller SM. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. J Med Internet Res 2018 Jun 26;20(6):e10148 [FREE Full text] [doi: 10.2196/10148] [Medline: 29945856]
88. Shum HY, He XD, Li D. From Eliza to XiaoIce: challenges and opportunities with social chatbots. Front Inf Technol Electron Eng 2018 Jan 8;19(1):10-26. [doi: 10.1631/fitee.1700826]
89. Narain J, Quach T, Davey M, Park HW, Breazeal C, Picard R. Promoting wellbeing with sunny, a chatbot that facilitates positive messages within social groups. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI EA '20; April 25-30, 2020; Honolulu, HI, USA p. 1-8. [doi: 10.1145/3334480.3383062]
90. de Gennaro M, Krumhuber EG, Lucas G. Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. Front Psychol 2019 Jan 23;10:3061 [FREE Full text] [doi: 10.3389/fpsyg.2019.03061] [Medline: 32038415]
91. Ta V, Griffith C, Boatfield C, Wang X, Civitello M, Bader H, et al. User experiences of social support from companion chatbots in everyday contexts: thematic analysis. J Med Internet Res 2020 Mar 06;22(3):e16235 [FREE Full text] [doi: 10.2196/16235] [Medline: 32141837]

Abbreviations

AUDIT: Alcohol Use Disorders Identification Test
BIDR: Balanced Inventory of Desirable Responding
BSI-18: Brief Symptom Inventory-18
ICC: intraclass correlation coefficient
K10: Kessler Psychological Distress Scale
rmANCOVA: repeated-measures analysis of covariance
rmANOVA: repeated-measures ANOVA
SDR: socially desirable responding

Edited by L Buis; submitted 19.02.21; peer-reviewed by G Hsieh, J Martínez-Miranda, H Tanaka, T Bickmore, H Mehdizadeh; comments to author 16.05.21; revised version received 09.10.21; accepted 09.05.22; published 31.10.22.

Please cite as:
Schick A, Feine J, Morana S, Maedche A, Reininghaus U
Validity of Chatbot Use for Mental Health Assessment: Experimental Study
JMIR Mhealth Uhealth 2022;10(10):e28082
URL: https://mhealth.jmir.org/2022/10/e28082
doi:10.2196/28082
PMID:36315228

XSL•FO
RenderX

©Anita Schick, Jasper Feine, Stefan Morana, Alexander Maedche, Ulrich Reininghaus. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 31.10.2022. 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.

XSL•FO
RenderX

Original Paper

Designing, Developing, Evaluating, and Implementing a Smartphone-Delivered, Rule-Based Conversational Agent (DISCOVER): Development of a Conceptual Framework

Dhakshenya Ardhithy Dhinagaran[1*], BSc (Hons); Laura Martinengo[1*], MD, PhD; Moon-Ho Ringo Ho[2], PhD; Shafiq Joty[3], MSc, PhD; Tobias Kowatsch[4,5,6,7], PhD; Rifat Atun[8,9,10], MBBS, MBA, DIC; Lorainne Tudor Car[1,11], MD, PhD

[1]Lee Kong Chian School of Medicine, Nanyang Technological University Singapore, Singapore, Singapore

[2]School of Social Sciences, Nanyang Technological University Singapore, Singapore, Singapore

[3]School of Computer Sciences and Engineering, Nanyang Technological University Singapore, Singapore, Singapore

[4]Institute for Implementation Science in Health Care, University of Zurich, Zurich, Switzerland

[5]School of Medicine, University of St Gallen, St Gallen, Switzerland

[6]Centre for Digital Health Interventions, Department of Management, Technology, and Economics, ETH Zurich, Zurich, Switzerland

[7]Future Health Technologies Programme, Campus for Research Excellence and Technological Enterprise, Singapore-ETH Centre, Singapore, Singapore

[8]Department of Global Health & Population, Department of Health Policy & Management, Harvard TH Chan School of Public Health, Harvard University, Cambridge, MA, United States

[9]Department of Global Health and Social Medicine, Harvard Medical School, Harvard University, Cambridge, MA, United States

[10]Health Systems Innovation Lab, Harvard TH Chan School of Public Health, Harvard University, Cambridge, MA, United States

[11]Department of Primary Care and Public Health, School of Public Health, Imperial College London, London, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Lorainne Tudor Car, MD, PhD
Lee Kong Chian School of Medicine
Nanyang Technological University Singapore
11 Mandalay Rd69047142
Clinical Sciences Building Level 18
Singapore, 308232
Singapore
Phone: 65 69047142
Email: lorainne.tudor.car@ntu.edu.sg

Abstract

Background: Conversational agents (CAs), also known as chatbots, are computer programs that simulate human conversations by using predetermined rule-based responses or artificial intelligence algorithms. They are increasingly used in health care, particularly via smartphones. There is, at present, no conceptual framework guiding the development of smartphone-based, rule-based CAs in health care. To fill this gap, we propose structured and tailored guidance for their design, development, evaluation, and implementation.

Objective: The aim of this study was to develop a conceptual framework for the design, evaluation, and implementation of smartphone-delivered, rule-based, goal-oriented, and text-based CAs for health care.

Methods: We followed the approach by Jabareen, which was based on the grounded theory method, to develop this conceptual framework. We performed 2 literature reviews focusing on health care CAs and conceptual frameworks for the development of mobile health interventions. We identified, named, categorized, integrated, and synthesized the information retrieved from the literature reviews to develop the conceptual framework. We then applied this framework by developing a CA and testing it in a feasibility study.

Results: The Designing, Developing, Evaluating, and Implementing a Smartphone-Delivered, Rule-Based Conversational Agent (DISCOVER) conceptual framework includes 8 iterative steps grouped into 3 stages, as follows: design, comprising defining the goal, creating an identity, assembling the team, and selecting the delivery interface; development, including developing the content and building the conversation flow; and the evaluation and implementation of the CA. They were complemented by 2 cross-cutting

XSL•FO
RenderX

considerations—user-centered design and privacy and security—that were relevant at all stages. This conceptual framework was successfully applied in the development of a CA to support lifestyle changes and prevent type 2 diabetes.

Conclusions: Drawing on published evidence, the DISCOVER conceptual framework provides a step-by-step guide for developing rule-based, smartphone-delivered CAs. Further evaluation of this framework in diverse health care areas and settings and for a variety of users is needed to demonstrate its validity. Future research should aim to explore the use of CAs to deliver health care interventions, including behavior change and potential privacy and safety concerns.

(JMIR Mhealth Uhealth 2022;10(10):e38740) doi:10.2196/38740

KEYWORDS

conceptual framework; conversational agent; chatbot; mobile health; mHealth; digital health; mobile phone

Introduction

Background

Conversational agents (CAs) are computer programs that use text, speech, and other input modalities to enable communication with users [1]. They can be accessed through a variety of ways, such as social media platforms (eg, Facebook Messenger), websites, and smartphone apps, or deployed using stand-alone digital devices (eg, Alexa, Google Assistant, and Siri). The interactive nature of CAs makes them acceptable to a diverse group of users [2-4] and a preferred tool in a number of disciplines, including customer service, retail, and e-commerce [5-7].

In health care, CAs are increasingly used to assist in various tasks, such as patient education, self-management of chronic conditions, and routine task automation (eg, appointment booking), and support health professionals' decision-making for diagnosis and triage [3,8-10]. More recently, CAs have seen large-scale implementation with the introduction of Babylon's artificial intelligence (AI)–based symptom checker CA to the UK National Health Service and to Rwanda's National Health Insurance Scheme [11]. CAs have the potential to support health care delivery, improve access to health care services, and automate tasks [12], and they may also reduce health professionals' workload [13].

CAs vary in complexity and capability. There are 3 design dimensions used to classify CAs: purpose, communication channels, and response generation architecture [6]. According to *purpose*, CAs can be classified into *task- or goal-oriented CAs*, which respond to a limited number of tasks within a prespecified domain, or *non–task- or non–goal-oriented CAs*, which are potentially able to respond to an unrestricted variety of user requests [6]. *Communication channels* can commonly be divided into 2 main types: *text-based* or *voice-based* CAs. *Response generation architecture* can be broadly classified into 3 groups: *rule-based* and *retrieval-based CAs*, which produce a response by selecting it from a pool of predetermined responses either following simple rules to match phrases or identifying specific keywords in the text [6,14,15], and *generative-based CAs*, which use AI algorithms to develop a contextual response informed by the system's previous and ongoing learning [6,14-16]. Although all 3 groups may involve the use of AI algorithms [6], rule-based CAs allow developers greater control over the conversation content and flow, which is a useful feature when developing CAs for health care. By contrast, AI algorithms, particularly neural networks, may develop decisions that are not explainable or understood by the end user, a phenomenon referred to as the *black box* [17]. In health care settings, the *black box* effect may lead to biased or erroneous decision-making and patient harm [18], which may limit the use of AI. A new field of explainable AI is currently emerging that aims to provide justification for algorithm predictions and increase system transparency, although the validity of results for individual patients should be carefully considered [19].

CAs can be deployed using a variety of digital devices, including smartphones. The widespread availability of smartphones in high-income countries and increasingly in low- and middle-income countries [20] makes them an ideal interface to deliver CA interventions. Smartphones offer users the possibility of continuous and dynamic monitoring of health conditions in a private space and at the time of their convenience [21] not only of subjective, self-reported data but also of objective, sensor-based data. Furthermore, smartphones allow for the delivery of interventions according to user needs [22]. CA interventions are complex and often require lengthy, costly design and development processes led by multidisciplinary teams of health care professionals, computer scientists, and app developers, which may limit the number of teams able to engage in CA development, particularly in low- and middle-income countries. However, mobile health (mHealth) interventions, particularly SMS text messages delivered using mobile phones, are effective in delivering health care interventions, especially in low-resource settings [23,24].

Several frameworks for the design and development of mHealth interventions currently exist, offering guidance at every step of the cycle, from the conceptualization of user needs [25,26] to the development of the digital health intervention [25-27]. These frameworks focus on generic, app-based interventions without a conversational interface. However, Zhang et al [28] described a framework for the development of AI-based CAs to deliver behavior change interventions that may require significant deployment of resources, including a large, multidisciplinary team, and close supervision of the AI algorithms to prevent unintended and potentially harmful effects on the users. However, to date, no conceptual framework for the design, development, and evaluation of rule-based CAs has been published despite a growing interest in the use of CAs in health care settings.

XSL•FO
RenderX

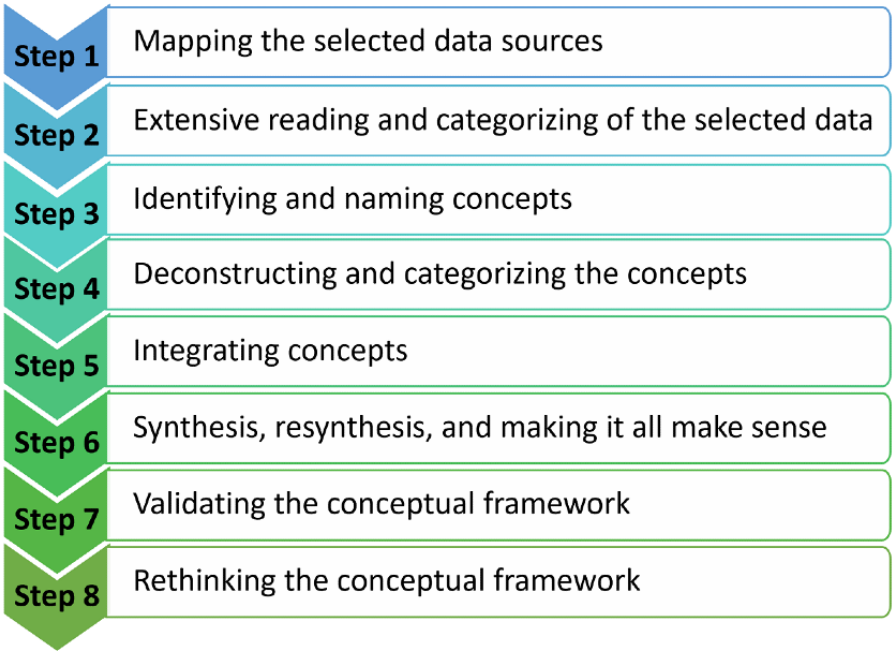
Objectives

CAs constitute a specific type of digital intervention characterized by the use of a conversational interface, often led by an agent with a distinct personality as evidenced by its tone of speech, method of interaction, and visual representation, which is often associated with higher levels of engagement with the user. These features and the ubiquity of smartphones support the need for a framework that is accessible to large as well as smaller research teams with limited resources to guide CA development, including the distinct design and development challenges of CAs such as the creation of dialogs and the look and personality of the agent, grounded in current best evidence. Therefore, this research aimed to develop a conceptual framework for the design, development, evaluation, and implementation of smartphone-delivered, rule-based, goal-oriented, and text-based CAs for health care.

Methods

We developed the Designing, Developing, Evaluating, and Implementing a Smartphone-Delivered, Rule-Based Conversational Agent (DISCOVER) conceptual framework according to the methodology described by Jabareen [29], consisting of the iterative, qualitative analysis of multidisciplinary data based on the grounded theory method. It comprises 8 interlinked steps aimed at integrating and analyzing the data and developing and validating the conceptual framework [29] (Figure 1).

Figure 1. The 8 phases of the methodology by Jabareen [29] for conceptual framework development.



Step 1

We conducted 2 literature reviews. The first review aimed to summarize the current literature on conceptual frameworks for the design, development, and evaluation of mHealth interventions, and the second review focused on smartphone-delivered, rule-based CAs. A description of these literature reviews can be found in Multimedia Appendix 1 [5,30-62] and Multimedia Appendix 2 [5,30-62]. Multimedia Appendix 3 presents the search strategy used to retrieve the studies for the review of CAs.

Step 2 and Step 3

The screening of retrieved citations was performed in 2 stages, independently and in parallel, by DD and LM. The same 2 reviewers extracted data from all the included studies independently and in parallel. At all stages of screening and data extraction, the results were compared, and discrepancies were resolved by consensus between the reviewers.

Step 4

The data analysis followed qualitative meta-synthesis to systematically summarize the findings across all the included studies. This step involved grouping the concepts extracted from both literature reviews into overarching domains.

Step 5 and Step 6

The next 2 steps involved linking the overarching domains and developing the first iteration of the conceptual framework.

Step 7 and Step 8

The conceptual framework was further amended based on discussions among the research team members and feedback from colleagues collected in a seminar. We subsequently applied the conceptual framework to develop a rule-based, text-based, smartphone-delivered CA prototype (*Precilla*) designed to support healthy lifestyle changes and educate participants about diabetes. The development, feasibility, and acceptability of Precilla have been reported elsewhere [63,64].

The feedback received from team members and colleagues and the lessons learned during the application study led to the refinement of concepts and domain labels, definitions, order, and grouping that were derived in the current version of DISCOVER presented in this paper.

XSL•FO
RenderX

Ethical Considerations

This study was approved by the Nanyang Technological University Institutional Review Board (IRB-2018-11-032).

Results

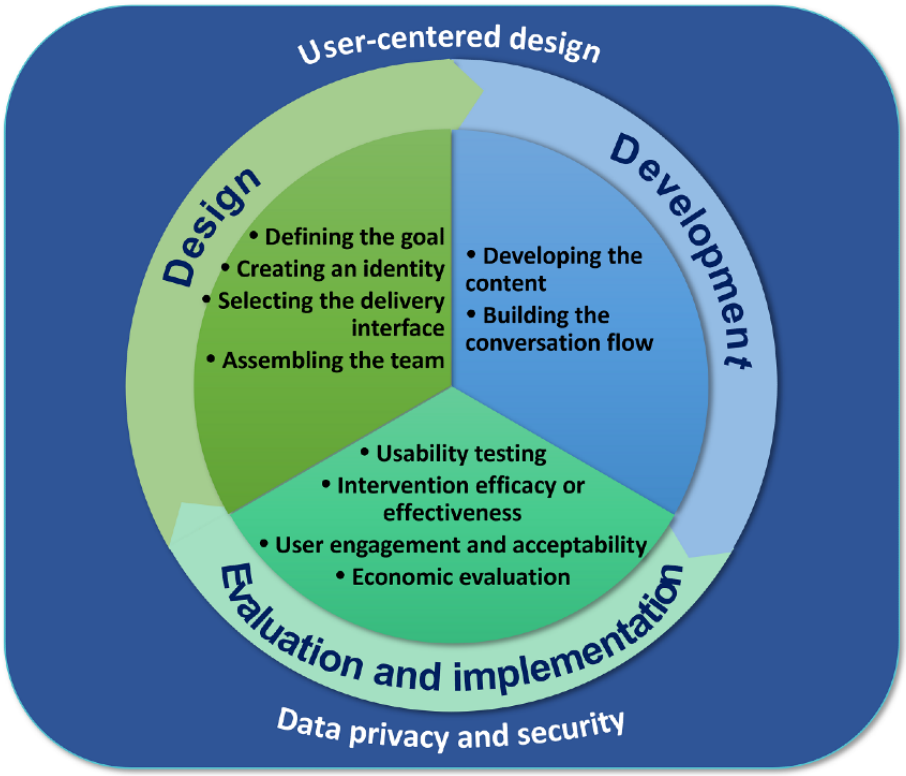
A Framework for Guiding the Design, Development, Evaluation, and Implementation of Smartphone-Delivered, Rule-Based CAs in Health Care: Overview

The conceptual framework development was informed by the 2 literature reviews and iterative consultations within the research team. Further refinements were also informed by the development of our CA prototype (*Precilla*) [63,64] as well as by presentations at clinical seminars and conferences. Multimedia Appendix 4 outlines the methodology applied in the development of the DISCOVER framework according to each step described by Jabareen [29]. Multimedia Appendix 5 [63,64] presents the steps to develop CA Precilla mapped to the steps of the current version of the conceptual framework.

The 2 literature searches retrieved a total of 55 studies, of which 41 (75%) described conceptual frameworks for the design, development, and evaluation of mHealth interventions and 14 (25%) were clinical trials evaluating smartphone- and rule-based CAs. The findings from these reviews are presented in Multimedia Appendices 1 and 2. The "Characteristics of included studies" tables are presented in Multimedia Appendix 6 [47-58], Multimedia Appendix 7 [5,32,65-67], and Multimedia Appendix 8 [3,30,31,33,34,68-80].

The initial framework contained 8 steps. They were subsequently condensed into 5 steps augmented by 2 overarching themes relevant to all phases of the development process. Further refinements led to the framework presented in this paper consisting of an iterative process of design, development, evaluation, and implementation steps, each comprising several components, as presented in Figure 2 and described in the following sections.

Figure 2. The DISCOVER conceptual framework for the design, development, and evaluation of rule-based, smartphone-based conversational agents in health care.



Step 1: Design

The first stage comprised 4 interlinked steps encapsulating the initial conceptual work of identifying the health care focus of the CA, target users, multidisciplinary team members, and the CA delivery interface.

Defining the Goal

Overview

A clearly defined goal is the first step in the design process and the foundation that will guide the development and evaluation of the CA. This step consists of 3 interlinked areas of evaluation—completing a thorough needs assessment, defining the aim, and characterizing the end user and objectives—which, in turn, determine the parameters to be tested and reported. The CA goal was described in 64% (35/55) of the papers in our reviews [3,5,25-28,65,68-76,81-99].

Needs Assessment

The design process should commence with an in-depth needs assessment to understand existing gaps that may be filled by the CA. These may be informed by a literature review [83,90,91,96,100] to assess potential research areas and the needs and challenges of the target population, including not

XSL•FO
RenderX

only patients but also caregivers, health care providers (HCPs), and other experts [25,26,87,89,95,98]. Researchers should also involve end users in this initial phase by using surveys and a variety of qualitative methods [70,87] such as in-depth interviews and focus group discussions to gather their views.

The Aim

Aligned with the needs assessment, the design team should formulate clear, attainable, and relevant objectives to drive the CA design and development process. It is important to consider the CA temporal profile, which characterizes 4 types of CAs according to the type and frequency of CA-user dialogs [101]. The CA temporal profile will also determine the type of objectives included, broadly classified as short term or long term [101]. A short-term objective refers to an outcome to be completed as soon as the interaction with the CA ends, such as medication reminders [30]. A long-term goal would involve several CA-user interactions being completed over a period, as in mental health interventions to promote mental well-being in the general population [3] or young people with cancer [70]. Complex CA interventions may include short- and long-term goals, such as CA Vik [30] providing medication reminders (short-term goal) and health education (long-term goal) to patients with breast cancer. Furthermore, Kowatsch et al [73] used prompts and reminder SMS text messages to enhance children's discipline and routine, which are essential for the self-management of asthma.

Determining the End User

The next important design consideration is to determine the target population. An initial assessment should establish whether the CA will be offered to healthy users or individuals with a specific medical condition, caregivers, or HCPs. It is important to generate a detailed and accurate portrayal of the target user, including gender, age group, cultural beliefs and socioeconomic concerns, digital and health literacy, access to digital devices, and smartphone penetration rate. If the intervention is educational, a knowledge test should be implemented [73]. The acceptability of CAs by the target population and the perceived risk of using a CA for health care matters should be evaluated, particularly for severe or highly stigmatizing conditions [102] such as mental health disorders [103,104].

Creating the CA Identity

This step involves determining the CA's name, appearance, tone of communication, language, and other characteristics that define its identity. This step was discussed in 25% (14/55) of the papers in our reviews [5,31,32,66,69,70,72-78].

CA Personality

User interaction with CAs appears to be enhanced when the CA displays a well-defined, positive, and empathic personality [105,106]. In general, giving a name and profile picture to the CA may enhance its social presence and user acceptance [107], although its effect appears to be small [106]. In health care settings, using a human-like avatar rendering realistic features, including medical attire, may increase user satisfaction [105], although avatars displaying highly realistic features may upset users and decrease engagement, an experience referred to as the "uncanny valley" [108].

Studies have consistently shown that CAs displaying empathy, relational behavior, and self-disclosure enhance the user experience [78,105] and increase the working alliance [109]. Conversely, users would notice if the CA did not convey empathy [69].

Acceptability may be further enhanced if the CA design acknowledges the specific cultural or demographic traits of the target population [73] or offers options to personalize the interface (eg, offering a male and female persona) [31,73]. Alternatively, CAs may explicitly disclose their identity [70] to reduce user expectations about their capabilities. Finally, CA personality should align with its intended function. For example, health care CAs often display one of two personality types: a more approachable, empathic coach-like personality, particularly if delivering behavior change interventions [77] supporting self-management of chronic disorders [73,78] and mental health conditions [3], or a health care professional persona to emphasize the legitimacy of the CA and its content [1].

Tone and Language

The language recommended for text-based interventions should be encouraging, positive, friendly, polite, and light-hearted and may include light humor while at the same time being formal [110]. To maintain the flow of the conversation, it may be advisable to use visual cues such as successive moving dots signaling that the CA is "typing" the next message.

The text should be written in clear, short sentences using simple language and avoid scientific jargon. The National Institutes of Health recommends that patient education materials be written at or below the sixth-grade reading level (ages of 11 and 12 years) to reach a diverse range of individuals with varying levels of literacy [111]. The readability of the text can be assessed using a scale such as the Flesch-Kincaid grade level to determine its suitability [112]. Furthermore, the CA should use the target population's native language in its communications [75] and, if needed, the conversational content may be translated to one or more languages, particularly if the CA will be deployed in multiethnic, multilanguage societies.

With regard to the tone of the conversation, despite the text-based nature of the CA, it may be advisable to simulate more casual, verbal speech while avoiding the use of "textese" [113], a form of abbreviated written or typed language characterized by unconventional spelling and grammar (eg, "tonite" instead of "tonight") and abbreviations and contractions (eg, "pls" instead of "please" or "wanna" for "want to") [114]. Furthermore, words written in full capital letters should be avoided as they equate to shouting [110].

Emojis may be used to articulate emotions or other expressions more efficiently than text [70]. However, emojis are vulnerable to varied interpretations across cultures and contexts and should be used mindfully. Fadhil et al [115] noted a context-specific nature of emojis whereby they increased efficacy in a mental health intervention but did not help in promoting physical well-being.

CAs designed to address sensitive topics such as HIV and AIDS, sexually transmitted infections, or mental health disorders may emphasize the confidential nature of the messages or include

XSL•FO
RenderX

code words to protect users' privacy. This is particularly relevant in low- and middle-income settings, where family members may share a single smartphone [116].

Selecting the Delivery Interface

Human Involvement

Conditional to the CA's aim, the design and development team may consider a "hybrid" intervention where the interaction with the CA would be complemented by regular interactions with HCPs offering timely feedback on a self-management technique or regular support and motivation [33,72,73]. Alternatively, as presented in the study by Stasinaki et al [79], the CA may be fitted with multiple channels, where the user can converse with the CA in one channel and directly with an HCP in another.

Peer support is recognized to play an important role in promoting adherence to self-management interventions [117] and a further point of human involvement to be considered. The CA intervention may include an additional communication channel for users to interact, share experiences, and receive peer support. For example, Wang et al [75] developed a WeChat intervention to support smoking cessation where the CA not only responded to individual users but also acted as a group moderator.

Delivery Channel

CAs may be delivered through a variety of channels, such as stand-alone apps [3,73] and existing messaging platforms [68,71,75] such as Facebook Messenger, Telegram, WeChat, and WhatsApp, or embedded in a website [69]. Each channel possesses its own set of complexities, and the decision regarding the delivery channel should be based on the target population needs and the expertise of the CA development team [118]. If the research team does not include app developers or computer scientists, the CA may be embedded in a messaging platform or may be developed using a CA development platform that offers templates or other design solutions for individuals with no previous programming knowledge [5,118], such as Chatfuel, ManyChat, and others. CAs are generally web-based, and some of these platforms are free of charge. Alternatively, if the team expertise or project budget allows, the CA may be delivered through a stand-alone app. This approach offers design flexibility, such as a variety of data collection sources including smartphone sensors, health programming interfaces, connected medical devices, and patient self-reported data [119]. The combination of subjective patient reports with objective, real-time data may reduce users' responsibility to update their progress and at the same time receive relevant, dynamic coaching based on the current data [120], which in turn may increase adherence to the intervention.

In addition, factors associated with the target population may also affect the selection of the most suitable delivery channel and operating system (eg, Android or Apple's iOS). For example, Kamita et al [71] implemented their CA on the messaging platform "LINE" as it was the most popular social network service in Japan, and Wang et al [75] selected WeChat, the most common messaging app in Hong Kong.

Communication Modalities

Aligned with the framework focus, text would be the CA's main input and output modality. Messages should be brief, fit the mobile screen without scrolling [69], and be of an adequate font size to allow for comfortable reading. Moreover, if the CA targets populations for whom reading might be challenging, such as older adults or visually impaired individuals, text-to-speech assistive technology may be incorporated into the app.

Visual aids such as images or videos are useful to adapt content to audiences with lower educational attainment [121], deliver personal narratives relevant to the end users (eg, young people with cancer), or decrease the amount of textual information [76]. When using multimedia content, it is important to use high-resolution files to avoid pixelated or blurred images. Furthermore, if pictures are obtained from the web, developers should abide by copyright regulations and either source the pictures from free stock photo repositories, acquire the image rights, or produce the images in-house.

Assembling a Multidisciplinary Team

The composition of the design and development team would be based on the objectives of the intervention. In addition to the inclusion of health professionals with the relevant expertise, it is recommended to include end users as well [69,70]. For example, a CA to support a lifestyle intervention in overweight adolescents was developed by a multidisciplinary team including computer scientists, physicians, a psychotherapist, and diet and sports experts [72]. End-user involvement in the intervention design is critical to ensure that it aligns with user needs. User involvement was reported in a large number of studies in our review (36/55, 65%); for example, young people with cancer participated in focus groups to refine the content of a CA aimed at delivering positive psychology to enhance well-being [70], and young patients with asthma and their parents were part of a multidisciplinary team of experts who developed a CA to improve cognitive and behavioral skills [73]. In general, studies that mentioned the composition of their multidisciplinary teams often reported computer scientists and physicians as key members [72-75], although other health professionals such as physiotherapists [78], psychologists [3], and music therapists [76] may be included as well.

Step 2: Development

Developing the Content

Content development may involve determining the sources of information, adapting content to the target audience, defining the behavior change theories and techniques guiding the intervention [28,94], and establishing error management and safety-netting strategies [26-28, 30, 67-70, 73, 75, 77, 79, 80, 82, 83, 85, 87, 90, 92-97, 99, 100].

Evidence-Based Information

All health-related information included in the CA should be derived from reputable sources and adequately referenced. Sources of evidence-based information include comprehensive literature reviews; clinical practice guidelines; Cochrane systematic reviews; and reputable organization websites such

XSL•FO
RenderX

as the World Health Organization, MEDLINE Plus, and the Centers for Disease Control and Prevention in the United States or the National Health Service Health A to Z in the United Kingdom [65]. For example, Kowatsch et al [73] used evidence from multiple sources such as published literature on the improvement of asthma management in children [122], technology acceptance research [123], and user-CA working alliances [124] to inform their intervention for asthma management.

Managing Errors

Another important aspect of content development is to ensure an adequate understanding of user requests, particularly for potentially serious or life-threatening health conditions. Safeguards to be implemented within the dialog include the request for clarification if the CA receives an unfamiliar input or directing the user to contact an HCP or a human administrator [125,126]. These strategies were included in TensioBot, an intervention to facilitate self-measurement of blood pressure where, after obtaining confirmation of a blood pressure measurement value outside the normal range, the CA alerted the attending physician [68]. Important strategies to manage unintended errors include using validated data entry fields; limiting the data input to predetermined number ranges, words, or characters; or including predefined options for the user to select.

Safety Netting

In general, health care CAs should include a disclaimer clearly stating that the intervention "does not replace healthcare provider's advice." Furthermore, in the case of health conditions associated with rapid deterioration of patient status leading to medical emergencies, such as cardiovascular conditions, diabetes, chronic pulmonary disorders, or mental health conditions that increase the risk of suicide, information should be included to assist users in managing an emergency situation, such as the provision of emergency services or crisis helpline telephone numbers [127], links to contact their primary physician, or clear advice on first aid treatments such as offering a sugary drink to manage a hypoglycemic event in a person with diabetes [128].

Types of Messages

The content and style of the messages should be aligned with the health condition and CA aim. Broadly, the messages may be educational [30,78] or motivational [34,77,79] or deliver reminders to perform a self-management task [68], input data [77], comply with preset tasks [73], take a medication, or attend an HCP appointment [68]. For CAs tasked with engaging with the user during clinic visits, it may be useful to include a status report or summary of the consultation [126].

CAs assuming a coach-like persona might emphasize sympathy, empathy, and participants' achievements [78]. Interventions attempting to modify users' behavior may deliver messages with higher emotional content, as reported in the study by Carfora et al [80], where only emotional messages led users to reduce red meat consumption. In addition, the Wang et al [75] CA used 4 types of messages to deliver a smoking cessation intervention: group announcements, health-related information, reminders to share positive results and progress, and fixed answers to frequently asked questions or requests.

Behavior Change Theories

CAs are increasingly used to promote behavior change [1,129]. Behavior change interventions are complex [130] and often comprise one or more behavior change techniques (BCTs) to induce change. In our assessment, 4% (2/55) of the studies used a behavior change theory to guide the intervention design, including the Health Action Process Approach [78] and the technology acceptance model [71]. In addition, 13% (7/55) of the studies [31,72-75,77,80] reported the use of specific BCTs such as goal setting, self-monitoring, tracking and feedback, social support, use of rewards, and anticipated regret.

For example, a study described a multicomponent behavior change intervention incorporating several BCTs, such as goal setting, self-monitoring, stimulus control, and behavioral contract, to support a healthy lifestyle for adolescents with obesity [34,79]. Furthermore, including group chats where peers or HCPs offer relevant information and emotional support may also assist in promoting positive behavior change, such as using a CA-led WeChat peer group to promote smoking cessation [75].

Optional Add-ons

Depending on the purpose of the CA, it may be appropriate to integrate data from external devices such as glucometers [131] or activity trackers [119]. Alternatively, access to smartphone sensor data [132] may facilitate passive monitoring of the user's activity [79] or determine novel digital biomarkers to assess the user's mood [133] or disease status [134]. The use of smartphone sensors for passive monitoring may further allow for real-time information sharing with HCPs, caregivers, or peers, a feature that may be particularly useful to monitor older people living alone, who may be at higher risk of falling, or individuals with severe chronic illnesses and multiple hospital admissions.

Building the Conversation Flow

A good CA is eloquent and knowledgeable and, thus, requires a meticulously crafted script. Conversation flow building was discussed in 35% (19/55) of the papers in our literature search [3,27,28,30-32,65,73,78,79,82,85,87,92-96,99].

Providing Suitable Answer Options

For a good conversation flow, the predefined answer options should be sufficient and appropriate to align with the user intent, defined as the user goals or intentions in each conversation turn. Constructing a mind map outlining the different facets associated with a topic (eg, medication adherence) and the likely influencing factors (lifestyle components or emotional state) would help predict the most relevant answer options to provide to the user [135].

Selecting a Mapping Tool

A mind map is a diagram representing concepts, ideas, or tasks generated from a key concept, which is generally represented in the center of the graph [136]. Mind maps are an effective method of brainstorming [137] that can be applied to building the conversation flows. Several web-based programs and platforms are available to organize the conversation flow,

XSL•FO
RenderX

including tools specifically designed to build the CA conversation, such as SAP Conversational AI [138] or MobileCoach [35]. Conversation flows may also be built using nonspecific mind mapping software such as Xmind [139]. Mind mapping is useful to assist in recording the flow of conversations between different topics or different user interactions. A well-constructed conversation flow leads the conversation, guides the user, and can address all relevant questions about its purpose. Furthermore, interactivity, personalization, and consistent messaging have been noted as valued qualities [140].

Personalizing Content and Delivery

Interventions should be tailored to individual participant needs [110]. When compared with generic CAs, context, situational, or individually aware agents promote a more positive user experience [132]. Personalized interventions include addressing the user by their name or nickname [141]; delivering notifications and reminders tailored to individual needs [110], such as medication or appointment reminders; and notifications for missed activities or unread messages [30,78]. For example, an intervention promoting self-management of chronic pain offered personalized content based on the user's type and duration of pain and personal interests [78].

An important caveat involves the design of interventions offering personalized advice based on user measurements, such as suggesting a treatment based on individually reported data (eg, blood glucose levels or blood pressure readings), as these interventions may require regulatory oversight and be considered a "mobile medical application" [142].

Selecting Appropriate Message Timing and Frequency

The timing and frequency of messages are important components when planning the intervention and may be determined by the intervention scope as well as user preference. Earlier studies on SMS text messaging interventions have suggested a preference for weekly messaging [113]. However, different intervention types may require a more adaptive message delivery system, such as smoking cessation programs that often require an increased volume of messages close to the desired quit date [143] or high-risk behavior prevention programs targeting binge drinking or inappropriate sexual behaviors timing their messages to when the risky behavior is expected to occur, for example, on a Friday night [116,141]. Therefore, strategies for message delivery and frequency could be adapted to suit the CA intervention.

Just-in-time adaptive interventions (JITAIs) leverage smartphone sensor data to "provide the right type (or amount) of support at the right time" [22]. Smartphone sensor data would determine and even predict "states of vulnerability" (susceptibility to negative health outcomes) [144] and "states of receptivity" (the capacity to receive, process, and use the intervention) [120] in the user when the intervention may be required and more useful. This novel approach may be particularly useful for behavior change interventions supporting a healthy lifestyle, such as increasing physical activity or adhering to a healthier diet, or supporting substance use remission [22,120]. Nevertheless, researchers considering this approach should take into account human and economic resources as JITAI design may require a larger development team that includes computer scientists and app developers.

Using Engagement Strategies

Strategies to keep the users engaged for the intended duration of the intervention are particularly important in health care settings. These aspects were discussed in 11% (6/55) of the studies in our reviews [3,30,31,73,78,79]. Reported strategies included notifications, weekly summaries, reminders, motivational statements, persuasive techniques, a high frequency of messages to promote habit formation, and daily encouragement. In addition, CA-specific engagement strategies included building rapport and attachment with the user [72,73] or adding gamified components to incentivize CA use for rewards and points [73,79].

Step 3: Evaluation and Implementation

Evaluation

The evaluation of digital interventions, including CAs, starts early in the development process and comprises several iterative steps. To ensure the validity of the results, the process must use a robust methodology that is adequate for the intervention design [15]. In digital health interventions, a commonly used evaluation methodology is the multiphase optimization strategy by Collins et al [145,146].

The CA evaluation follows 3 distinct stages representing the intervention development process. The initial iterations of the CA may be evaluated using one or more usability testing methods [147] aiming to produce a minimum viable prototype. Once this working prototype is ready, pilot and randomized trials may ensue to assess the effectiveness of the CA [148]. Several aspects of CA evaluation were discussed in 36% (20/55) of the studies in our reviews [25, 26,28, 33, 71,73, 83, 85, 86, 88, 89, 91-95, 97-100].

The evaluation design may include one or more aspects of the CA functionalities, including clinical or technical attributes and user experience. The outcomes should be clearly defined and include widely used and validated outcome measurement tools whenever possible to improve the comparability and reproducibility of the research results. Examples of outcome measurement tools include the Patient Health Questionnaire-9 [149] to screen for depression, the Flourishing Scale [150] to assess psychological well-being, the Brief Pain Inventory [151] to assess pain intensity and its interference in activities of daily living, and the Working Alliance Inventory-Short Revised [152] to evaluate the CA-user working alliance.

Usability Testing

The evaluation of the CA should start early in the development cycle [153]. In the initial stages, formative evaluation aims to assess the viability of the digital tool by assessing its usability, usefulness, and user experience [154] using one or more qualitative or quantitative research designs. Qualitative methods include surveys, interviews, focus group discussions, and "think aloud" protocols [147] in which users express their opinions about the product as they use it. Quantitative methods include closed-ended questionnaires, task completion assessments, and A/B testing [147,155]. An A/B test, split test, or controlled

XSL•FO
RenderX

experiment compares two or more versions of a product to evaluate the intervention components that perform better or are preferred by the user [155]. This stage relates to the screening and confirming stages in the multiphase optimization strategy [145,146], which use a fractional factorial design to assess which components should be included in the digital intervention and the best dosages to use in a more cost-effective fashion. Finally, microrandomized trials are another novel methodology that is particularly useful for assessing and optimizing the delivery of JITAIs [156]. Microrandomized trials allow the randomization of multiple components to occur at multiple times triggered by predefined decision points [156] and have been used to evaluate CA interventions, as reported by Kramer et al [119,157].

Efficacy and Effectiveness of the CA Intervention

Once initial evaluations have determined the components that should be included in the intervention and the frequency of administration, a traditional randomized trial design should be implemented to assess the effectiveness of the CA intervention compared with current best practices [145,146,148]. Given the complexities and cost that a full-powered randomized controlled trial often entails, researchers may consider conducting a pilot study to refine the study methodology or assess the feasibility of the study design and participant recruitment strategies, among other aspects [158]. For example, Casas et al [77] conducted a pilot study to preliminarily assess a CA aimed at coaching participants to make healthier food choices, whereas Greer et al [70] evaluated a CA delivering a positive psychological intervention to young people with cancer.

User Engagement and Acceptability

Overview

Digital health interventions often report high rates of participant attrition, which may limit the validity of research findings and, more importantly, the effectiveness of the intervention. Therefore, the assessment of the CA-led intervention should be complemented by regular evaluations of end-user adherence to as well as engagement with and acceptability of the intervention. Several assessment methods are commonly used, including quantitative, data-driven analyses and qualitative assessments of users' opinions.

Data-Driven Analyses

The definition of adherence to digital health interventions refers to the extent to which a user has interacted with the intervention [159]. This term may be used to define the degree to which a user interacts with the CA (greater adherence equals more time engaging with the intervention) or the degree to which the user-CA interaction complies with the prescribed recommendation (intended use of the intervention) [159]. In health care interventions, the concept of "intended use" is preferred, and it should be clearly defined during the CA design and development stage for the subsequent adherence measurements to be meaningful. Increased adherence to an intervention may be related to its increased effectiveness [75,160], although the data are not conclusive [3,161,162].

User engagement with the CA may be evaluated using data metrics such as the times the user opened the app, time spent interacting with the CA, the extent of the dialog, or the number of screens opened if the CA also includes other functions [70]. Chaix et al [30] measured use duration, interest in various educational contents, and level of interactivity as indicators of engagement. Nevertheless, researchers should consider the challenges of defining engagement with digital interventions, which may include other user-related variables such as the severity or stage of the disease as well as the long-term engagement with the CA [163].

Other aspects of CA use, such as underused or missing topics or CA functionalities not working as intended, may also be assessed. CA use analytics are often embedded in host platforms. Commercial platforms such as ManyChat [164] may offer a variety of built-in analytics tools such as the number of times the CA is accessed. Some of these platforms offer free-of-charge services. For health care CAs, the open-source MobileCoach platform [35] offers flexible, customizable use analytics.

Qualitative Evaluation

Acceptability refers to the "affective attitudes towards a new digital health intervention" [165]. It is a dynamic concept comprising the intention to engage with the novel CA, the actual interaction with the CA, and the postengagement satisfaction [165].

Acceptability is a subjective term that is generally assessed using questionnaires or other qualitative methods such as focus groups or interviews. For example, Kowatsch et al [73] evaluated the acceptance of a CA to support asthma self-management using a 7-point Likert scale (strongly agree-strongly disagree) for perceived usefulness, ease of use, enjoyment, and use intention, and Echeazarra et al [68] used a survey with questions on ease of use, preference for the CA over existing methods, CA usefulness for its intended purpose, and whether the user had stopped using it as measures of acceptability and satisfaction. Furthermore, Gabrielli et al [69] facilitated a participatory design workshop where suggestions for improvement were provided via open-ended questions, and Ly et al [3] conducted semistructured interviews on the benefits, opportunities, and challenges associated with the CA for mental health. Yan et al [166] described a very involved process of evaluation of an mHealth intervention to promote physical activity. A focus group discussion was organized whereby each SMS text message was displayed and participants were required to respond either with "Yes, I like it" or "No, let's change it to make it better." This voting was then followed by a discussion in which suboptimal messages were improved and the strengths of effective messages were noted. Finally, participants may also be questioned about their willingness to recommend the conversation to others, which is a good indicator of satisfaction and acceptability [70].

Several aspects of user engagement and acceptability may be measured using one of several app quality rating tools, of which the most commonly used one is the Mobile App Rating Scale [167]. The use of standardized, validated rating scales may improve the reproducibility of this research area and facilitate the reporting of trial results, although they are not specific for CAs.

XSL•FO
RenderX

Economic Evaluation

The economic evaluation includes not only the affordability of the project but also the cost-benefits associated with developing the CA. These analyses should consider the end-user perspective as well as the potential benefits for the health care system in general [168,169]. Digital health interventions appear to be cost-effective [170], although reports often present varying, inconclusive results [171]. Although it is often mentioned that one of the potential advantages of digital health interventions, particularly in the long term, may be a significant decrease in health care costs [172], the upfront expenses of developing the digital intervention might be substantial. For example, Kowatsch et al [73] reported upfront expenses of approximately US $250,000 to develop a CA to support asthma self-management in young patients. The development costs will vary conditional to the type and functionalities of the CA, the use of a messaging platform or development as a stand-alone app, and the number of team members, among other aspects. Despite the increasing importance of conducting economic evaluations of digital health care interventions, only 2% (1/55) of the studies included in our reviews reported economic evaluation data [73]. Recent documents from the World Health Organization [168] and the International Training and Education Center for Health [169] at the University of Washington in the United States, as well as a recent review [171], present a practical overview of how to perform economic evaluations.

Implementation

Once the effectiveness of the CA intervention has been determined in rigorous clinical trials, the research team should consider implementing the intervention in the broader population. Implementation research aims to integrate research and practice [173] and understand the users and context in which an intervention would be implemented. The research methods, including pragmatic trials, participatory action research, and mixed methods studies, aim to assess the intervention "acceptability, adoption, appropriateness, feasibility, fidelity, implementation cost, coverage, and sustainability" [174-177]. Important considerations include the need to upgrade the systems to adapt to higher traffic, personnel to provide long-term system maintenance and updates, and the costs these changes may incur [25,26]. Furthermore, the team should consider CA intervention commercialization strategies, including engaging HCPs, health insurers, or governmental organizations if aligned with the health care focus of the intervention [26].

Finally, the team should be aware of and comply with the current regulatory frameworks for digital health interventions. Increasingly, countries are developing national policy frameworks to regulate the evaluation, use, and commercialization of digital health interventions [178], particularly if the intervention is considered a digital therapeutic [179]. Digital therapeutics refer to "evidence-based therapeutic interventions that are driven by high-quality software programs to prevent, manage, or treat a medical disorder or disease" [179], may require a provider's prescription to be accessed [179], and often require approval from official regulatory bodies such as the Food and Drug Administration in the United States [180] and the Conformité Européenne mark in the European Union (EU) [181].

Cross-Cutting Considerations

The themes described in this section are relevant throughout all the design stages referred to in the previous sections.

User-Centered Design and Co-design

User-centered design refers to design practices that include the end users' views to guide the process, either in a passive, consultive manner or as active participants in the design process (co-design) [182]. Several approaches to user-centered design have been described. They share the general principles of involving users during the design process, although the steps involved in the process and the type and extent of end-user involvement may differ. They include but may not be limited to human-centered design [183,184] and design thinking [185] (often considered synonyms), user-centered design [186], co-design [182], and participatory action research [187].

End users include patients, caregivers, HCPs, or other relevant stakeholders. There are several benefits of including end users as part of the CA development team, such as a better understanding of users' and communities' needs, development of culturally sensitive products, and improved communication between the different stakeholders [188,189]. This, in turn, may increase compliance with the intervention and improve health-related outcomes [190]. For example, to develop a CA to promote positivity and well-being in young people after cancer treatment, Greer et al [70] conducted interviews and focus groups with young adults treated for cancer to refine the informational content.

During the evaluation stage, thinking-out-loud usability testing is another example of a user-centered design methodology in the design of digital health interventions, including CAs [191].

The role of user-centered design in the development of digital health interventions has been repeatedly emphasized by several frameworks included in our review (36/55, 65%) [25,26,36-39,41-46,62,69,70,73,76,82-85,87-100,110].

Privacy and Security

Overview

Safeguarding the privacy and security of CA users' data is essential and should be a part of the entire design and development cycle. Health information is considered personal, sensitive information that should be protected at all times. The level of data protection should align with the data collected by the CA, if any. Therefore, the functionalities of the CA will determine the type of sensitive data to be collected and guide the inclusion of data protection software such as firewalls and encryption.

In general, developers should minimize the amount of personal and sensitive information collected from users by asking specific questions to avoid oversharing or simply providing predetermined responses instead of using free text. Furthermore, all CAs should include a privacy policy that is brief and written in clear language outlining the data collected and the uses of these data. All data must be encrypted during transit (when the

XSL•FO
RenderX

message is being sent) and at rest (when the message has been delivered) [192]. The platform on which the CA will be deployed may also vary according to the CA functionalities. For example, a CA collecting users' personal data should not be deployed on proprietary or messaging platforms as the platform data management policies may not be clearly reported [65] or data sharing with third parties may occur without informing the user [193]. This might create an ever-increasing digital footprint, potentially allowing for user identification from data aggregation rather than actually identifiable information [194].

A 2020 framework for governing the responsible use of CAs in health care highlighted the importance of safeguarding data privacy, including user health data, history of interactions, and disclosure of user data even if unintended [195]. In addition, the framework highlighted the user's right to access their personally identifiable information, the requirement of user consent before recording or saving health-related data, and the preclusion of using the stored data as a means of surveillance or to discriminate users against health care privileges or coverage [195].

Compliance With Data Privacy Laws

Health care CAs that collect users' sensitive data must comply with country-relevant data privacy laws, such as the Health Insurance Portability and Accountability Act in the United States [192] or the General Data Protection Regulation (GDPR) in the EU [196]. These laws' jurisdiction is generally limited to the issuing country; however, the GDPR applies to any EU citizen within or outside the EU. The GDPR, which went into effect in 2018, is an overarching law that aims to enhance the rights of individuals over their personal data, defined as any data that may allow for the identification of a person on their own or combined with other data, including pseudonymized data [196]. Alternatively, the Health Insurance Portability and Accountability Act is industry-specific and applies only to health-related data [197]. Other countries have adopted their own data protection laws and regulations. In Singapore, the Personal Data Protection Act is a baseline regulatory framework informing the collection, distribution, and use of personal data [198].

In addition to the aforementioned GDPR, children's data are generally more stringently safeguarded. For example, in the United States, the Children's Online Privacy Protection Act [199] requires that verifiable parental consent be obtained by all digital operators (not restricted to health care) collecting data from children (aged <13 years). Similar considerations are included within the GDPR and the Singapore Personal Data Protection Act, with the caveat that, in some European countries, parental consent is required for children and adolescents aged <16 years.

Discussion

Principal Findings

We present a new conceptual framework for the design, development, evaluation, and implementation of smartphone-delivered, rule-based, and text-based CAs. The DISCOVER conceptual framework includes 8 iterative steps arranged in three main groups: (1) design, which includes defining the goal, creating an identity, assembling the team, and selecting the delivery interface; (2) development, which comprises developing the content and building the conversation flow; and (3) evaluation and implementation. User-centered design and privacy and security were included as cross-cutting considerations, which are relevant at every stage of the framework.

This framework was based on the comprehensive analysis of 36 mHealth frameworks, 5 CA taxonomies, and 14 primary studies reporting on the design and development of rule-based health care CAs. The framework was applied in a web-based pilot study using a CA deployed on Facebook Messenger. The existing mHealth frameworks provided general guidelines to develop mHealth interventions for health care, from the characterization of the target population to evaluation, with emphasis on the application of user-centered design techniques in all stages of development. Concurrently, the CA taxonomies provided focused on several aspects of CA design and evaluation as well as the impact of design features on CA-user interactions.

Considering the multifaceted nature of embodied CAs, we decided to focus on CAs that are nonembodied.

Comparisons With Prior Work

The existing frameworks for the design and development of mHealth interventions provide detailed guidance in all steps of the intervention development, starting with an understanding of the needs and the profile of the end users through a review of existing literature or formative research [67], and they emphasize the need for patient and public involvement to make the intervention as relevant to the target population as possible [90,98]. These frameworks also described the importance of conducting iterative evaluations to identify limitations before testing the mHealth intervention in a larger-scale trial [28,98,99]. However, the literature on the design and development of CAs was restricted to the development of taxonomies that were not limited to health care describing CA design platforms [5], classification of CAs according to the approach to conversation design [67], characteristics of embodied agents [66], or the impact of CA characteristics on user interactions [32]. Moreover, the taxonomy by Denecke et al [65] referred to health care CAs, but they focused exclusively on CA evaluation. Therefore, a conceptual framework guiding the development of health care CAs was needed to expand previous mHealth frameworks with elements particularly relevant to CAs, such as personality development, converting evidence-based content into conversations, and using novel research designs for evaluation. Furthermore, our framework focused particularly on the development of the CA, including personality, display of empathy, and disclosure of its identity as a computer-generated entity without human involvement, and on the development of dialogs guided by up-to-date evidence-based information sources.

This framework described the development of rule-based CAs, allowing the research team total control of the conversation and dialog flow. There are several reasons for this. First, our framework presents easy-to-follow steps that could be applied

XSL•FO
RenderX

by smaller research teams that do not include computer science or AI specialists or that undertake the CA development project under restricted financial resources. Second, we aimed to provide guidance for the development of goal-oriented CAs aimed at delivering health education content or simple interventions aimed at improving healthy lifestyle choices or self-management behavior and, therefore, prioritize control over the conversation content using a rule-based paradigm, albeit less engaging, over AI algorithms that have yet to become truly explainable.

Implications for Future Research

Future research should apply the DISCOVER conceptual framework to the development of CAs offering behavior change interventions aimed at different specialties, settings (hospital or outpatient), target groups, and cultures. Moreover, although the use of theories in the design of behavior change interventions is favored and may increase the effectiveness of the intervention [77,200], it is still unclear which behavior change theories or techniques are better suited for CA-led interventions. Alternatively, because of the interactive nature of CAs, it would be appropriate to assess whether behavior change interventions previously proved effective in traditional face-to-face settings are equally effective when led by a CA.

Although the concepts of identity creation, conversational flow, and delivery are important, their relative relevance to varying target populations is still unknown. In addition, more research on the assessment of health care chatbot interventions can help inform the ideal health-related outcome measures and digital data sets required for a comprehensive evaluation. Finally, although this framework is comprehensive and many components may apply to AI CAs, a separate framework is needed to describe specific aspects relevant to AI CAs, such as dialog development using machine learning or natural language processing techniques, voice versus text parsing, and many others.

Strengths

This is, to the best of our knowledge, the first conceptual framework outlining the steps required to develop a smartphone-delivered, rule-based health care CA offering clear yet comprehensive guidelines to accommodate health care researchers with varying computer science expertise.

The DISCOVER framework builds on an analysis of existing mHealth frameworks and a stringent analysis of rule-based CA literature complemented by the team's demonstration of its applicability in the development of a rule-based CA to support lifestyle changes in people at risk of developing diabetes.

Limitations

Much of the information provided is anecdotal or derived from research conducted on SMS text messaging and other mHealth interventions because of the scarcity of research on the evidence-based development of rule-based CAs for health care. Therefore, this framework provides an overview of the main steps required to develop a rule-based CA.

The descriptions and examples presented in the conceptual framework focused on CA interventions for end users to support either a healthy lifestyle or the management of a chronic condition, as derived from the literature reviews and our experience developing a CA. Nevertheless, the design and development principles discussed in this study could apply to other relevant user groups such as caregivers and health care professionals.

Furthermore, this framework is focused on rule-based CAs and, although it may guide researchers in the development of particular aspects of AI CAs, it does not provide guidance on the development of AI-based conversations. In addition, the economic, social, and behavioral characteristics of different populations may limit its generalizability.

Conclusions

The interest in and potential for CAs in health care are growing, but guidelines to design, develop, evaluate, and implement these interventions are currently lacking. Drawing on published evidence, the DISCOVER conceptual framework provides the first attempt to fill this void. The process was divided into 8 iterative steps arranged in 3 overarching groups and complemented by 2 cross-cutting considerations. Future research should explore aspects of CA development such as the use of behavior change theories and privacy and safety concerns. Further evaluation of this framework in diverse health care areas and settings and for a variety of users is needed to demonstrate its validity.

Acknowledgments

This research was supported by the Ageing Research Institute for Society and Education (ARISE), Nanyang Technological University, Singapore. This study was also supported by the Singapore Ministry of Education under the Singapore Ministry of Education Academic Research Fund Tier 1 (RG36/20). This research was conducted as part of the Future Health Technologies program, which was established collaboratively between ETH Zürich and the National Research Foundation, Singapore. This research was supported by the National Research Foundation, Prime Minister's Office, Singapore, under its Campus for Research Excellence and Technological Enterprise program.

Authors' Contributions

DAD designed the study, extracted the data, conducted the analysis, and wrote the manuscript. LM conducted the analysis and wrote the manuscript. M-HRH, SJ, TK, and RA provided a critical review of the manuscript. LTC conceptualized and designed

XSL•FO
RenderX

the study, provided a critical review of the manuscript, and provided supervision at all steps of the research. All authors approved the final version of the manuscript, and they take accountability for all aspects of this work.

Conflicts of Interest

TK is affiliated with the Centre for Digital Health Interventions, a joint initiative of the Institute for Implementation Science in Health Care at the University of Zurich; the Department of Management, Technology and Economics at ETH Zurich; the Future Health Technologies Programme at the Singapore-ETH Centre; and the School of Medicine and Institute of Technology Management at the University of St. Gallen. Centre for Digital Health Interventions is funded in part by CSS, a Swiss health insurer. TK is also a cofounder of Pathmate Technologies, a university spin-off company that creates and delivers digital clinical pathways. However, neither CSS nor Pathmate Technologies were involved in this study. SJ is also affiliated with Salesforce Research. However, Salesforce Research was not involved in this study. The other authors declare that they have no conflicts of interest.

Multimedia Appendix 1
Literature review of conceptual frameworks for the design, development, and evaluation of mobile health interventions.
[DOCX File , 122 KB - mhealth_v10i10e38740_app1.docx]

Multimedia Appendix 2
Literature review of smartphone-delivered, rule-based conversational agents.
[DOCX File , 30 KB - mhealth_v10i10e38740_app2.docx]

Multimedia Appendix 3
Search strategy for the conversational agent research trial review.
[DOCX File , 30 KB - mhealth_v10i10e38740_app3.docx]

Multimedia Appendix 4
Methodology implemented for conceptual framework development using the conceptual framework development steps described by Jabareen [29].
[DOCX File , 19 KB - mhealth_v10i10e38740_app4.docx]

Multimedia Appendix 5
Mapping of the steps of the conceptual framework applied to the design, development, and evaluation of Precilla.
[DOCX File , 22 KB - mhealth_v10i10e38740_app5.docx]

Multimedia Appendix 6
Design, development, and evaluation frameworks for mobile health interventions.
[DOCX File , 41 KB - mhealth_v10i10e38740_app6.docx]

Multimedia Appendix 7
Classification systems for conversational agents.
[DOCX File , 20 KB - mhealth_v10i10e38740_app7.docx]

Multimedia Appendix 8
Characteristics of clinical trials on rule-based conversational agents.
[DOCX File , 27 KB - mhealth_v10i10e38740_app8.docx]

References

1. Tudor Car L, Dhinagaran DA, Kyaw BM, Kowatsch T, Joty S, Theng YL, et al. Conversational agents in health care: scoping review and conceptual analysis. J Med Internet Res 2020 Aug 07;22(8):e17158 [FREE Full text] [doi: 10.2196/17158] [Medline: 32763886]
2. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health chatbots: scoping review. J Med Internet Res 2021 Jan 13;23(1):e17828 [FREE Full text] [doi: 10.2196/17828] [Medline: 33439133]
3. Ly KH, Ly AM, Andersson G. A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. Internet Interv 2017 Dec;10:39-46 [FREE Full text] [doi: 10.1016/j.invent.2017.10.002] [Medline: 30135751]

XSL•FO
RenderX

4. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res 2020 Oct 22;22(10):e20346 [FREE Full text] [doi: 10.2196/20346] [Medline: 33090118]
5. Diederich S, Brendel AB, Kolbe LM. Towards a taxonomy of platforms for conversational agent design. In: Proceedings of the 14th International Conference on Wirtschaftsinformatik. 2019 Presented at: WI '19; February 24-27, 2019; Siegen, Germany p. 1100-1114 URL: https://web.archive.org/web/20200619062226id_/https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1247&context=wi2019
6. Mohamad Suhaili S, Salim N, Jambli MN. Service chatbots: a systematic review. Expert Syst Appl 2021 Dec;184:115461. [doi: 10.1016/j.eswa.2021.115461]
7. Bavaresco R, Silveira D, Reis E, Barbosa J, Righi R, Costa C, et al. Conversational agents in business: a systematic literature review and future research directions. Comput Sci Rev 2020 May;36:100239. [doi: 10.1016/j.cosrev.2020.100239]
8. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 06;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
9. Gaffney H, Mansell W, Edwards R, Wright J. Manage Your Life Online (MYLO): a pilot trial of a conversational computer-based intervention for problem solving in a student sample. Behav Cogn Psychother 2014 Nov;42(6):731-746. [doi: 10.1017/S135246581300060X] [Medline: 23899405]
10. Ireland D, Liddle J, McBride S, Ding H, Knuepffer C. Chat-bots for people with Parkinson's disease: science fiction or reality? Stud Health Technol Inform 2015;214:128-133. [Medline: 26210429]
11. Jack A. Rwanda venture tests digital health potential in developing world. Financial Times. 2021 Jan 18. URL: https://www.ft.com/content/4fe33c92-cbd5-459a-8df6-20d0d1f57ec8 [accessed 2022-01-19]
12. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians' perceptions of chatbots in health care: cross-sectional Web-based survey. J Med Internet Res 2019 Apr 05;21(4):e12887 [FREE Full text] [doi: 10.2196/12887] [Medline: 30950796]
13. Joshi I. Waiting for deep medicine. Lancet 2019 Mar 23;393(10177):1193-1194. [doi: 10.1016/S0140-6736(19)30579-3]
14. Ramesh K, Ravishankaran S, Joshi A, Chandrasekaran K. A survey of design techniques for conversational agents. In: Proceedings of the 2nd International Conference on Information, Communication and Computing Technology. 2017 Presented at: ICICCT '17; May 13, 2017; New Delhi, India p. 336-350. [doi: 10.1007/978-981-10-6544-6_31]
15. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 01;25(9):1248-1258 [FREE Full text] [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
16. Gao J, Galley M, Li L. Neural Approaches to Conversational AI: Question Answering, Task-Oriented Dialogues and Social Chatbots. Delft, The Netherlands: Now Publishers; 2019.
17. Wadden JJ. Defining the undefinable: the black box problem in healthcare artificial intelligence. J Med Ethics (forthcoming) 2021 Jul 21:medethics-2021-107529-2021-107529. [doi: 10.1136/medethics-2021-107529] [Medline: 34290113]
18. Quinn TP, Jacobs S, Senadeera M, Le V, Coghlan S. The three ghosts of medical AI: can the black-box present deliver? Artif Intell Med 2022 Feb;124:102158. [doi: 10.1016/j.artmed.2021.102158] [Medline: 34511267]
19. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021 Nov;3(11):e745-e750 [FREE Full text] [doi: 10.1016/S2589-7500(21)00208-9] [Medline: 34711379]
20. Silver L. Smartphone ownership is growing rapidly around the world, but not always equally. Pew Research Center. 2019 Feb 5. URL: https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/ [accessed 2022-03-28]
21. Dogan E, Sander C, Wagner X, Hegerl U, Kohls E. Smartphone-based monitoring of objective and subjective data in affective disorders: where are we and where are we going? Systematic review. J Med Internet Res 2017 Jul 24;19(7):e262 [FREE Full text] [doi: 10.2196/jmir.7006] [Medline: 28739561]
22. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
23. Marcolino MS, Oliveira JA, D'Agostino M, Ribeiro AL, Alkmim MB, Novillo-Ortiz D. The impact of mHealth interventions: systematic review of systematic reviews. JMIR Mhealth Uhealth 2018 Jan 17;6(1):e23 [FREE Full text] [doi: 10.2196/mhealth.8873] [Medline: 29343463]
24. Eze P, Lawani LO, Acharya Y. Short message service (SMS) reminders for childhood immunisation in low-income and middle-income countries: a systematic review and meta-analysis. BMJ Glob Health 2021 Jul;6(7):e005035 [FREE Full text] [doi: 10.1136/bmjgh-2021-005035] [Medline: 34290051]
25. Whittaker R, Merry S, Dorey E, Maddison R. A development and evaluation process for mHealth interventions: examples from New Zealand. J Health Commun 2012;17 Suppl 1:11-21. [doi: 10.1080/10810730.2011.649103] [Medline: 22548594]

XSL•FO
RenderX

26. Mummah SA, Robinson TN, King AC, Gardner CD, Sutton S. IDEAS (Integrate, Design, Assess, and Share): a framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. J Med Internet Res 2016 Dec 16;18(12):e317 [FREE Full text] [doi: 10.2196/jmir.5927] [Medline: 27986647]
27. Mohr DC, Schueller SM, Montague E, Burns MN, Rashidi P. The behavioral intervention technology model: an integrated conceptual and technological framework for eHealth and mHealth interventions. J Med Internet Res 2014 Jun 05;16(6):e146 [FREE Full text] [doi: 10.2196/jmir.3077] [Medline: 24905070]
28. Zhang J, Oh YJ, Lange P, Yu Z, Fukuoka Y. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet: viewpoint. J Med Internet Res 2020 Sep 30;22(9):e22845 [FREE Full text] [doi: 10.2196/22845] [Medline: 32996892]
29. Jabareen Y. Building a conceptual framework: philosophy, definitions, and procedure. Int J Qual Methods 2009 Dec 01;8(4):49-62. [doi: 10.1177/160940690900800406]
30. Chaix B, Bibault JE, Pienkowski A, Delamon G, Guillemassé A, Nectoux P, et al. When chatbots meet patients: one-year prospective study of conversations between patients with breast cancer and a chatbot. JMIR Cancer 2019 May 02;5(1):e12856 [FREE Full text] [doi: 10.2196/12856] [Medline: 31045505]
31. Heldt K, Büchter D, Brogle B, Chen-Hsuan IS, Rüegger D, Filler A, et al. Telemedicine therapy for overweight adolescents: first results of a novel smartphone app intervention using a behavioural health platform. Obes Facts 2018 May 26;11(Suppl 1):214-215 [FREE Full text] [doi: 10.1159/000489691] [Medline: 29804106]
32. Zierau N, Elshan E, Visini C, Janson A. A review of the empirical literature on conversational agents and future research directions. In: Proceedings of the 2020 International Conference on Information Systems. 2020 Presented at: ICIS '20; December 13-16, 2020; Hyderabad, India p. 1414.
33. l'Allemand D, Shih CH, Heldt K, Büchter D, Brogle B, Rüegger D, et al. Design and interim evaluation of a smartphone app for overweight adolescents using a behavioural health intervention platform. Obes Rev 2018 Dec 03;19(Suppl 1):102. [doi: 10.1111/obr.12778] [Medline: 30511506]
34. Stasinaki A, Brogle B, Buchter D, Shih CH, Heldt K, White C, et al. A novel digital health intervention improves physical performance in obese youth. In: Joint Annual Meeting Swiss Society of Paediatrics, Swiss Society of Paediatric Surgery and Swiss Society of Child and Adolescent Psychiatry and Psychotherapy. 2018 Presented at: SMW '18; May 24-25, 2018; Lausanne, Switzerland p. 10S. [doi: 10.4414/smw.2018.20430]
35. MobileCoach. URL: https://www.mobile-coach.eu/ [accessed 2022-03-02]
36. Bauerle Bass S, Jessop A, Gashat M, Maurer L, Alhajji M, Forry J. Take Charge, Get Cured: the development and user testing of a culturally targeted mHealth decision tool on HCV treatment initiation for methadone patients. Patient Educ Couns 2018 Nov;101(11):1995-2004. [doi: 10.1016/j.pec.2018.07.007] [Medline: 30055893]
37. Chen Y, Wu F, Wu Y, Li J, Yue P, Deng Y, et al. Development of interventions for an intelligent and individualized mobile health care system to promote healthy diet and physical activity: using an intervention mapping framework. BMC Public Health 2019 Oct 17;19(1):1311 [FREE Full text] [doi: 10.1186/s12889-019-7639-7] [Medline: 31623589]
38. Curtis KE, Lahiri S, Brown KE. Targeting parents for childhood weight management: development of a theory-driven and user-centered healthy eating app. JMIR Mhealth Uhealth 2015 Jun 18;3(2):e69 [FREE Full text] [doi: 10.2196/mhealth.3857] [Medline: 26088692]
39. Farao J, Malila B, Conrad N, Mutsvangwa T, Rangaka MX, Douglas TS. A user-centred design framework for mHealth. PLoS One 2020 Aug 19;15(8):e0237910 [FREE Full text] [doi: 10.1371/journal.pone.0237910] [Medline: 32813711]
40. Jindal D, Gupta P, Jha D, Ajay VS, Goenka S, Jacob P, et al. Development of mWellcare: an mHealth intervention for integrated management of hypertension and diabetes in low-resource settings. Glob Health Action 2018;11(1):1517930 [FREE Full text] [doi: 10.1080/16549716.2018.1517930] [Medline: 30253691]
41. Kazemi DM, Borsari B, Levine MJ, Lamberson KA, Dooley B. REMIT: development of a mHealth theory-based intervention to decrease heavy episodic drinking among college students. Addict Res Theory 2018;26(5):377-385 [FREE Full text] [doi: 10.1080/16066359.2017.1420783] [Medline: 32694964]
42. McBride B, Nguyen LT, Wiljer D, Vu NC, Nguyen CK, O'Neil J. Development of a maternal, newborn and child mHealth intervention in Thai Nguyen Province, Vietnam: protocol for the mMom project. JMIR Res Protoc 2018 Jan 11;7(1):e6 [FREE Full text] [doi: 10.2196/resprot.7912] [Medline: 29326095]
43. Patel S, Arya M. The BUS framework: a comprehensive tool in creating an mHealth app utilizing behavior change theories, user-centered design, and social marketing. J Mob Technol Med 2017 Apr;6(1):39-45 [FREE Full text] [doi: 10.7309/jmtm.6.1.6] [Medline: 29104674]
44. Sporrel K, De Boer RD, Wang S, Nibbeling N, Simons M, Deutekom M, et al. The design and development of a personalized leisure time physical activity application based on behavior change theories, end-user perceptions, and principles from empirical data mining. Front Public Health 2020 Feb 2;8:528472 [FREE Full text] [doi: 10.3389/fpubh.2020.528472] [Medline: 33604321]
45. Sun CJ, Anderson KM, Kuhn T, Mayer L, Klein CH. A sexual health promotion app for transgender women (trans women connected): development and usability study. JMIR Mhealth Uhealth 2020 May 12;8(5):e15888 [FREE Full text] [doi: 10.2196/15888] [Medline: 32396131]

XSL•FO
RenderX

46. Verbiest ME, Corrigan C, Dalhousie S, Firestone R, Funaki T, Goodwin D, et al. Using codesign to develop a culturally tailored, behavior change mHealth intervention for indigenous and other priority communities: a case study in New Zealand. Transl Behav Med 2019 Jul 16;9(4):720-736. [doi: 10.1093/tbm/iby093] [Medline: 30388262]
47. Bartholomew LK, Parcel GS, Kok G. Intervention mapping: a process for developing theory- and evidence-based health education programs. Health Educ Behav 1998 Oct;25(5):545-563. [doi: 10.1177/109019819802500502] [Medline: 9768376]
48. Norman DA. User Centered System Design: New Perspectives on Human-Computer Interaction. Boca Raton, FL, USA: CRC Press; 1986.
49. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
50. Hevner AR, March ST, Park J, Ram S. Design science in information systems research. MIS Q 2004 Mar;28(1):75-105. [doi: 10.2307/25148625]
51. Get Started with Design Thinking. Hasso Plattner Institute of Design at Stanford University. 2022. URL: https://dschool.stanford.edu/resources/getting-started-with-design-thinking [accessed 2022-02-03]
52. Michie S, Johnston M, Abraham C, Lawton R, Parker D, Walker A, "Psychological Theory" Group. Making psychological theory useful for implementing evidence based practice: a consensus approach. Qual Saf Health Care 2005 Feb;14(1):26-33 [FREE Full text] [doi: 10.1136/qshc.2004.011155] [Medline: 15692000]
53. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061 [FREE Full text] [doi: 10.1136/bmj.n2061] [Medline: 34593508]
54. Dijkstra A, De Vries H. The development of computer-generated tailored interventions. Patient Educ Couns 1999 Feb;36(2):193-203. [doi: 10.1016/s0738-3991(98)00135-9] [Medline: 10223023]
55. Oinas-Kukkonen H, Harjumaa M. Persuasive systems design: key issues, process model, and system features. Commun Assoc Inf Syst 2009 May 1;24:485-500. [doi: 10.17705/1CAIS.02428]
56. Michie S, Hyder N, Walia A, West R. Development of a taxonomy of behaviour change techniques used in individual behavioural support for smoking cessation. Addict Behav 2011 Apr;36(4):315-319. [doi: 10.1016/j.addbeh.2010.11.016] [Medline: 21215528]
57. Bratteteig T, Bødker K, Dittrich Y, Holst Mogensen P, Simonsen J. Methods: organising principles and general guidelines for participatory design projects. In: Simonsen J, Robertson T, editors. Routledge International Handbook of Participatory Design. London, UK: Routledge; 2012:117-144.
58. Wagner EH, Austin BT, Davis C, Hindmarsh M, Schaefer J, Bonomi A. Improving chronic illness care: translating evidence into action. Health Aff (Millwood) 2001;20(6):64-78. [doi: 10.1377/hlthaff.20.6.64] [Medline: 11816692]
59. Provoost S, Lau HM, Ruwaard J, Riper H. Embodied conversational agents in clinical psychology: a scoping review. J Med Internet Res 2017 May 09;19(5):e151 [FREE Full text] [doi: 10.2196/jmir.6553] [Medline: 28487267]
60. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
61. mHealth: new horizons for health through mobile technologies: second global survey on eHealth. World Health Organization. 2011. URL: https://apps.who.int/iris/bitstream/handle/10665/44607/9789241564250_eng.pdf?sequence=1&isAllowed=y [accessed 2022-09-21]
62. Modi D, Gopalan R, Shah S, Venkatraman S, Desai G, Desai S, et al. Development and formative evaluation of an innovative mHealth intervention for improving coverage of community-based maternal, newborn and child health services in rural areas of India. Glob Health Action 2015 Feb 16;8:26769 [FREE Full text] [doi: 10.3402/gha.v8.26769] [Medline: 25697233]
63. Dhinagaran DA, Sathish T, Kowatsch T, Griva K, Best JD, Tudor Car L. Public perceptions of diabetes, healthy living, and conversational agents in Singapore: needs assessment. JMIR Form Res 2021 Nov 11;5(11):e30435 [FREE Full text] [doi: 10.2196/30435] [Medline: 34762053]
64. Dhinagaran DA, Sathish T, Soong A, Theng Y, Best J, Tudor Car L. Conversational agent for healthy lifestyle behavior change: Web-based feasibility study. JMIR Form Res 2021 Dec 03;5(12):e27956 [FREE Full text] [doi: 10.2196/27956] [Medline: 34870611]
65. Denecke K, Warren J. How to evaluate health applications with conversational user interface? Stud Health Technol Inform 2020 Jun 16;270:976-980. [doi: 10.3233/SHTI200307] [Medline: 32570527]
66. ter Stal S, Kramer LL, Tabak M, op den Akker H, Hermens H. Design features of embodied conversational agents in eHealth: a literature review. Int J Human Comput Stud 2020 Jun;138:102409. [doi: 10.1016/j.ijhcs.2020.102409]
67. Hussain S, Sianaki OA, Ababneh N. A survey on conversational agents/chatbots classification and design techniques. In: Proceedings of the Workshops of the 33rd International Conference on Advanced Information Networking and Applications. 2019 Presented at: WAINA '19; March 27-29, 2019; Matsue, Japan p. 946-956. [doi: 10.1007/978-3-030-15035-8_93]
68. Echeazarra L, Pereira J, Saracho R. TensioBot: a chatbot assistant for self-managed in-house blood pressure checking. J Med Syst 2021 Mar 15;45(4):54. [doi: 10.1007/s10916-021-01730-x] [Medline: 33723721]
69. Gabrielli S, Rizzi S, Carbone S, Donisi V. A chatbot-based coaching intervention for adolescents to promote life skills: pilot study. JMIR Hum Factors 2020 Feb 14;7(1):e16762 [FREE Full text] [doi: 10.2196/16762] [Medline: 32130128]

XSL•FO
RenderX

70. Greer S, Ramo D, Chang YJ, Fu M, Moskowitz J, Haritatos J. Use of the chatbot "Vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: randomized controlled feasibility trial. JMIR Mhealth Uhealth 2019 Oct 31;7(10):e15018 [FREE Full text] [doi: 10.2196/15018] [Medline: 31674920]
71. Kamita T, Ito T, Matsumoto A, Munakata T, Inoue T. A chatbot system for mental healthcare based on SAT counseling method. Mob Inf Syst 2019 Mar 03;2019:1-11. [doi: 10.1155/2019/9517321]
72. Kowatsch T, Nißen M, Shih CH, Rüegger D, Volland D, Filler A, et al. Text-based healthcare chatbots supporting patient and health professional teams: preliminary results of a randomized controlled trial on childhood obesity. In: Persuasive Embodied Agents for Behavior Change. 2017 Presented at: PEACH '17; August 27, 2017; Stockholm, Sweden. [doi: 10.3929/ethz-b-000218776]
73. Kowatsch T, Schachner T, Harperink S, Barata F, Dittler U, Xiao G, et al. Conversational agents as mediating social actors in chronic disease management involving health care professionals, patients, and family members: multisite single-arm feasibility study. J Med Internet Res 2021 Feb 17;23(2):e25060 [FREE Full text] [doi: 10.2196/25060] [Medline: 33484114]
74. Kowatsch T, Volland D, Shih I, Rüegger D, Künzler F, Barata F, et al. Design and evaluation of a mobile chat app for the open source behavioral health intervention platform MobileCoach. In: Proceedings of the 12th International Conference on Designing the Digital Transformation. 2017 Presented at: DESRIST '17; May 30-June 1, 2017; Karlsruhe, Germany p. 485-489. [doi: 10.1007/978-3-319-59144-5_36]
75. Wang H, Zhang Q, Ip M, Fai Lau JT. Social media–based conversational agents for health management and interventions. Computer 2018 Aug;51(8):26-33. [doi: 10.1109/mc.2018.3191249]
76. Denecke K, Hochreutener SL, Pöpel A, May R. Self-anamnesis with a conversational user interface: concept and usability study. Methods Inf Med 2018 Nov;57(05-06):243-252. [doi: 10.1055/s-0038-1675822] [Medline: 30875703]
77. Casas J, Mugellini E, Khaled O. Food diary coaching chatbot. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. 2018 Presented at: UbiComp '18; October 8-12, 2018; Singapore, Singapore p. 1676-1680. [doi: 10.1145/3267305.3274191]
78. Hauser-Ulrich S, Künzli H, Meier-Peterhans D, Kowatsch T. A smartphone-based health care chatbot to promote self-management of chronic pain (SELMA): pilot randomized controlled trial. JMIR Mhealth Uhealth 2020 Apr 03;8(4):e15806 [FREE Full text] [doi: 10.2196/15806] [Medline: 32242820]
79. Stasinaki A, Büchter D, Shih CH, Heldt K, Güsewell S, Brogle B, et al. Effects of a novel mobile health intervention compared to a multi-component behaviour changing program on body mass index, physical capacities and stress parameters in adolescents with obesity: a randomized controlled trial. BMC Pediatr 2021 Jul 09;21(1):308 [FREE Full text] [doi: 10.1186/s12887-021-02781-2] [Medline: 34243738]
80. Carfora V, Bertolotti M, Catellani P. Informational and emotional daily messages to reduce red and processed meat consumption. Appetite 2019 Oct 01;141:104331. [doi: 10.1016/j.appet.2019.104331] [Medline: 31276710]
81. Depp C, Torous J, Thompson W. Technology-based early warning systems for bipolar disorder: a conceptual framework. JMIR Ment Health 2016 Sep 07;3(3):e42 [FREE Full text] [doi: 10.2196/mental.5798] [Medline: 27604265]
82. Direito A, Walsh D, Hinbarji M, Albatal R, Tooley M, Whittaker R, et al. Using the intervention mapping and behavioral intervention technology frameworks: development of an mHealth intervention for physical activity and sedentary behavior change. Health Educ Behav 2018 Jun;45(3):331-348. [doi: 10.1177/1090198117742438] [Medline: 29216765]
83. Duff O, Walsh D, Malone S, McDermott L, Furlong B, O'Connor N, et al. MedFit app, a behavior-changing, theoretically informed mobile app for patient self-management of cardiovascular disease: user-centered development. JMIR Form Res 2018 Apr 27;2(1):e8 [FREE Full text] [doi: 10.2196/formative.9550] [Medline: 30684426]
84. Fedele DA, McConville A, Moon J, Thomas JG. Topical review: design considerations when creating pediatric mobile health interventions: applying the IDEAS framework. J Pediatr Psychol 2019 Apr 01;44(3):343-348 [FREE Full text] [doi: 10.1093/jpepsy/jsy084] [Medline: 30346561]
85. Good A, Omisade O. Linking activity theory with user centred design: a human computer interaction framework for the design and evaluation of mHealth interventions. Stud Health Technol Inform 2019 Jul 30;263:49-63. [doi: 10.3233/SHTI190110] [Medline: 31411152]
86. Horvath KJ, Ecklund AM, Hunt SL, Nelson TF, Toomey TL. Developing Internet-based health interventions: a guide for public health researchers and practitioners. J Med Internet Res 2015 Jan 23;17(1):e28 [FREE Full text] [doi: 10.2196/jmir.3770] [Medline: 25650702]
87. Jennings HM, Morrison J, Akter K, Kuddus A, Ahmed N, Kumer Shaha S, et al. Developing a theory-driven contextually relevant mHealth intervention. Glob Health Action 2019;12(1):1550736 [FREE Full text] [doi: 10.1080/16549716.2018.1550736] [Medline: 31154988]
88. Pham Q, Graham G, Lalloo C, Morita PP, Seto E, Stinson JN, et al. An analytics platform to evaluate effective engagement with pediatric mobile health apps: design, development, and formative evaluation. JMIR Mhealth Uhealth 2018 Dec 21;6(12):e11447 [FREE Full text] [doi: 10.2196/11447] [Medline: 30578179]
89. Ravn Jakobsen P, Hermann AP, Søndergaard J, Wiil UK, Clemensen J. Development of an mHealth application for women newly diagnosed with osteoporosis without preceding fractures: a participatory design approach. Int J Environ Res Public Health 2018 Feb 13;15(2):330 [FREE Full text] [doi: 10.3390/ijerph15020330] [Medline: 29438343]

XSL•FO
RenderX

90. Skrabal Ross X, Gunn KM, Patterson P, Olver I. Development of a smartphone program to support adherence to oral chemotherapy in people with cancer. Patient Prefer Adherence 2019 Dec 20;13:2207-2215 [FREE Full text] [doi: 10.2147/PPA.S225175] [Medline: 31908427]
91. Schellong J, Lorenz P, Weidner K. Proposing a standardized, step-by-step model for creating post-traumatic stress disorder (PTSD) related mobile mental health apps in a framework based on technical and medical norms. Eur J Psychotraumatol 2019 May 15;10(1):1611090 [FREE Full text] [doi: 10.1080/20008198.2019.1611090] [Medline: 31143412]
92. Schnall R, Rojas M, Bakken S, Brown W, Carballo-Dieguez A, Carry M, et al. A user-centered model for designing consumer mobile health (mHealth) applications (apps). J Biomed Inform 2016 Apr;60:243-251 [FREE Full text] [doi: 10.1016/j.jbi.2016.02.002] [Medline: 26903153]
93. Tay I, Garland S, Gorelik A, Wark JD. Development and testing of a mobile phone app for self-monitoring of calcium intake in young women. JMIR Mhealth Uhealth 2017 Mar 07;5(3):e27 [FREE Full text] [doi: 10.2196/mhealth.5717] [Medline: 28270379]
94. van Agteren JE, Lawn S, Bonevski B, Smith BJ. Kick.it: the development of an evidence-based smoking cessation smartphone app. Transl Behav Med 2018 Mar 01;8(2):243-267. [doi: 10.1093/tbm/ibx031] [Medline: 29447386]
95. Vilardaga R, Rizo J, Zeng E, Kientz JA, Ries R, Otis C, et al. User-centered design of learn to quit, a smoking cessation smartphone app for people with serious mental illness. JMIR Serious Games 2018 Jan 16;6(1):e2 [FREE Full text] [doi: 10.2196/games.8881] [Medline: 29339346]
96. Wilhide Iii CC, Peeples MM, Anthony Kouyaté RC. Evidence-based mHealth chronic disease mobile app intervention design: development of a framework. JMIR Res Protoc 2016 Feb 16;5(1):e25 [FREE Full text] [doi: 10.2196/resprot.4838] [Medline: 26883135]
97. Wittenberg E, Xu J, Goldsmith J, Mendoza Y. Caregiver communication about cancer: development of a mHealth resource to support family caregiver communication burden. Psychooncology 2019 Feb;28(2):365-371 [FREE Full text] [doi: 10.1002/pon.4950] [Medline: 30471236]
98. Woods L, Cummings E, Duff J, Walker K. Design thinking for mHealth application co-design to support heart failure self-management. Stud Health Technol Inform 2017;241:97-102. [Medline: 28809190]
99. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
100. Fjeldsoe BS, Miller YD, O'Brien JL, Marshall AL. Iterative development of MobileMums: a physical activity intervention for women with young children. Int J Behav Nutr Phys Act 2012 Dec 20;9:151 [FREE Full text] [doi: 10.1186/1479-5868-9-151] [Medline: 23256730]
101. Nißen M, Selimi D, Janssen A, Cardona DR, Breitner MH, Kowatsch T, et al. See you soon again, chatbot? A design taxonomy to characterize user-chatbot relationships with different time horizons. Comput Human Behav 2022 Feb;127:107043. [doi: 10.1016/j.chb.2021.107043]
102. Burkill S, Copas A, Couper MP, Clifton S, Prah P, Datta J, et al. Using the Web to collect data on sensitive behaviours: a study looking at mode effects on the British national survey of sexual attitudes and lifestyles. PLoS One 2016 Feb 11;11(2):e0147983 [FREE Full text] [doi: 10.1371/journal.pone.0147983] [Medline: 26866687]
103. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. Digit Health 2019 Aug 21;5:2055207619871808 [FREE Full text] [doi: 10.1177/2055207619871808] [Medline: 31467682]
104. Boucher EM, Harake NR, Ward HE, Stoeckl SE, Vargas J, Minkel J, et al. Artificially intelligent chatbots in digital mental health interventions: a review. Expert Rev Med Devices 2021 Dec;18(sup1):37-49. [doi: 10.1080/17434440.2021.2013200] [Medline: 34872429]
105. Curtis RG, Bartel B, Ferguson T, Blake HT, Northcott C, Virgara R, et al. Improving user experience of virtual health assistants: scoping review. J Med Internet Res 2021 Dec 21;23(12):e31737 [FREE Full text] [doi: 10.2196/31737] [Medline: 34931997]
106. De Cicco R, Silva SC, Alparone FR. Millennials' attitude toward chatbots: an experimental study in a social relationship perspective. Int J Retail Distrib Manag 2020 Jul 08;48(11):1213-1233. [doi: 10.1108/IJRDM-12-2019-0406]
107. Etemad-Sajadi R. The impact of online real-time interactivity on patronage intention: the use of avatars. Comput Human Behav 2016 Aug;61:227-232. [doi: 10.1016/j.chb.2016.03.045]
108. Thaler M, Schlögl S, Groth A. Agent vs. avatar: comparing embodied conversational agents concerning characteristics of the uncanny valley. In: Proceedings of the 2020 IEEE International Conference on Human-Machine Systems. 2020 Presented at: ICHMS '20; September 7-9, 2020; Rome, Italy p. 1-6. [doi: 10.1109/ICHMS49158.2020.9209539]
109. Bickmore T, Gruber A, Picard R. Establishing the computer-patient working alliance in automated health behavior change interventions. Patient Educ Couns 2005 Oct;59(1):21-30. [doi: 10.1016/j.pec.2004.09.008] [Medline: 16198215]
110. Morton K, Sutton S, Hardeman W, Troughton J, Yates T, Griffin S, et al. A text-messaging and pedometer program to promote physical activity in people at high risk of type 2 diabetes: the development of the PROPELS follow-on support program. JMIR Mhealth Uhealth 2015 Dec 15;3(4):e105 [FREE Full text] [doi: 10.2196/mhealth.5026] [Medline: 26678750]

XSL•FO
RenderX

111. Mac OA, Thayre A, Tan S, Dodd RH. Web-based health information following the renewal of the cervical screening program in Australia: evaluation of readability, understandability, and credibility. J Med Internet Res 2020 Jun 26;22(6):e16701 [FREE Full text] [doi: 10.2196/16701] [Medline: 32442134]
112. Kincaid JP, Fishburne Jr RP, Rogers RL, Chissom BS. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for navy enlisted personnel. Institute for Simulation and Training, University of Central Florida. 1975. URL: https://stars.library.ucf.edu/istlibrary/56/ [accessed 2022-01-19]
113. Bobrow K, Farmer A, Cishe N, Nwagi N, Namane M, Brennan TP, et al. Using the Medical Research Council framework for development and evaluation of complex interventions in a low resource setting to develop a theory-based treatment support intervention delivered via SMS text message to improve blood pressure control. BMC Health Serv Res 2018 Jan 23;18(1):33 [FREE Full text] [doi: 10.1186/s12913-017-2808-9] [Medline: 29361934]
114. van Dijk CN, van Witteloostuijn M, Vasić N, Avrutin S, Blom E. The influence of texting language on grammar and executive functions in primary school children. PLoS One 2016 Mar 31;11(3):e0152409 [FREE Full text] [doi: 10.1371/journal.pone.0152409] [Medline: 27031708]
115. Fadhil A, Schiavo G, Wang Y, Yilma BA. The effect of emojis when interacting with conversational interface assisted health coaching system. In: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare. 2018 Presented at: PervasiveHealth '18; May 21-24, 2018; New York, NY, USA p. 378-383. [doi: 10.1145/3240925.3240965]
116. Menacho LA, Blas MM, Alva IE, Roberto Orellana E. Short text messages to motivate HIV testing among men who have sex with men: a qualitative study in Lima, Peru. Open AIDS J 2013 Apr 5;7:1-6 [FREE Full text] [doi: 10.2174/1874613601307010001] [Medline: 23802032]
117. Bock BC, Rosen RK, Barnett NP, Thind H, Walaska K, Foster R, et al. Translating behavioral interventions onto mHealth platforms: developing text message interventions for smoking and alcohol. JMIR Mhealth Uhealth 2015 Feb 24;3(1):e22 [FREE Full text] [doi: 10.2196/mhealth.3779] [Medline: 25714907]
118. Adamopoulou E, Moussiades L. An overview of chatbot technology. In: Proceedings of the 16th International Conference on Artificial Intelligence Applications and Innovations. 2020 Presented at: AIAI '20; June 5–7, 2020; Neos Marmaras, Greece p. 373-383. [doi: 10.1007/978-3-030-49186-4_31]
119. Kramer JN, Künzler F, Mishra V, Smith SN, Kotz D, Scholz U, et al. Which components of a smartphone walking app help users to reach personalized step goals? Results from an optimization trial. Ann Behav Med 2020 Jun 12;54(7):518-528 [FREE Full text] [doi: 10.1093/abm/kaaa002] [Medline: 32182353]
120. Künzler F, Mishra V, Kramer JN, Kotz D, Fleisch E, Kowatsch T. Exploring the state-of-receptivity for mHealth interventions. Proc ACM Interact Mob Wearable Ubiquitous Technol 2019 Dec 11;3(4):1-27. [doi: 10.1145/3369805]
121. Meppelink CS, van Weert JC, Haven CJ, Smit EG. The effectiveness of health animations in audiences with different health literacy levels: an experimental study. J Med Internet Res 2015 Jan 13;17(1):e11 [FREE Full text] [doi: 10.2196/jmir.3979] [Medline: 25586711]
122. Klok T, Kaptein AA, Brand PL. Non-adherence in children with asthma reviewed: the need for improvement of asthma care and medical education. Pediatr Allergy Immunol 2015 May;26(3):197-205. [doi: 10.1111/pai.12362] [Medline: 25704083]
123. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003 Sep;27(3):425-478. [doi: 10.2307/30036540]
124. Bickmore TW, Picard RW. Establishing and maintaining long-term human-computer relationships. ACM Trans Comput Hum Interact 2005 Jun 01;12(2):293-327. [doi: 10.1145/1067860.1067867]
125. Fadhil A. Can a chatbot determine my diet?: addressing challenges of chatbot application for meal recommendation. arXiv 2022 Preprint posted online on February 25, 2018. [doi: 10.48550/arXiv.1802.09100]
126. Radziwill NM, Benton MC. Evaluating quality of chatbots and intelligent conversational agents. arXiv 2022 Preprint posted online on April 15, 2017. [doi: 10.48550/arXiv.1704.04579]
127. Martinengo L, Van Galen L, Lum E, Kowalski M, Subramaniam M, Car J. Suicide prevention and depression apps' suicide risk assessment and management: a systematic assessment of adherence to clinical guidelines. BMC Med 2019 Dec 19;17(1):231 [FREE Full text] [doi: 10.1186/s12916-019-1461-z] [Medline: 31852455]
128. Lum E, Jimenez G, Huang Z, Thai L, Car J. Appropriateness of action prompts for hypoglycaemia and hyperglycaemia in type 2 diabetes self-management apps. Diabetes Metab Res Rev 2020 Feb;36(2):e3235. [doi: 10.1002/dmrr.3235] [Medline: 31721421]
129. Pereira J, Díaz Ó. Using health chatbots for behavior change: a mapping study. J Med Syst 2019 Apr 04;43(5):135. [doi: 10.1007/s10916-019-1237-1] [Medline: 30949846]
130. Michie S, Wood CE, Johnston M, Abraham C, Francis JJ, Hardeman W. Behaviour change techniques: the development and evaluation of a taxonomic method for reporting and describing behaviour change interventions (a suite of five studies involving consensus methods, randomised controlled trials and analysis of qualitative data). Health Technol Assess 2015 Nov;19(99):1-188 [FREE Full text] [doi: 10.3310/hta19990] [Medline: 26616119]

XSL•FO
RenderX

131. Krishnakumar A, Verma R, Chawla R, Sosale A, Saboo B, Joshi S, et al. Evaluating glycemic control in patients of South Asian origin with type 2 diabetes using a digital therapeutic platform: analysis of real-world data. J Med Internet Res 2021 Mar 25;23(3):e17908 [FREE Full text] [doi: 10.2196/17908] [Medline: 33764306]
132. Chatterjee S, Price A. Healthy living with persuasive technologies: framework, issues, and challenges. J Am Med Inform Assoc 2009;16(2):171-178 [FREE Full text] [doi: 10.1197/jamia.M2859] [Medline: 19074300]
133. Rykov Y, Thach TQ, Bojic I, Christopoulos G, Car J. Digital biomarkers for depression screening with wearable devices: cross-sectional study with machine learning modeling. JMIR Mhealth Uhealth 2021 Oct 25;9(10):e24872 [FREE Full text] [doi: 10.2196/24872] [Medline: 34694233]
134. Tinschert P, Rassouli F, Barata F, Steurer-Stey C, Fleisch E, Puhan MA, et al. Nocturnal cough and sleep quality to assess asthma control and predict attacks. J Asthma Allergy 2020 Dec 14;13:669-678 [FREE Full text] [doi: 10.2147/JAA.S278155] [Medline: 33363391]
135. Fadhil A. A conversational interface to improve medication adherence: towards AI support in patient's treatment. arXiv 2022 Preprint posted online on March 3, 2018. [doi: 10.48550/arXiv.1803.09844]
136. Burgess-Allen J, Owen-Smith V. Using mind mapping techniques for rapid qualitative data analysis in public participation processes. Health Expect 2010 Dec;13(4):406-415 [FREE Full text] [doi: 10.1111/j.1369-7625.2010.00594.x] [Medline: 20550595]
137. Davies M. Concept mapping, mind mapping and argument mapping: what are the differences and do they matter? High Educ 2011;62(3):279-301. [doi: 10.1007/s10734-010-9387-6]
138. Low code chatbot building platform. SAP Conversational AI. URL: https://cai.tools.sap/ [accessed 2022-03-02]
139. XMind. URL: https://www.xmind.net/ [accessed 2022-01-19]
140. Hall AK, Cole-Lewis H, Bernhardt JM. Mobile text messaging for health: a systematic review of reviews. Annu Rev Public Health 2015 Mar 18;36:393-415 [FREE Full text] [doi: 10.1146/annurev-publhealth-031914-122855] [Medline: 25785892]
141. Coomes CM, Lewis MA, Uhrig JD, Furberg RD, Harris JL, Bann CM. Beyond reminders: a conceptual framework for using short message service to promote prevention and improve healthcare quality and clinical outcomes for people living with HIV. AIDS Care 2012;24(3):348-357. [doi: 10.1080/09540121.2011.608421] [Medline: 21933036]
142. Policy for Device Software Functions and Mobile Medical Applications: Guidance for Industry and Food and Drug Administration Staff. U.S. Food & Drug Administration. 2019 Sep 27. URL: https://www.fda.gov/media/80958/download [accessed 2022-03-03]
143. Mbuagbaw L, Mursleen S, Lytvyn L, Smieja M, Dolovich L, Thabane L. Mobile phone text messaging interventions for HIV and other chronic diseases: an overview of systematic reviews and framework for evidence transfer. BMC Health Serv Res 2015 Jan 22;15:33 [FREE Full text] [doi: 10.1186/s12913-014-0654-6] [Medline: 25609559]
144. Mishra V, Künzler F, Kramer JN, Fleisch E, Kowatsch T, Kotz D. Detecting receptivity for mHealth interventions in the natural environment. Proc ACM Interact Mob Wearable Ubiquitous Technol 2021 Jun;5(2):74 [FREE Full text] [doi: 10.1145/3463492] [Medline: 34926979]
145. Collins LM, Murphy SA, Nair VN, Strecher VJ. A strategy for optimizing and evaluating behavioral interventions. Ann Behav Med 2005 Aug;30(1):65-73. [doi: 10.1207/s15324796abm3001_8] [Medline: 16097907]
146. Collins LM, Kugler KC. Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: Advanced Topics. Cham, Switzerland: Springer; 2018.
147. Maramba I, Chatterjee A, Newman C. Methods of usability testing in the development of eHealth applications: a scoping review. Int J Med Inform 2019 Jun;126:95-104. [doi: 10.1016/j.ijmedinf.2019.03.018] [Medline: 31029270]
148. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016 Nov;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
149. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
150. Diener E, Wirtz D, Tov W, Kim-Prieto C, Choi D, Oishi S, et al. New well-being measures: short scales to assess flourishing and positive and negative feelings. Soc Indic Res 2009 May 28;97(2):143-156. [doi: 10.1007/s11205-009-9493-y]
151. Cleeland CS, Ryan KM. Pain assessment: global use of the Brief Pain Inventory. Ann Acad Med Singap 1994 Mar;23(2):129-138. [Medline: 8080219]
152. Munder T, Wilmers F, Leonhart R, Linster HW, Barth J. Working Alliance Inventory-Short Revised (WAI-SR): psychometric properties in outpatients and inpatients. Clin Psychol Psychother 2010;17(3):231-239. [doi: 10.1002/cpp.658] [Medline: 20013760]
153. Guo C, Ashrafian H, Ghafur S, Fontana G, Gardner C, Prime M. Challenges for the evaluation of digital health solutions-a call for innovative evidence generation approaches. NPJ Digit Med 2020 Aug 27;3:110 [FREE Full text] [doi: 10.1038/s41746-020-00314-2] [Medline: 32904379]
154. Blandford A, Gibbs J, Newhouse N, Perski O, Singh A, Murray E. Seven lessons for interdisciplinary research on interactive digital health interventions. Digit Health 2018 May 3;4:2055207618770325 [FREE Full text] [doi: 10.1177/2055207618770325] [Medline: 29942629]

XSL•FO
RenderX

155. Kohavi R, Longbotham R. Online controlled experiments and A/B testing. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning and Data Mining. Boston, MA, USA: Springer; 2017:922-929.
156. Klasnja P, Hekler EB, Shiffman S, Boruvka A, Almirall D, Tewari A, et al. Microrandomized trials: an experimental design for developing just-in-time adaptive interventions. Health Psychol 2015 Dec;34S:1220-1228 [FREE Full text] [doi: 10.1037/hea0000305] [Medline: 26651463]
157. Kramer JN, Künzler F, Mishra V, Presset B, Kotz D, Smith S, et al. Investigating intervention components and exploring states of receptivity for a smartphone app to promote physical activity: protocol of a microrandomized trial. JMIR Res Protoc 2019 Jan 31;8(1):e11540 [FREE Full text] [doi: 10.2196/11540] [Medline: 30702430]
158. Thabane L, Ma J, Chu R, Cheng J, Ismaila A, Rios LP, et al. A tutorial on pilot studies: the what, why and how. BMC Med Res Methodol 2010 Jan 06;10:1 [FREE Full text] [doi: 10.1186/1471-2288-10-1] [Medline: 20053272]
159. Sieverink F, Kelders SM, van Gemert-Pijnen JE. Clarifying the concept of adherence to eHealth technology: systematic review on when usage becomes adherence. J Med Internet Res 2017 Dec 06;19(12):e402 [FREE Full text] [doi: 10.2196/jmir.8578] [Medline: 29212630]
160. Cao W, Milks MW, Liu X, Gregory ME, Addison D, Zhang P, et al. mHealth interventions for self-management of hypertension: framework and systematic review on engagement, interactivity, and tailoring. JMIR Mhealth Uhealth 2022 Mar 02;10(3):e29415 [FREE Full text] [doi: 10.2196/29415] [Medline: 35234655]
161. Paz Castro R, Haug S, Filler A, Kowatsch T, Schaub MP. Engagement within a mobile phone-based smoking cessation intervention for adolescents and its association with participant characteristics and outcomes. J Med Internet Res 2017 Nov 01;19(11):e356 [FREE Full text] [doi: 10.2196/jmir.7928] [Medline: 29092811]
162. Achilles MR, Anderson M, Li SH, Subotic-Kerry M, Parker B, O'Dea B. Adherence to e-mental health among youth: considerations for intervention development and research design. Digit Health 2020 May 21;6:2055207620926064 [FREE Full text] [doi: 10.1177/2055207620926064] [Medline: 32547775]
163. Torous J, Michalak EE, O'Brien HL. Digital health and engagement-looking behind the measures and methods. JAMA Netw Open 2020 Jul 01;3(7):e2010918 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.10918] [Medline: 32678446]
164. ManyChat. URL: https://manychat.com/ [accessed 2022-03-04]
165. Perski O, Short CE. Acceptability of digital health interventions: embracing the complexity. Transl Behav Med 2021 Jul 29;11(7):1473-1480 [FREE Full text] [doi: 10.1093/tbm/ibab048] [Medline: 33963864]
166. Yan AF, Stevens P, Wang Y, Weinhardt L, Holt CL, O'Connor C, et al. mHealth text messaging for physical activity promotion in college students: a formative participatory approach. Am J Health Behav 2015 May;39(3):395-408. [doi: 10.5993/AJHB.39.3.12] [Medline: 25741684]
167. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015 Mar 11;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
168. Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. World Health Organization. Geneva, Switzerland: World Health Organization; 2016. URL: https://apps.who.int/iris/bitstream/handle/10665/252183/9789241511766-eng.pdf [accessed 2022-03-29]
169. Babigumira JB, Dolan S, Shade S, Puttkammer N, Bale J, Tolentino H, et al. Applied Economic Evaluation of Digital Health Interventions. Centers for Disease Control and Prevention. 2021 Jan. URL: https://www.go2itech.org/wp-content/uploads/2021/02/I-TECH_HIS_Economic_Evaluation.pdf [accessed 2022-03-29]
170. Jiang X, Ming WK, You JH. The cost-effectiveness of digital health interventions on the management of cardiovascular diseases: systematic review. J Med Internet Res 2019 Jun 17;21(6):e13166 [FREE Full text] [doi: 10.2196/13166] [Medline: 31210136]
171. Gomes M, Murray E, Raftery J. Economic evaluation of digital health interventions: methodological issues and recommendations for practice. Pharmacoeconomics 2022 Apr;40(4):367-378 [FREE Full text] [doi: 10.1007/s40273-022-01130-0] [Medline: 35132606]
172. Topol E. The Topol Review: Preparing the healthcare workforce to deliver the digital future: an independent report on behalf of the Secretary of State for Health and Social Care. National Health Service. 2019 Feb. URL: https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-2019.pdf [accessed 2022-03-17]
173. Theobald S, Brandes N, Gyapong M, El-Saharty S, Proctor E, Diaz T, et al. Implementation research: new imperatives and opportunities in global health. Lancet 2018 Nov 17;392(10160):2214-2228. [doi: 10.1016/S0140-6736(18)32205-0] [Medline: 30314860]
174. Peters DH, Adam T, Alonge O, Agyepong IA, Tran N. Implementation research: what it is and how to do it. BMJ 2013 Nov 20;347:f6753. [doi: 10.1136/bmj.f6753] [Medline: 24259324]
175. Peters DH, Tran NT, Adam T. Implementation research in health: a practical guide. Alliance for Health Policy and Systems Research & World Health Organization. 2013. URL: https://apps.who.int/iris/handle/10665/91758 [accessed 2022-04-11]
176. Bauer MS, Damschroder L, Hagedorn H, Smith J, Kilbourne AM. An introduction to implementation science for the non-specialist. BMC Psychol 2015 Sep 16;3(1):32 [FREE Full text] [doi: 10.1186/s40359-015-0089-9] [Medline: 26376626]

XSL•FO
RenderX

177. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 07;4:50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
178. Essén A, Stern AD, Haase CB, Car J, Greaves F, Paparova D, et al. Health app policy: international comparison of nine countries' approaches. NPJ Digit Med 2022 Mar 18;5(1):31 [FREE Full text] [doi: 10.1038/s41746-022-00573-1] [Medline: 35304561]
179. Understanding DTx. Digital Therapeutics Alliance. URL: https://dtxalliance.org/understanding-dtx/ [accessed 2022-04-11]
180. Device Software Functions Including Mobile Medical Applications. U.S. Food & Drug Administration. 2019. URL: https://tinyurl.com/5n896nhk [accessed 2022-05-31]
181. Clever Compliance. 2020 Mar 17. URL: https://blog.clevercompliance.io/medical-product-compliance/ce-marking-medical-device/ [accessed 2022-05-31]
182. Sanders EB, Stappers PJ. Co-creation and the new landscapes of design. CoDesign 2008 Mar;4(1):5-18. [doi: 10.1080/15710880701875068]
183. Matheson GO, Pacione C, Shultz RK, Klügl M. Leveraging human-centered design in chronic disease prevention. Am J Prev Med 2015 Apr;48(4):472-479. [doi: 10.1016/j.amepre.2014.10.014] [Medline: 25700655]
184. Melles M, Albayrak A, Goossens R. Innovating health care: key characteristics of human-centered design. Int J Qual Health Care 2021 Jan 12;33(Supplement_1):37-44 [FREE Full text] [doi: 10.1093/intqhc/mzaa127] [Medline: 33068104]
185. Plattner H, Meinel C, Leifer L. Design Thinking : Understand - Improve - Apply. Berlin, Germany: Springer; 2011.
186. Still B, Crane K. Fundamentals of User-Centered Design: A Practical Approach. Boca Raton, FL, USA: CRC Press; 2017.
187. Baum F, MacDougall C, Smith D. Participatory action research. J Epidemiol Community Health 2006 Oct;60(10):854-857 [FREE Full text] [doi: 10.1136/jech.2004.028662] [Medline: 16973531]
188. Steen M, Manschot M, De Koning N. Benefits of co-design in service design projects. Int J Des 2011;5(2):53-60 [FREE Full text]
189. Pirinen A. The barriers and enablers of co-design for services. Int J Des 2016;10(3):27-42 [FREE Full text]
190. Georgsson M, Staggers N. An evaluation of patients' experienced usability of a diabetes mHealth system using a multi-method approach. J Biomed Inform 2016 Feb;59:115-129 [FREE Full text] [doi: 10.1016/j.jbi.2015.11.008] [Medline: 26639894]
191. Martinengo L, Spinazze P, Car J. Mobile messaging with patients. BMJ 2020 Mar 16;368:m884. [doi: 10.1136/bmj.m884] [Medline: 32179548]
192. Huckvale K, Torous J, Larsen ME. Assessment of the data sharing and privacy practices of smartphone apps for depression and smoking cessation. JAMA Netw Open 2019 Apr 05;2(4):e192542 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.2542] [Medline: 31002321]
193. Grande D, Luna Marti X, Feuerstein-Simon R, Merchant RM, Asch DA, Lewson A, et al. Health policy and privacy challenges associated with digital technology. JAMA Netw Open 2020 Jul 01;3(7):e208285 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.8285] [Medline: 32644138]
194. Sundareswaran V, Sarkar A. Chatbots RESET: a framework for governing responsible use of conversational AI in healthcare. World Economic Forum. 2020 Dec. URL: http://www3.weforum.org/docs/WEF_Governance_of_Chatbots_in_Healthcare_2020.pdf [accessed 2022-03-17]
195. Rumbold JM, Pierscionek B. The effect of the general data protection regulation on medical research. J Med Internet Res 2017 Feb 24;19(2):e47 [FREE Full text] [doi: 10.2196/jmir.7108] [Medline: 28235748]
196. Summary of the HIPAA Privacy Rule. U.S. Department of Health and Human Services. 2013. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html [accessed 2022-03-17]
197. Wong BY. Data privacy law in Singapore: the personal data protection act 2012. Int Data Priv Law 2017 Nov;7(4):287-302. [doi: 10.1093/idpl/ipx016]
198. Children's Online Privacy Protection Rule ("COPPA"). Federal Trade Commission. URL: https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa [accessed 2022-05-31]
199. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010 Feb 17;12(1):e4 [FREE Full text] [doi: 10.2196/jmir.1376] [Medline: 20164043]
200. Dalgetty R, Miller CB, Dombrowski SU. Examining the theory-effectiveness hypothesis: a systematic review of systematic reviews. Br J Health Psychol 2019 May;24(2):334-356. [doi: 10.1111/bjhp.12356] [Medline: 30793445]

Abbreviations

AI: artificial intelligence
BCT: behavior change technique
CA: conversational agent
DISCOVER: Designing, Developing, Evaluating, and Implementing a Smartphone-Delivered, Rule-Based Conversational Agent
EU: European Union

XSL•FO
RenderX

GDPR: General Data Protection Regulation
HCP: health care provider
JITAI: just-in-time adaptive intervention
mHealth: mobile health

Edited by L Buis; submitted 13.04.22; peer-reviewed by A Islam, L Agrawal, M Jalan; comments to author 24.05.22; revised version received 02.08.22; accepted 26.08.22; published 04.10.22.

Please cite as:
Dhinagaran DA, Martinengo L, Ho MHR, Joty S, Kowatsch T, Atun R, Tudor Car L
Designing, Developing, Evaluating, and Implementing a Smartphone-Delivered, Rule-Based Conversational Agent (DISCOVER): Development of a Conceptual Framework
JMIR Mhealth Uhealth 2022;10(10):e38740
URL: https://mhealth.jmir.org/2022/10/e38740
doi:10.2196/38740
PMID:36194462

©Dhakshenya Ardhithy Dhinagaran, Laura Martinengo, Moon-Ho Ringo Ho, Shafiq Joty, Tobias Kowatsch, Rifat Atun, Lorainne Tudor Car. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 04.10.2022. 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.

XSL•FO
RenderX

Original Paper

Mental Health Mobile Apps in the French App Store: Assessment Study of Functionality and Quality

Florence Carrouel[1*], PhD; Benjamin du Sartz de Vigneulles[1*], MD; Denis Bourgeois[1,2], PhD, Prof Dr; Bernard Kabuth[3,4], PhD; Nicolas Baltenneck[5], PhD; Fanny Nusbaum[1], PhD; Valérie Burge[3,4], MD; Sylvain Roy[1], PhD; Sophie Buchheit[3,4], PhD; Marie-Line Carrion-Martinaud[6], PhD; Catherine Massoubre[7], PhD; Laurie Fraticelli[1*], PhD; Claude Dussart[1,2*], PhD, Prof Dr

[1]Health Systemic Process, Research Unit UR4129, University Claude Bernard Lyon 1, University of Lyon, Lyon, France

[2]Hospices Civils de Lyon, Lyon, France

[3]Prisme Team, Interpsy Laboratory, Research Unit EA4432, University of Lorraine, Nancy, France

[4]Nancy Psychotherapeutic Center, Laxou, France

[5]Development, Individual, Process, Disability, University Lyon 2, Lyon, France

[6]The Constellation, Health and Wellness Center, Rive-de-Gier, France

[7]Department of Psychiatry, Research Unit EA7423, Saint-Etienne University Hospital Center of Saint Etienne, University Jean Monnet, Saint Etienne, France

[*]these authors contributed equally

Corresponding Author:
Florence Carrouel, PhD
Health Systemic Process, Research Unit UR4129
University Claude Bernard Lyon 1
University of Lyon
7 rue Guillaume Paradin
Lyon, 69372
France
Phone: 33 0478785745
Email: florence.carrouel@univ-lyon1.fr

Abstract

Background: Approximately 800 million people, representing 11% of the world's population, are affected by mental health problems. The COVID-19 pandemic exacerbated problems and triggered a decline in well-being, with drastic increase in the incidence of conditions such as anxiety, depression, and stress. Approximately 20,000 mental health apps are listed in mobile app stores. However, no significant evaluation of mental health apps in French, spoken by approximately 300 million people, has been identified in the literature yet.

Objective: This study aims to review the mental health mobile apps currently available on the French Apple App Store and Google Play Store and to evaluate their quality using Mobile App Rating Scale–French (MARS-F).

Methods: Screening of mental health apps was conducted from June 10, 2022, to June 17, 2022, on the French Apple App Store and Google Play Store. A shortlist of 12 apps was identified using the criteria of selection and assessed using MARS-F by 9 mental health professionals. Intraclass correlation was used to evaluate interrater agreement. Mean (SD) scores and their distributions for each section and item were calculated.

Results: The highest scores for MARS-F quality were obtained by *Soutien psy avec Mon Sherpa* (mean 3.85, SD 0.48), Evoluno (mean 3.54, SD 0.72), and Teale (mean 3.53, SD 0.87). Mean engagement scores (section A) ranged from 2.33 (SD 0.69) for *Reflexe reussite* to 3.80 (SD 0.61) for *Soutien psy avec Mon Sherpa*. Mean aesthetics scores (section C) ranged from 2.52 (SD 0.62) for Mental Booster to 3.89 (SD 0.69) for *Soutien psy avec Mon Sherpa*. Mean information scores (section D) ranged from 2.00 (SD 0.75) for Mental Booster to 3.46 (SD 0.77) for *Soutien psy avec Mon Sherpa*. Mean Mobile App Rating Scale subjective quality (section E) score varied from 1.22 (SD 0.26) for *VOS – journal de l'humeur* to 2.69 (SD 0.84) for *Soutien psy avec Mon Sherpa*. Mean app specificity (section F) score varied from 1.56 (SD 0.97) for Mental Booster to 3.31 (SD 1.22) for Evoluno. For all the mental health apps studied, except *Soutien psy avec Mon Sherpa* (11/12, 92%), the subjective quality score was always

XSL•FO
RenderX

lower than the app specificity score, which was always lower than the MARS-F quality score, and that was lower than the rating score from the iPhone Operating System or Android app stores.

Conclusions: Mental health professionals assessed that, despite the lack of scientific evidence, the mental health mobile apps available on the French Apple App Store and Google Play Store were of good quality. However, they are reluctant to use them in their professional practice. Additional investigations are needed to assess their compliance with recommendations and their long-term impact on users.

(JMIR Mhealth Uhealth 2022;10(10):e41282) doi:10.2196/41282

KEYWORDS

mobile apps; behavior change; mental; prevention; mobile health; mHealth; lifestyle; French; well-being

Introduction

Approximately 800 million people, representing 11% of the world's population, are affected by mental health problems [1]. In addition, the COVID-19 pandemic exacerbated problems and triggered a decline in well-being, with a drastic increase in the incidence of conditions such as anxiety, depression, and stress [2]. However, within the World Health Organization's classification of mental, behavioral, and neurodevelopmental disorders (International Classification of Diseases–11), a broad spectrum of mental conditions exists, varying from mild, one-time disorders to severe, chronic, and disabling disorders [3]. Similarly, the mental well-being concept encompasses more than the absence of mental health disorders and symptoms and may include psychological parameters such as subjective autonomy, well-being, personal fulfillment, and positive relationships [4]. There is growing recognition that mental health and mental well-being are distinct entities with specific determinants [5]. Preserving positive identity, maintaining good self-esteem, being able to control and adapt to one's own life, and preventing social isolation and solitude are all positive aspects of well-being that can contribute to optimizing autonomy [6,7].

Mental disorders are a real public health problem [8]. Each year, approximately 20% of adults are affected by anxiety disorders [9]. Depression affects >300 million people worldwide [10,11]. In France, the mental health of French adults deteriorated between 2015 and 2020, partly because of the COVID-19 crisis, as indicated by a longitudinal study of adults [12]. French adults were more affected by depressive or anxiety symptoms. In addition, in recent years, the number of workers reporting complaints of severe stress has increased. Thus, stress is one of the most common occupational health problems. In Europe, the prevalence of men and women reporting work-related stress *always* or *most of the time* is 26% and 27%, respectively [13]. Stress at an early stage of working life can contribute to burnout, depression, and unfavorable work outcomes in later life, depending on the life course perspective [14,15]. Various psychosocial interventions have been suggested to overcome stress, such as relaxation, mindfulness practices, or social engagement, ideally performed in groups [16,17].

New technologies have the potential to overcome these challenges by providing large-scale health literacy programs, low-threshold approaches [18], or education for the population and health care workers [19]. In this sense, disruptive technologies provide a fantastic opportunity for enhancing mental health approaches [20]. Beyond supporting people with mental disorders, apps can also be used to improve their overall well-being by encouraging behavior changes, including the practice of meditation and mindfulness. Mobile apps related to mental health are known as mental health apps and vary in type and number [21]. In 2022, global spending on mental health mobile apps is estimated, not including China, to be approximately US $500 million [22]. Currently, there are reportedly up to 20,000 mental health apps [22].

Growing digital device use for mental health support also suggests mental health apps as a possible pertinent aspect for a proactive mental health and wellness model in the coming years [23]. Apps that capture the user's moods or emotional conditions and then deliver information have significant and positive support, as do apps offering regular, short meditation sessions [24]. Mental health apps promote positive mental health and well-being, involving a decrease in symptoms of mental health conditions such as anxiety, stress, and depression. Mental health apps have the potential to be effective in improving the symptoms of certain mental disorders [25] and to improve well-being [26] and life satisfaction, with a more effective emotional management [21,24]. During the first 10 months of 2020, consumers spent a record of US $1.1 billion on wellness apps worldwide [22].

Nevertheless, no exploratory studies, which identify the development and application trends of evidence-based apps on topics such as outcomes of mental health or well-being apps, are currently available [25]. Further comprehensive validation and evaluation of these apps from a clinical perspective is needed [27].

The study aimed to review the mental health mobile apps currently available on the French Apple App Store and Google Play Store and to evaluate their quality using Mobile App Rating Scale–French (MARS-F).

Methods

Selection of the French Mobile Health Apps

Mental health–related apps were screened in the French App Store (iPhone Operating System [iOS]) and in the French Google Play Store (Android) from June 10, 2022, to June 17, 2022, by 2 academic researchers. The following search terms were used: "bien être mental" (mental well-being), "santé mentale" (mental health), and "bien être" (well-being). Each search term was introduced separately in the Apple App Store and Google Play

XSL•FO
RenderX

Store because no truncation or use of logic operators (AND, OR, and NOT) was possible.

Each researcher eliminated duplicate apps by cross-checking the names of apps and developers. Then, both researchers checked that they had the same list of apps and downloaded them. They checked the following inclusion criteria: (1) mainly in the French language, (2) mental well-being as subject matter, (3) targeting adult users, and (4) self-personalized programs. The exclusion criteria were apps (1) focusing on content unrelated to mental health services, such as mental training, yoga, physical activity, and nutrition; (2) targeting people with specific disorders such as suicidal tendencies, eating disorders, or addiction; and (3) providing a single function only.

Selection of Mental Health Professionals

The inclusion criteria for raters were the following: (1) mental health professional and (2) practicing at a hospital or performing private clinical activity in France. The exclusion criteria were (1) not having a mobile phone; (2) not being able to download apps from the Apple or Google stores; (3) never having used a mobile app; and (4) having hearing, visual, or motor disabilities.

Selection of a Standardized Rating Scale for Mobile Apps

MARS-F [28] was used in this study. The first part of this scale, named *App classification*, includes the main characteristics of the app, such as the name, version, developer, focus or target, theoretical background or strategies, age group, and so on. This part was reviewed by the 2 academic researchers. The Mobile App Rating Scale (MARS) scale is composed of a main part (23 items organized into 5 sections, named A, B, C, D, and E) and an additional part (section F with 6 items).

Section A (engagement section; 5 items) determines whether the app is interesting, fun, customizable, and interactive (sends alerts, feedback, reminders, and messages and allows sharing). Section B (functionality section; 4 items) analyzes the app operation, ease of learning, flow logic, navigation, and gestural design of the app. Section C (aesthetics section; 3 items) focuses on the graphic design of the app, color palette, overall visual appeal, and stylistic consistency. Section D (information quality; 7 items) assesses whether the app contains high-quality information (feedback, text, references, and measurements) from a credible source. Section E (subjective section; 4 items) determines the interest of the user in the app. Section F (mobile app specificities; 6 items) analyzes the point of view of mental health professionals regarding the effect of selected apps on knowledge, possible changes in user attitudes and intentions to change, and probability of changing the identified targeted behaviors. In our study, the targeted health behavior was *mental well-being*.

Each item was rated on a 5-point Likert scale (1=strongly disagree to 5=strongly agree). The score for each section was obtained by calculating the mean score of the items. The mean of section scores (sections A, B, C, and D) corresponds to the overall quality MARS score. The mean score of section E corresponds to the subjective quality score, and the mean score of section F evaluates the specificities of the app. The scores range from a minimum of 1 (poor quality) to a maximum of 5 (high quality).

Methodology of Evaluation

Training of the Raters

The raters were 9 mental health professionals (Multimedia Appendix 1). Before rating the mental well-being apps, the raters had to train in the use of MARS-F. For this, they viewed a training video in French (available on request to the corresponding author), which was developed for MARS-F [28] and adapted from the English training video of Stoyanov et al [29]. In this video, each item and answer is explained based on examples. At the end of the video, a training exercise with an app not included in the sample of the study was proposed. The raters downloaded the app, tested it for at least 10 minutes, completed MARS-F, and then compared their results with those in the video. When the individual score for an item differed by >2 points, the raters discussed until a consensus was reached to ensure that they had the same understanding.

Evaluation of the Selected Apps by the Raters

The mental health apps were evaluated by the 9 mental health professionals during the month of July 2022. They downloaded all the included apps, used each app for at least 10 minutes, and immediately evaluated the app using a web-based MARS-F questionnaire.

Statistical Analysis

To assess interrater reliability, the intraclass correlations (ICCs; 2-way random, average measures and absolute concordance) were calculated [30,31]. For each item, each section, and MARS-F quality score (sections A, B, C, and D), the 95% CIs were calculated. On the basis of the 95% CI of the ICC estimate, values <0.5, between 0.5 and 0.75, between 0.75 and 0.9, and >0.90 are indicative of poor, moderate, good, and excellent reliability, respectively [31]. Mean values and SDs were calculated for each item and each mental health app section (presented as mean [SD]). Owing to missing values, item 19 was excluded from all analyses, and the mean for section D was adjusted accordingly.

To assess the differences between the quality of the apps, by item and by section, box plots were generated. The points that appear on either side of the boxes are calculated using 1.5 times the IQR (the distance between the first and third quartiles) to highlight the extreme values.

To provide an overview of the average scores for each item (row) and each app (column), a heat map was constructed. The color gradients represent whether the score is low (near 1 [yellow]) or high (near 5 [green]).

To assess the correlation between average quality and subjective item 23 (“What is your overall star rating of the app?”), the Pearson coefficient (*r*) was calculated. To provide a complete overview of the popularity of each mobile app and the number of reviewers, the number of stars awarded by users in the iOS and Android stores was reported.

Statistical analyses were performed using R, with the *dplyr*, *psych*, and *ggplot2* packages from the R Project for Statistical

XSL•FO
RenderX

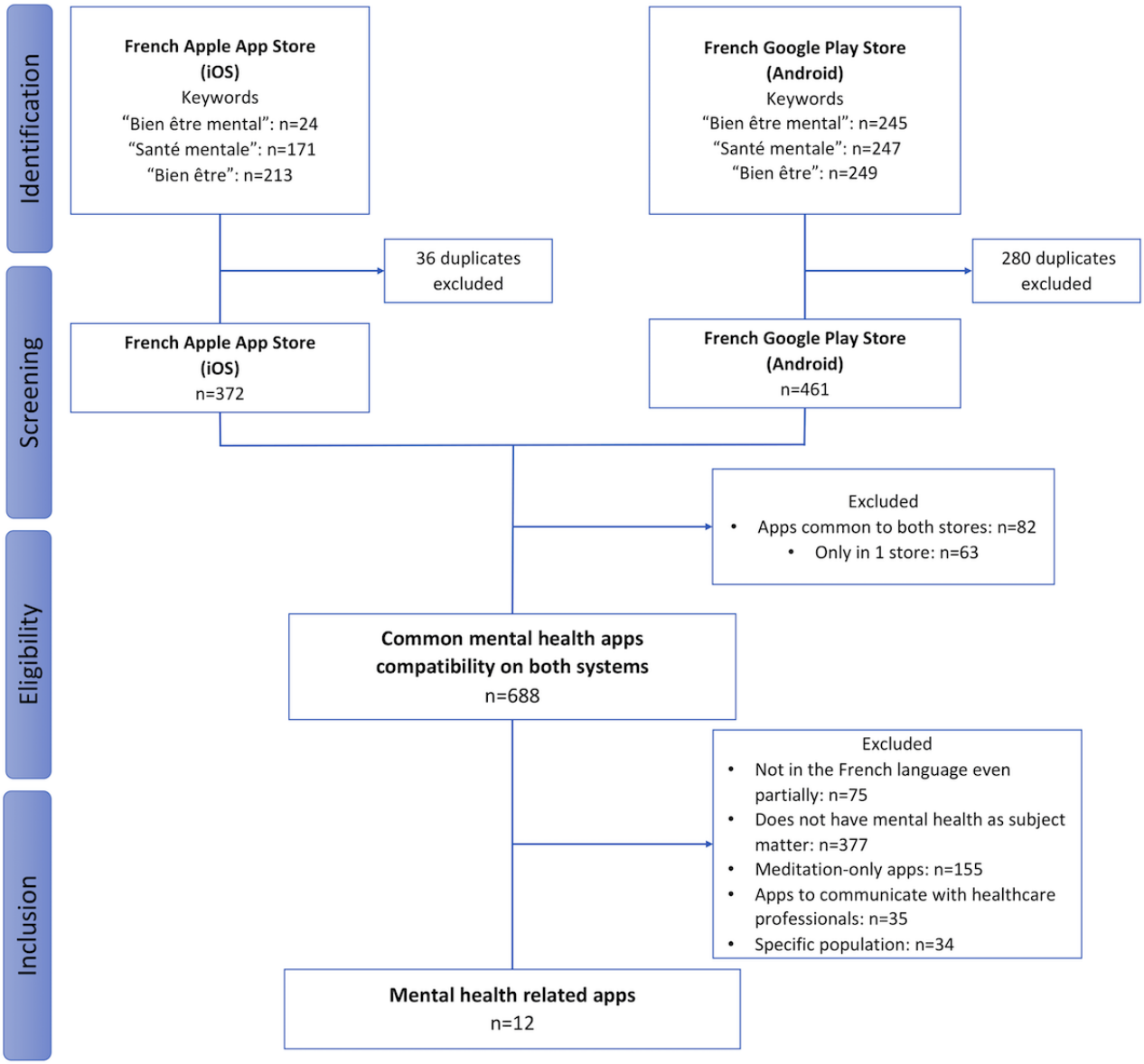
Computing (version 4.1.1; R Foundation for Statistical Computing).

Results

Selection of Mental Health Mobile Apps

The use of keywords allowed the identification of 35.51% (408/1149) of apps from Apple App Store and 64.49% (741/1149) of apps from Google Play Store (Figure 1). Once the 2 lists were cross-checked according to the name of the app and developer, 59.87% (688/1149) of the apps were common to both systems. After a thorough review of the remaining apps' download page and application of the inclusion and exclusion criteria, 1.04% (12/1149) of the apps were finally selected.

Figure 1. Flowchart of mental health mobile apps selection. iOS: iPhone Operating System.



Characteristics of Mental Health Mobile Apps

Descriptive and technical information regarding the mobile mental health apps is provided in Tables S1-S6 in Multimedia Appendix 2. None of these apps had the same developer. An app had a different name depending on the store (*VOS – journal de l'humeur* in the Apple App Store and *VOS – journal intime* in the Google Play Store), whereas all the other apps had the same name. Of the 12 apps, 1 (8%) was fully free of charge, 8 (67%) were free with in-app purchases, and 3 (25%) were accessible exclusively with a company code. The number of downloads depends on the app. According to Google Play Store (as the data were not available for the Apple store), Sanvello was the most downloaded app, followed by *VOS – journal de l'humeur*.

Table 1 presents the characteristics of the mental health mobile apps. All the apps (12/12, 100%) targeted increasing happiness or well-being; mindfulness, meditation, or relaxation; negative emotion reduction; anxiety or stress; anger; behavior change; and goal setting, whereas only 17% (2/12) of the apps targeted alcoholism or substance use, 17% (2/12) focused on

XSL•FO
RenderX

relationships, and another 17% (2/12) aimed at physical health. The theoretical background and strategies of all apps were the following: assessment; information or education; monitoring or tracking; goal setting; advice, tips, strategies, or skills training; cognitive behavioral therapy (CBT)—behavioral (positive events); CBT—cognitive (thought challenging); acceptance commitment therapy; and mindfulness or meditation. Most mental health apps (11/12, 92%) also used monitoring or tracking as strategies. The apps were designed for young adults (12/12, 100%) and adults (12/12, 100%) but not for adolescents and children (aged <12 years). All the mental health apps (12/12, 100%) allowed password protection, required log-in, sent reminders, and needed web access to function, and 67% (10/15) of the apps required web access to function. Of the 12 apps, 6 (50%) allowed sharing (Facebook, Twitter, etc) and 5 (42%) had an app community.

XSL•FO
RenderX

Table 1. Characteristics of the mental health mobile apps (N=12).

Characteristics	Apps, n (%)
Focus—what the app targets[a]	
Increase happiness or well-being	12 (100)
Mindfulness, meditation, or relaxation	12 (100)
Reduce negative emotions	12 (100)
Anxiety or stress	12 (100)
Anger	12 (100)
Behavior change	12 (100)
Alcohol or substance use	2 (17)
Goal setting	12 (100)
Relationships	2 (17)
Physical health	2 (17)
Theoretical background or strategies[a]	
Assessment	12 (100)
Information or education	12 (100)
Monitoring or tracking	11 (92)
Goal setting	12 (100)
Advice, tips, strategies, or skills training	12 (100)
CBT[b]—behavioral (positive events)	12 (100)
CBT—cognitive (thought challenging)	12 (100)
Acceptance commitment therapy	12 (100)
Mindfulness or meditation	12 (100)
Relaxation	12 (100)
Age group[a]	
Children (<12 years)	0 (0)
Adolescents (13-17 years)	0 (0)
Young adults (18-25 years)	12 (100)
Adults (>25 years)	12 (100)
Technical aspects of the app[a]	
Allows sharing (Facebook, Twitter, etc)	6 (50)
Has an app community	5 (42)
Allows password protection	12 (100)
Requires log-in	12 (100)
Sends reminders	12 (100)
Needs web access to function	12 (100)

[a]Participants could choose several answers.

[b]CBT: cognitive behavioral therapy.

Reliability of the Evaluation

The ICC, calculated based on the 12 apps, was considered as good, with 0.79 (95% CI 0.73-0.83) for the A, B, C, and D sections. The ICC per section was good for section A (0.68, 95% CI 0.54-0.79), section B (0.62, 95% CI 0.44-0.76), section C (0.63, 95% CI 0.42-0.78), and section D (0.81, 95% CI 0.73-0.87).

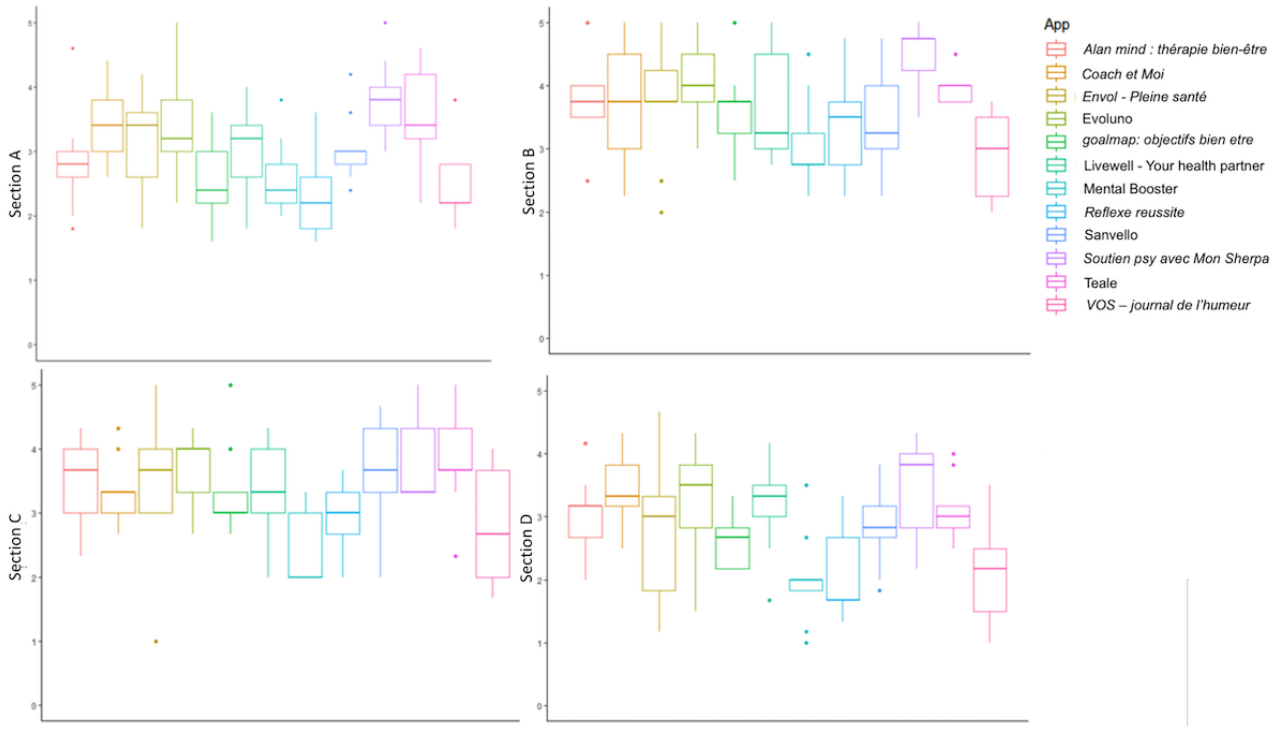
XSL•FO
RenderX

Assessing the Quality of Content of Mental Health Mobile Apps

MARS-F quality score for each section (sections A, B, C, and D) and each app are presented in Figure 2 and Tables S1 and S2 in Multimedia Appendix 3. Mean engagement scores (section A) ranged from 2.33 (SD 0.69) for *Reflexe reussite* to 3.80 (SD 0.61) for *Soutien psy avec Mon Sherpa*. Mean functionality scores (section B) ranged from 2.89 (SD 0.63) for *VOS – journal de l'humeur* to 4.47 (SD 0.53) for *Soutien psy avec Mon Sherpa*. Mean esthetics scores (section C) ranged from 2.52 (SD 0.62) for Mental Booster to 3.89 (SD 0.69) for *Soutien psy avec Mon Sherpa*. Mean information scores (section D) ranged from 2 (SD 0.75) for Mental Booster to 3.46 (SD 0.77) for *Soutien psy avec Mon Sherpa*. For all mental health apps, except Sanvello (11/12, 92%), the mean score for functionality was consistently higher than that for the other sections.

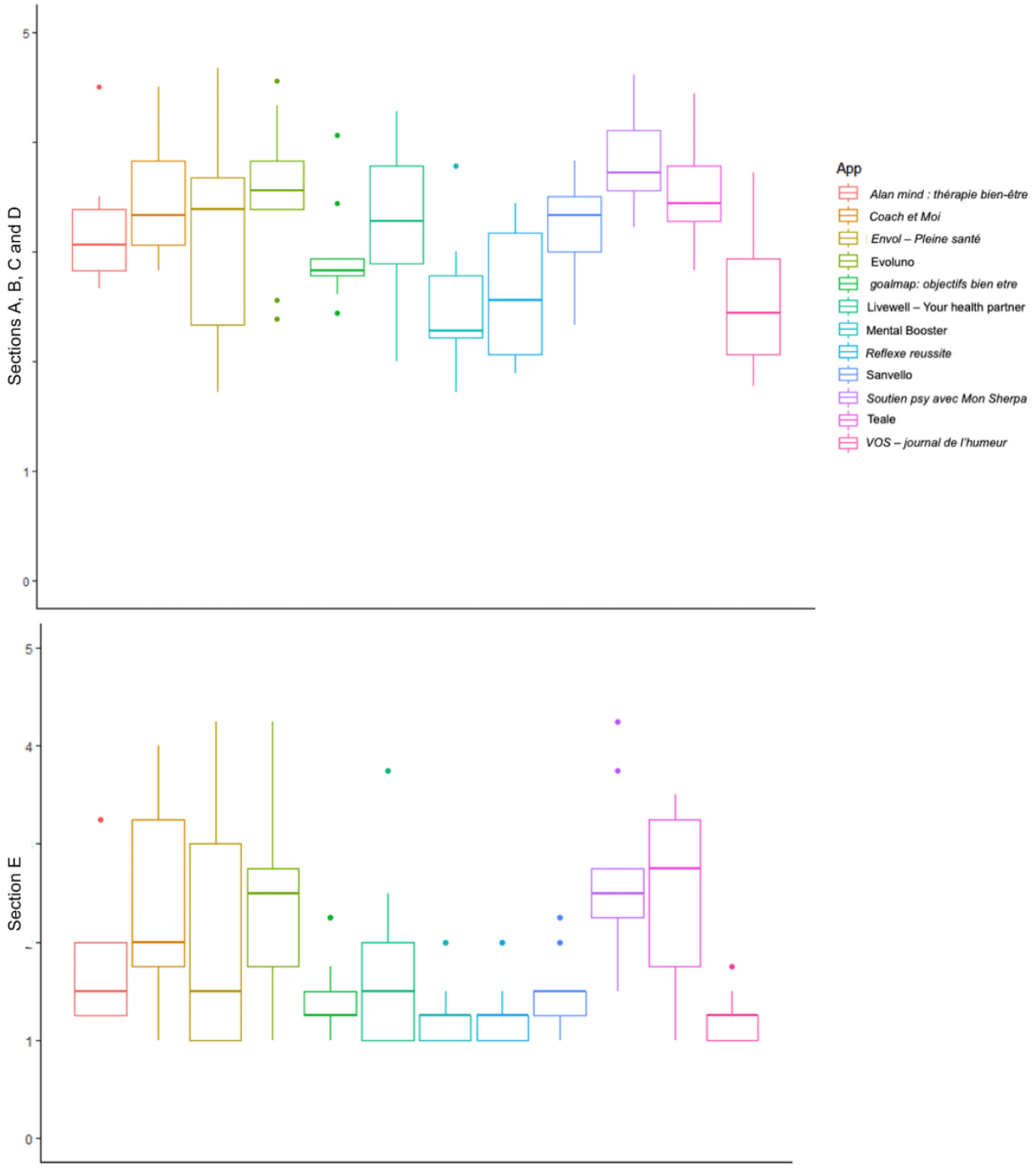
MARS-F quality score (sections A, B, C, and D) and MARS-F subjective quality scores (section E) for each app are presented in Figure 3 and Table S1 in Multimedia Appendix 3. The best MARS-F quality scores were obtained by *Soutien psy avec Mon Sherpa* (mean 3.85, SD 0.48), Evoluno (mean 3.54, SD 0.72), and Teale (mean 3.53, SD 0.47), whereas the worst quality scores were obtained by *Reflexe reussite* (mean 2.59, SD 0.61), *VOS – journal de l'humeur* (mean 2.55, SD 0.71), and Mental Booster (mean 2.49, SD 0.61). Mean MARS-F subjective quality score varied from 1.22 (SD 0.26) for *VOS – journal de l'humeur* to 2.69 (SD 0.84) for *Soutien psy avec Mon Sherpa*. The best subjective quality scores were obtained by *Soutien psy avec Mon Sherpa* (mean 2.69, SD 0.84), Teale (mean 2.53, SD 0.91), and Evoluno (mean 2.42, SD 1), whereas the worst quality scores were obtained by Mental Booster (mean 1.25, SD 0.33), *Reflexe reussite* (mean 1.25, SD 0.33), and *VOS – journal de l'humeur* (mean 1.22, SD 0.26).

Figure 2. Qualitative evaluation of mental health mobile apps—section A: engagement, section B: functionality, section C: esthetics, and section D: information.



XSL•FO
RenderX

Figure 3. Qualitative (sections A, B, C, and D) and subjective qualitative (section E) evaluation of mental health mobile apps.

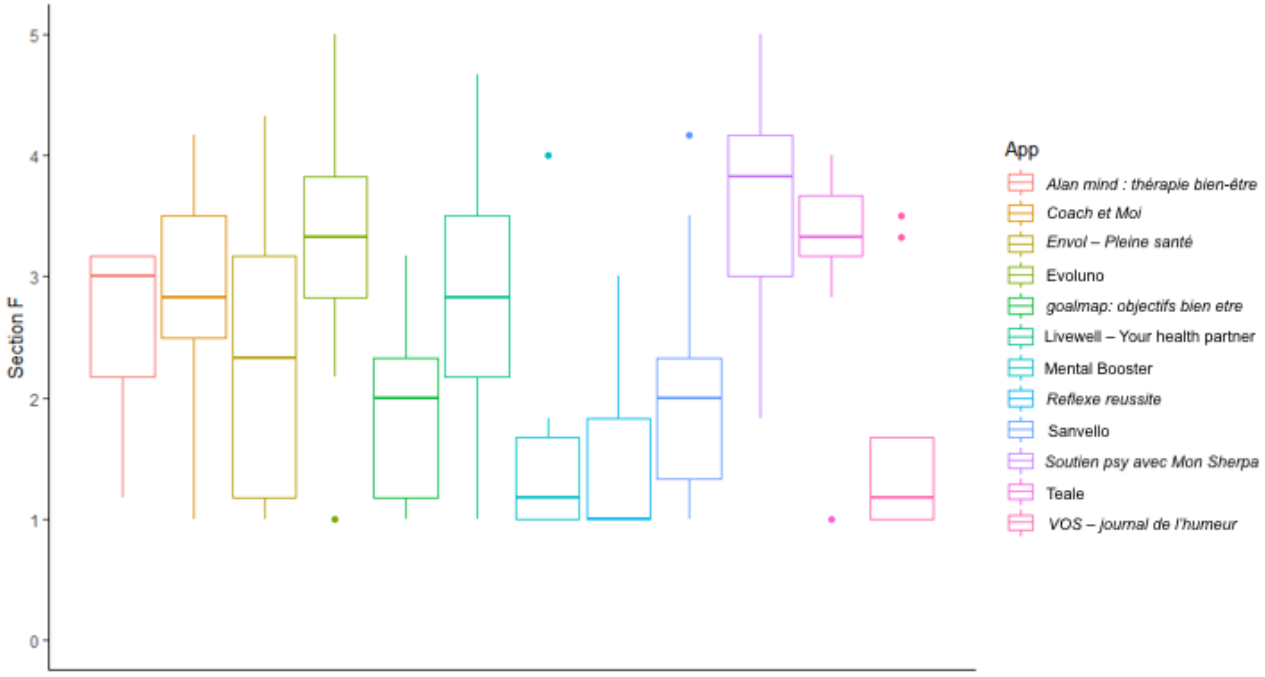


Assessing the Content Specificity of Mental Health Mobile Apps

The evaluation of the specificity (section F) of the mental health apps is summarized in Figure 4 and Table S1 in Multimedia Appendix 3. This score (mean) ranged from 1.56 (SD 0.97) for Mental Booster to 3.31 (SD 1.22) for Evoluno. The best specificity scores were achieved by *Soutien psy avec Mon Sherpa* (mean 2.59, SD 0.92), Teale (mean 3.20, SD 0.90), and Evoluno (mean 3.31, SD 1.22), whereas the worst quality scores were obtained by Mental Booster (mean 1.56, SD 0.97), *Reflexe reussite* (mean 1.57, SD 0.76), and *VOS – journal de l'humeur* (mean 1.65, SD 1.02).

XSL•FO
RenderX

Figure 4. App-specific scores of mental health mobile apps (section F).



Evaluation of the Strengths and Weaknesses of Each App

The app-specific score (section F) was always lower than the subjective quality score (section E), which was always lower than the MARS-F quality score (sections A, B, C, and D), except for *Soutien psy avec Mon Sherpa*. This score was lower than the rating from Apple App Store or Google Play Store, when the apps are rated in the stores (Multimedia Appendix 2).

The graphical comparison of the average scores for each item and each app (Table S2 in Multimedia Appendix 3) is shown in the heat map (Figure 5). The weakness of the quality score (sections A, B, C, and D) was mainly owing to sections A and D, except *for Coach et Moi*, Evoluno, Livewell - Your health partner, and *Soutien psy avec Mon Sherpa*. More particularly, in section D, the worst scores were observed for the credibility of the app (item 18), except for *Coach et Moi*, *Soutien psy avec Mon Sherpa*, and Teale, with mean 2.78 (SD 1.30), mean 2.78 (SD 0.83), and mean 2.56 (SD 0.73), respectively. The subjective quality (section E) was affected by the weak score regarding the price of the app (item 22). For the specificity of the apps (section F), no real weakness was observed because all the apps had similar item scores. It should be noted that Evoluno and *Soutien psy avec Mon Sherpa* obtained the highest scores.

XSL•FO
RenderX

Figure 5. Heat map of the average scores for each item and each app. The colors are related to the scores and range from yellow (1=worst score) to green (5=best score).

	Alan mind : thérapie bien-être	*Coach et Moi*	*Envol – Pleine santé*	Evoluno	*goalmap: objectifs bien etre*	Livewell: Your health partner	Mental Booster	*Reflexe reussite*	Sanvello	*Soutien psy avec Mon sherpa*	Teale	*VOS – journal de l'humeur*
Section A												
Item 1: Entertainment	2.44	3.56	3.33	3.44	2.78	2.56	2.56	2	3	3.78	3.67	2.67
Item 2: Interest	2.78	3.67	3.33	3.33	2.56	3	2.56	1.89	3.33	4.11	3.89	2.56
Item 3: Customisation	2.78	3.11	2.89	3.22	2.44	3.44	2.78	2.89	2.78	3.22	3.22	2.56
Item 4: Interactivity	2.44	3	3	3	2.44	3.33	2.78	2.44	2.78	3.89	2.78	2.56
Item 5: Target group	3.67	3.56	3.44	3.89	2.78	2.89	2.22	2.44	3.33	4	3.89	2.44
Section B												
Item 6: Performance	3.89	3.67	4.22	3.67	3.67	3.67	4.22	3.44	3.89	4.56	3.78	3
Item 7: Ease of use	3.78	3.78	3.56	3.89	3.78	3.22	3.11	2.89	3.33	4.33	4.56	2.67
Item 8: Navigation	3.78	3.56	3.56	4.22	3.56	3.44	3.22	3.22	3.56	4.33	4.22	2.44
Item 9: Gestural design	3.67	4.11	3.78	4.11	3.56	4	3.33	3.78	3.44	4.67	3.56	3.44
Section C												
Item 10: Layout	3.56	3.22	3.44	4	3.67	3.44	2.78	3.22	3.56	4.33	4.44	2.78
Item 11: Graphics	3.44	3.33	3.56	3.78	3.22	3.44	2.44	3	3.89	3.67	3.78	2.78
Item 12: Visual appeal	3.44	3.33	3.44	3.33	3.11	3.11	2.33	2.67	3.33	3.67	3.33	3
Section D												
Item 13: Accuracy	3.22	3.56	3	3.89	2.67	3.22	2.22	2.56	3.22	3.67	2.89	2.78
Item 14: Goals	2.78	2.78	3	2.78	2.67	3.78	1.78	2	2.56	2.78	2.67	1.89
Item 15: Quality of information	3.44	3.56	3.11	3.78	2.89	3.44	2.22	2.67	3	3.56	3.44	2.22
Item 16: Quantity of information	3.22	3.89	2.67	3.67	2.67	3.67	2.67	2.56	3.11	3.89	3.44	2
Item 17: Visual information	3.89	4	3.11	3.44	3.44	3.44	3.11	2	3.56	4.11	3.56	2.56
Item 18: Credibility	1.67	2.78	1.78	2.33	1.56	1.78	1.11	1	1.67	2.78	2.56	1.56
Section E												
Item 20: Recommendations	1.67	2.44	2	2.67	1.44	1.67	1.22	1.33	1.67	3.22	2.56	1.11
Item 21: Usage	1.56	2.56	2.11	2.56	1.33	1.67	1.11	1	1	2.67	2.67	1
Item 22: Price	1.22	1.22	1.44	1.44	1	1.22	1	1	1	1.22	1.67	1
Item 23: Overall rating	2.44	2.89	2.67	3	1.78	2.44	1.67	1.67	2.22	3.67	3.22	1.78
Section F												
Awareness	3.11	2.78	2.22	3.56	1.78	3	1.44	1.56	2.11	3.56	3	1.89
Knowledge	2.89	2.89	1.78	3.33	1.89	2.89	1.44	1.44	2.33	3.56	3	1.33
Attitude	2.33	2.89	2.56	3.33	2	2.78	1.56	1.56	2.22	3.56	3.44	1.67
Intention to change	2.78	2.89	2.44	3.22	2.11	2.44	1.56	1.44	2.11	3.56	3.44	1.56
Help seeking	2.44	3.22	2.67	3.22	2.44	2.89	1.78	1.89	2.22	3.67	3.33	2
Behavior change	2.44	2.67	2.33	3.22	1.78	2.67	1.56	1.56	2	3.67	3	1.44

Correlation Between MARS-F Score and Star Rating

The correlation between the MARS-F quality score (sections A, B, C, and D) and MARS-F overall star rating (item 23: "What is your overall star rating of the app?") was considered as good (r=0.86; P<.001). The MARS-F quality score was higher than the MARS-F overall star rating for all the mental health apps tested (12/12, 100%). Correlation analysis between the MARS-F quality score of the apps and their respective star rating in the mobile app stores was limited by the availability of star rating in the stores and discrepancies among the number of raters. The store ratings were high compared with the MARS-F quality score. These user ratings fluctuate from 2.7 (*VOS – journal de l'humeur*) to 4.8 (*Alan Mind: thérapie bien-être*) for the iOS store and from 3.2 (*Coach et Moi*) to 4.6 (*Envol - Pleine santé*) for the Android store (Table 2).

XSL•FO
RenderX

Table 2. MARS-F[a] overall star rating, overall quality MARS-F score, star rating in the Apple App Store, and star rating in the Google Play Store.

Apps	MARS-F overall star rating (item 23), mean (SD)	MARS-F quality score (sections A, B, C, and D), mean (SD)	Apple App Store, star rating; number of raters	Google Play Store, star rating; number of raters
Alan Mind: thérapie bien-être	2.44 (0.53)	3.22 (0.56)	4.8; 3217	4.2; 125
Coach et Moi	2.89 (1.27)	3.47 (0.55)	4.6; 28	3.2; 32
Envol - Pleine santé	2.67 (1.41)	3.23 (0.97)	4.8; 30	4.6; 156
Evoluno	3.00 (1.22)	3.54 (0.72)	N/A[b]	N/A
goalmap: objectifs bien etre	1.78 (0.44)	2.97 (0.49)	4.2; 269	3.6; 2000
Livewell - Your health partner	2.44 (1.13)	3.28 (0.73)	N/A	N/A
Mental Booster	1.67 (0.71)	2.49 (0.61)	3.1; 11	N/A
Reflexe reussite	1.67 (0.71)	2.59 (0.61)	4.4; 10	4; 28
Sanvello	2.22 (1.09)	3.18 (0.54)	4.3; 339	4.2; 23,000
Soutien psy avec Mon Sherpa	3.67 (1.00)	3.85 (0.48)	4.6; 1409	4.3; 1000
Teale	3.22 (1.09)	3.53 (0.47)	4.2; 17	4.3; 14
VOS – journal de l'humeur	1.78 (0.83)	2.55 (0.71)	2.7; 14	4.4; 16,000

[a]MARS-F: Mobile App Rating Scale–French.

[b]N/A: not applicable.

Discussion

Principal Findings

Mental health apps aim to improve mental health and wellness, whether in terms of curing a mental illness or implementing habits to improve emotional health [32]. A public survey showed that 76% of the respondents would like to have access to free apps to self-manage and self-monitor their mental health [33]. Thus, mental health apps can play a critical role in mental health care [34]; making mental health care accessible; and reducing barriers to help-seeking, as few people with mood or anxiety problems seek professional help. In addition, geographic, financial, or social constraints may make it difficult to access support [21].

Mental health apps cover all stages of clinical care delivery, including prevention, diagnosis, primary treatment [35], and so on. Our study focused on mental health well-being apps that were consistent with health promotion. Thus, we included apps that enabled users' mental health self-management and that did or did not offer appointments with mental health professionals. We excluded apps whose only purpose was to schedule appointments with mental health professionals, because their focus was more treatment-oriented. Screening of mental health apps available in the French iOS and Android stores led to the inclusion of 12 apps. In Singapore, screening of iTunes and Google Play Store led to inclusion of 44 mental health apps [36]. Compared with our study, they used more keywords, and several of them were related to COVID-19. In China, the screening permitted to select 40 apps, but the authors screened the apps in Apple App Store (for iOS apps), Tencent My App, and Huawei App Gallery (for Android apps), and they selected apps in the Chinese language [37]. In our 2 previous studies using the same methodology but focusing on the risk factors of noncommunicable diseases, 15 apps related to nutrition [38] and 9 apps related to oral hygiene [39] were included.

All the included apps aimed to increase happiness, well-being, mindfulness, meditation, and relaxation and reduce negative emotions, anxiety, stress, and anger. Some apps additionally targeted alcohol or substance use (Livewell - Your health partner and Teale), relationships (Livewell - Your health partner and *VOS – journal de l'humeur*), or physical activity (*goalmap: objectifs bien etre* and Livewell - Your health partner). These mental health apps mainly used CBT, acceptance commitment therapy, and mindfulness or mediation to reach their goals. The use of CBT in mental health apps was one of the recommendations of Bakker et al [21]. CBT is a form of collaborative, individualized psychological treatment that is recognized as the most sustained approach for generating behavioral, cognitive, and emotional adaptation to a wide range of common psychological problems [40]. Several studies have concluded that CBT is an effective treatment for a wide range of psychological disorders, which can be administered via a mobile device and still retain its therapeutic validity [21]. Moreover, CBT can also prevent psychological problems from becoming worse [41-43].

The MARS-F quality score was >2.5, for all the mental health apps tested, except for Mental Booster (11/12, 92%; mean 2.49, SD 0.61). Mean score ranged from 2.49 (SD 0.61) to 3.85 (SD 0.48). Similar scores were observed for apps related to nutrition available in France (mean 3.34, SD 0.39 to mean 3.84, SD 0.32) [38] and oral hygiene (mean 1.80, SD 0.79 to mean 3.4, SD 0.97) [39]. *Soutien psy avec Mon Sherpa*, which obtained the highest score (mean 3.85, SD 0.48), had the particularity of offering real-time interactions with an artificial

XSL•FO
RenderX

intelligence–enabled chatbot to collect and record daily mood. Thus, the user can develop a relationship with the chatbot, similar to the one observed between a therapist and their patient, which involves collaborative empiricism [44]. The 2 apps (Evoluno: mean 3.54, SD 0.72 and Teale: mean 3.53, SD 0.47) that received the highest MARS-F scores had a different concept. They provided some information to the user (resilience, stress, anxiety, negative thoughts, etc) in the form of a podcast (Evoluno) or videos (Teale). Evoluno proposed to evaluate the user (sleep quality, risk of burnout, depression, etc) and suggested breathing exercises. Teale proposed to set goals for the user (self-control, self-esteem, self-fulfillment, etc) with various activities (meditation, breathing, writing, etc) to reach them. Moreover, these 2 apps permitted to contact mental health professionals.

The functionality (section B) was the strength for all the apps tested, except for Sanvello (11/12, 92%). Our previous studies demonstrated that it was also the strength of French nutrition–related and oral hygiene–related apps [38,39]. Information quality (section D) was the weakness of all the apps, except 4 apps (8/12, 67%; *Alan Mind: thérapie bien-être*, *Coach et Moi*, *goalmap: objectifs bien etre*, and Livewell - Your health partner). The weak point of Coach et Moi was the aesthetics (section C), which was not observed for any of the French apps related to nutrition or oral hygiene included in our previous studies [38,39]. The worst score in section D was observed for the credibility of the app because the mental health professionals identified the source of information, but they evaluated that its validity or reliability was questionable (eg, commercial enterprise with vested interest). In addition, as the level of scientific evidence is difficult to assess, mental health professionals chose "N/A The application has not been tested" in most cases to answer item 19, which therefore could not be included in the statistical analysis. To our knowledge, of the 12 apps, only 1 (8%) app included in this study (Sanvello) was indexed in PubMed [45-48]. Myers et al [45] evaluated apps for depression self-management, Mehdi et al [46] evaluated apps for Tinnitus, and Lau et al [47] evaluated apps for mental health. In these 3 studies, the observed MARS scores for Sanvello were 4.6, 4.6, and 4.28, respectively, which were higher than those observed in our study (mean 3.18, SD 0.54). This difference could be explained by the fact that, in these studies, the number of raters was 2 [45,47] or 4 [46] at most.

The MARS-F quality score (sections A, B, C, and D) was higher than the specificity of the app (section F), which was higher than the subjective quality (section E), except for *Soutien psy avec Mon Sherpa*. For this app, the specificity of app (section F) obtained the low score. The subjective quality score (section E) was >2.5 only for *Soutien psy avec Mon Sherpa* (mean 2.69, SD 0.84) and Teale (mean 2.53, SD 0.91). For all the apps tested (12/12, 100%), the main problem was the cost, because mental health professionals declared that they will not pay for this app. The specificity of the apps (section F) assessed the perceived impact of the app on the user's knowledge, attitudes, intent to change, and likelihood of actual change regarding mental health. Only 50% (6/12) of the mental health apps included obtained a score >2.5 for this section. Evoluno (mean 3.31, SD 1.22), Teale (mean 3.20, SD 0.90), and *Coach et Moi* (mean 2.89, SD 0.95) obtained the best scores, followed by Livewell - Your health partner (mean 2.78, SD 1.12), *Alan Mind: thérapie bien-être* (mean 2.67, SD 0.72), and *Soutien psy avec Mon Sherpa* (mean 2.59, SD 0.92). These results can be explained because professionals preferred apps containing information that promotes mental health and well-being than a chatbot.

In contrast to previous studies on French nutrition–related [38] and oral hygiene–related [39] apps, the star rating score from the iOS and Android stores was always lower than the subjective quality score (section E). This difference may be explained by the fact that, when the mental health apps were evaluated in the stores, they were evaluated by a very small number of evaluators, which is in contrast to the nutrition and oral hygiene apps, which were evaluated by a large number of evaluators. The difference between the star rating score and subjective quality score was owing to the fact that the evaluations in the mobile app stores are made by all the users, whereas the evaluation in this study was made by mental health professionals using the MARS-F. Thus, it could be interesting to evaluate the same apps using the user version of the MARS (uMARS) that is designed for users to compare the results.

Limitations

This study has several limitations. First, only mental health apps available on both French Apple App Store and Google Play Store were included, even though several other stores could be investigated (Huawei store, BlackBerry World, Windows Phone Store, or Samsung store). Second, regarding the protection of personal data, no criteria have been defined, except for password protection to log in to the apps. This criterion is not sufficient to prove that the user's personal data are protected.

Third, the commercial use of mobile mental health apps was not considered as a key criterion for assessing the acceptability of the app to users when many mobile apps are developed by anonymous companies but funded by large groups such as pharmaceutical companies. Fourth, MARS was selected for this investigation because it is the most commonly used scale in scientific literature for the assessment of mental health apps [38,39,47,49-59], and its use has been recommended by the French High Authority of Health [60]. Nevertheless, other scales, such as ENLIGHT, which also assesses mobile health apps, could have been used [61]. Correlation and good parallel validity between the MARS and ENLIGHT have been shown [62]. Fifth, our study is only applicable to the French-speaking community, where these mobile apps are available. Sixth, the evaluation was conducted by mental health professionals, whereas the mobile health apps were developed for the general population. Therefore, it will be interesting to have these apps evaluated by users, using the uMARS scale [63] that was developed for them, and then compare the results.

Perspectives

The COVID-19 pandemic has changed the use of digital technology in the field of mental health and, more particularly, by mental health professionals. The pandemic has made it possible to revisit the ways in which the most fragile people are cared for with connected tools and to move toward an evolution of practices [64]. Mental health apps can improve accessibility

XSL•FO
RenderX

to mental health services and be beneficial for mental health [32]. They have positive effects on depressive symptoms, anxiety, and other mental health problems [65] and help overcome some of the barriers that make it difficult to access mental health services (stigma, cost, capacity, and geography) [66]. Thus, this study can help mental health professionals and mobile app users to choose the best mental health apps available in France, in terms of quality.

Health care professionals generally have a different perspective than users. Thus, it may be interesting to have the mental health apps studied in this study evaluated by users via the uMARS scale and then compared.

This study qualitatively assessed mental health–related apps but did not focus on content analysis of the included apps. A review of the literature specific to apps supporting mental health and wellness concluded that apps for bipolar disorder were cost-effective and convenient but that most of them did not provide information on all the core principles of psychoeducation and did not adhere to best practice guidelines [67]. Similarly, Larsen et al [68] showed that many suicide prevention apps were not evidence-based and determined that some apps may be more harmful than helpful. They found that some apps included potentially extremely harmful content describing or supporting access to lethal means, encouraging people to end their lives, and presenting suicide in a trendy manner [66]. Moreover, as many apps are developed for commercial purposes, ethical questions arise regarding the use of personal data. Therefore, additional studies analyzing the content of the apps are needed.

Furthermore, the implementation of randomized clinical trials or longitudinal studies including the 12 mental health apps selected in this study should allow for the analysis of the change in mental health status after the use of the apps.

Conclusions

Mental health professionals assessed that, despite the lack of scientific evidence, the mental health mobile apps available on the French Apple App Store and Google Play Store were of good quality. However, they are reluctant to use them in their professional practice. Additional investigations are needed to assess their compliance with recommendations and their long-term impact on users.

Acknowledgments

The authors would like to thank the mental health professionals for their availability and engagement. The authors would also like to thank Union Prévention Santé pour la Fonction publique, a French player in preventive health care, as manager of the public service's compulsory health insurance scheme, because the subject of this study appeared during a reflection on the mental health support for civil servants. A recent ministerial circular (NOR: TFPF2133602C; February 23, 2022) calls for the development of a culture of prevention of situations of psychological distress and promotion of mental health, particularly in professional environments and within public jobs.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Characteristics of the raters and the hardware and software used.
[DOCX File , 18 KB - mhealth_v10i10e41282_app1.docx]

Multimedia Appendix 2
Descriptive and technical information about the mental health mobile apps.
[DOCX File , 40 KB - mhealth_v10i10e41282_app2.docx]

Multimedia Appendix 3
Mobile App Rating Scale scoring.
[DOCX File , 47 KB - mhealth_v10i10e41282_app3.docx]

References

1. Dattani S, Ritchie H, Roser M. Mental health. Our World in Data. 2018 Apr. URL: https://ourworldindata.org/mental-health [accessed 2022-07-15]
2. Nochaiwong S, Ruengorn C, Thavorn K, Hutton B, Awiphan R, Phosuya C, et al. Global prevalence of mental health issues among the general population during the coronavirus disease-2019 pandemic: a systematic review and meta-analysis. Sci Rep 2021 May 13;11(1):10173 [FREE Full text] [doi: 10.1038/s41598-021-89700-8] [Medline: 33986414]
3. CIM-11 pour les statistiques de mortalité et de morbidité (CIM-11 SMM). ICD. URL: https://icd.who.int/browse11/l-m/fr [accessed 2022-07-15]

XSL•FO
RenderX

4. Yeo ZZ, Suárez L. Validation of the mental health continuum-short form: the bifactor model of emotional, social, and psychological well-being. PLoS One 2022;17(5):e0268232 [FREE Full text] [doi: 10.1371/journal.pone.0268232] [Medline: 35584145]
5. McAneney H, Tully MA, Hunter RF, Kouvonen A, Veal P, Stevenson M, et al. Individual factors and perceived community characteristics in relation to mental health and mental well-being. BMC Public Health 2015 Dec 12;15:1237 [FREE Full text] [doi: 10.1186/s12889-015-2590-8] [Medline: 26655388]
6. Moksnes UK, Bjørnsen HN, Ringdal R, Eilertsen MB, Espnes GA. Association between loneliness, self-esteem and outcome of life satisfaction in Norwegian adolescents aged 15-21. Scand J Public Health 2022 Mar 31:14034948221081287 [FREE Full text] [doi: 10.1177/14034948221081287] [Medline: 35360987]
7. Santini ZI, Pisinger VS, Nielsen L, Madsen KR, Nelausen MK, Koyanagi A, et al. Social disconnectedness, loneliness, and mental health among adolescents in danish high schools: a nationwide cross-sectional study. Front Behav Neurosci 2021;15:632906 [FREE Full text] [doi: 10.3389/fnbeh.2021.632906] [Medline: 33927599]
8. Reijneveld S. Mental health as a public health issue. Eur J Public Health 2015 Apr 1;15(2):111. [doi: 10.1093/eurpub/cki088]
9. Munir S, Takov V. Generalized anxiety disorder. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2022.
10. Depression and other common mental disorders: global health estimates. World Health Organization. 2017. URL: https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf [accessed 2022-07-15]
11. GBD 2015 DiseaseInjury IncidencePrevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 2016 Oct 08;388(10053):1545-1602 [FREE Full text] [doi: 10.1016/S0140-6736(16)31678-6] [Medline: 27733282]
12. Ramiz L, Contrand B, Rojas Castro MY, Dupuy M, Lu L, Sztal-Kutas C, et al. A longitudinal study of mental health before and during COVID-19 lockdown in the French population. Global Health 2021 Mar 22;17(1):29 [FREE Full text] [doi: 10.1186/s12992-021-00682-8] [Medline: 33752717]
13. Eurofound, European Agency for Safety and Health at Work, Psychosocial risks in Europe : prevalence and strategies for prevention : a joint report from the European Foundation for the Improvement of Living and Working Conditions and the European Agency for Safety and Health at Work. Publications Office of the European Union. 2014. URL: https://op.europa.eu/en/publication-detail/-/publication/0567add6-db0e-4065-b13d-7e2599090851/language-en [accessed 2022-07-15]
14. Melchior M, Caspi A, Milne BJ, Danese A, Poulton R, Moffitt TE. Work stress precipitates depression and anxiety in young, working women and men. Psychol Med 2007 Aug;37(8):1119-1129 [FREE Full text] [doi: 10.1017/S0033291707000414] [Medline: 17407618]
15. Veldman K, Reijneveld SA, Verhulst FC, Ortiz JA, Bültmann U. A life course perspective on mental health problems, employment, and work outcomes. Scand J Work Environ Health 2017 Jul 01;43(4):316-325 [FREE Full text] [doi: 10.5271/sjweh.3651] [Medline: 28570743]
16. Mashaphu S, Talatala M, Seape S, Eriksson L, Chiliza B. Mental health, culture and resilience-approaching the COVID-19 pandemic from a South African perspective. Front Psychiatry 2021;12:611108 [FREE Full text] [doi: 10.3389/fpsyt.2021.611108] [Medline: 34305663]
17. Maric V, Mishra J, Ramanathan DS. Using mind-body medicine to reduce the long-term health impacts of COVID-specific chronic stress. Front Psychiatry 2021;12:585952 [FREE Full text] [doi: 10.3389/fpsyt.2021.585952] [Medline: 33692706]
18. Kola L, Kohrt BA, Hanlon C, Naslund JA, Sikander S, Balaji M, et al. COVID-19 mental health impact and responses in low-income and middle-income countries: reimagining global mental health. Lancet Psychiatry 2021 Feb 24;8(6):535-550. [doi: 10.1016/S2215-0366(21)00025-0] [Medline: 33639109]
19. Esponda GM, Hartman S, Qureshi O, Sadler E, Cohen A, Kakuma R. Barriers and facilitators of mental health programmes in primary care in low-income and middle-income countries. Lancet Psychiatry 2020 Jan;7(1):78-92. [doi: 10.1016/S2215-0366(19)30125-7] [Medline: 31474568]
20. Global governance toolkit for digital mental health. World Economic Forum. 2021 Apr 6. URL: https://www.weforum.org/whitepapers/global-governance-toolkit-for-digital-mental-health/ [accessed 2022-07-15]
21. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental health smartphone apps: review and evidence-based recommendations for future developments. JMIR Ment Health 2016 Mar 01;3(1):e7 [FREE Full text] [doi: 10.2196/mental.4984] [Medline: 26932350]
22. Auxier B, Westcott K, Bucaille A. Mental health goes mobile: The mental health app market will keep on growing. Deloitte Insights. 2021 Dec 1. URL: https://www2.deloitte.com/xe/en/insights/industry/technology/technology-media-and-telecom-predictions/2022/mental-health-app-market.html [accessed 2022-07-15]
23. Cangiano F, Parker S. Proactivity for mental health and well-being. In: The Wiley Blackwell Handbook of the Psychology of Occupational Safety and Workplace Health. Hoboken, New Jersey, United States: Wiley Blackwell; 2016.
24. Eisenstadt M, Liverpool S, Infanti E, Ciuvat RM, Carlsson C. Mobile apps that promote emotion regulation, positive mental health, and well-being in the general population: systematic review and meta-analysis. JMIR Ment Health 2021 Nov 08;8(11):e31170 [FREE Full text] [doi: 10.2196/31170] [Medline: 34747713]

XSL•FO
RenderX

25. Hwang WJ, Ha JS, Kim MJ. Research trends on mobile mental health application for general population: a scoping review. Int J Environ Res Public Health 2021 Mar 02;18(5):2459 [FREE Full text] [doi: 10.3390/ijerph18052459] [Medline: 33801537]
26. Gál É, tefan S, Cristea IA. The efficacy of mindfulness meditation apps in enhancing users' well-being and mental health related outcomes: a meta-analysis of randomized controlled trials. J Affect Disord 2021 Jan 15;279:131-142. [doi: 10.1016/j.jad.2020.09.134] [Medline: 33049431]
27. Eis S, Solà-Morales O, Duarte-Díaz A, Vidal-Alaball J, Perestelo-Pérez L, Robles N, et al. Mobile applications in mood disorders and mental health: systematic search in apple app store and google play store and review of the literature. Int J Environ Res Public Health 2022 Feb 15;19(4):2186 [FREE Full text] [doi: 10.3390/ijerph19042186] [Medline: 35206373]
28. Saliasi I, Martinon P, Darlington E, Smentek C, Tardivo D, Bourgeois D, et al. Promoting health via mHealth applications using a French version of the mobile app rating scale: adaptation and validation study. JMIR Mhealth Uhealth 2021 Aug 31;9(8):e30480 [FREE Full text] [doi: 10.2196/30480] [Medline: 34463623]
29. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015 Mar 11;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
30. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979 Mar;86(2):420-428. [doi: 10.1037//0033-2909.86.2.420] [Medline: 18839484]
31. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropractic Med 2016 Jun;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
32. Donker T, Petrie K, Proudfoot J, Clarke J, Birch M, Christensen H. Smartphones for smarter delivery of mental health programs: a systematic review. J Med Internet Res 2013 Nov 15;15(11):e247 [FREE Full text] [doi: 10.2196/jmir.2791] [Medline: 24240579]
33. Proudfoot J, Parker G, Hadzi Pavlovic D, Manicavasagar V, Adler E, Whitton A. Community attitudes to the appropriation of mobile phones for monitoring and managing depression, anxiety, and stress. J Med Internet Res 2010 Dec 19;12(5):e64 [FREE Full text] [doi: 10.2196/jmir.1475] [Medline: 21169174]
34. Simon GE, Ludman EJ. It's time for disruptive innovation in psychotherapy. Lancet 2009 Aug 22;374(9690):594-595. [doi: 10.1016/S0140-6736(09)61415-X] [Medline: 19699995]
35. Chandrashekar P. Do mental health mobile apps work: evidence and recommendations for designing high-efficacy mental health mobile apps. Mhealth 2018;4:6 [FREE Full text] [doi: 10.21037/mhealth.2018.03.02] [Medline: 29682510]
36. Li LS, Wong LL, Yap KY. Quality evaluation of stress, anxiety and depression apps for COVID-19. J Affect Disord Rep 2021 Dec;6:100255 [FREE Full text] [doi: 10.1016/j.jadr.2021.100255] [Medline: 34704086]
37. Wu X, Xu L, Li P, Tang T, Huang C. Multipurpose mobile apps for mental health in Chinese app stores: content analysis and quality evaluation. JMIR Mhealth Uhealth 2022 Jan 04;10(1):e34054 [FREE Full text] [doi: 10.2196/34054] [Medline: 34982717]
38. Martinon P, Saliasi I, Bourgeois D, Smentek C, Dussart C, Fraticelli L, et al. Nutrition-related mobile apps in the French app stores: assessment of functionality and quality. JMIR Mhealth Uhealth 2022 Mar 14;10(3):e35879 [FREE Full text] [doi: 10.2196/35879] [Medline: 35285817]
39. Carrouel F, Bourgeois D, Clément C, Tardivo D, Martinon P, Guiral S, et al. Oral-hygiene-related mobile apps in the French app stores: assessment of functionality and quality. Int J Environ Res Public Health 2022 Jun 14;19(12):7293 [FREE Full text] [doi: 10.3390/ijerph19127293] [Medline: 35742541]
40. Payne L, Ellard K, Fairholme C, Barlow D. Emotional disorders: a unified transdiagnostic protocol. In: Clinical Handbook of Psychological Disorders, Fourth Edition: A Step-by-Step Treatment Manual. New York, United States: The Guilford Press; 2008.
41. Bakker G. Practical CBT Using Functional Analysis and Standardised Homework in Everyday Therapy. Queensland, Australia: Australian Academic Press; 2008.
42. Moorey S. The six cycles maintenance model: growing a "vicious flower" for depression. Behav Cogn Psychother 2010 Mar;38(2):173-184. [doi: 10.1017/S1352465809990580] [Medline: 19925693]
43. Gyani A, Shafran R, Layard R, Clark DM. Enhancing recovery rates: lessons from year one of IAPT. Behav Res Ther 2013 Sep;51(9):597-606 [FREE Full text] [doi: 10.1016/j.brat.2013.06.004] [Medline: 23872702]
44. Person J. The Case Formulation Approach to Cognitive-Behavior Therapy. New York: The Guilford Press; 2008.
45. Myers A, Chesebrough L, Hu R, Turchioe MR, Pathak J, Creber RM. Evaluating commercially available mobile apps for depression self-management. AMIA Annu Symp Proc 2020;2020:906-914 [FREE Full text] [Medline: 33936466]
46. Mehdi M, Stach M, Riha C, Neff P, Dode A, Pryss R, et al. Smartphone and mobile health apps for tinnitus: systematic identification, analysis, and assessment. JMIR Mhealth Uhealth 2020 Aug 18;8(8):e21767 [FREE Full text] [doi: 10.2196/21767] [Medline: 32808939]
47. Lau N, O'Daffer A, Yi-Frazier JP, Rosenberg AR. Popular evidence-based commercial mental health apps: analysis of engagement, functionality, aesthetics, and information quality. JMIR Mhealth Uhealth 2021 Jul 14;9(7):e29689 [FREE Full text] [doi: 10.2196/29689] [Medline: 34259639]

XSL•FO
RenderX

48. Leong QY, Sridhar S, Blasiak A, Tadeo X, Yeo G, Remus A, et al. Characteristics of mobile health platforms for depression and anxiety: content analysis through a systematic review of the literature and systematic search of two app stores. J Med Internet Res 2022 Feb 04;24(2):e27388 [FREE Full text] [doi: 10.2196/27388] [Medline: 35119370]
49. Talwar D, Yeh Y, Chen W, Chen L. Characteristics and quality of genetics and genomics mobile apps: a systematic review. Eur J Hum Genet 2019 Jun;27(6):833-840 [FREE Full text] [doi: 10.1038/s41431-019-0360-2] [Medline: 30809045]
50. Chen R, Santo K, Wong G, Sohn W, Spallek H, Chow C, et al. Mobile apps for dental caries prevention: systematic search and quality evaluation. JMIR Mhealth Uhealth 2021 Jan 13;9(1):e19958 [FREE Full text] [doi: 10.2196/19958] [Medline: 33439141]
51. Ciapponi A, Donato M, Gülmezoglu AM, Alconada T, Bardach A. Mobile apps for detecting falsified and substandard drugs: a systematic review. PLoS One 2021;16(2):e0246061 [FREE Full text] [doi: 10.1371/journal.pone.0246061] [Medline: 33539433]
52. Mandracchia F, Llauradó E, Tarro L, Valls RM, Solà R. Mobile phone apps for food allergies or intolerances in app stores: systematic search and quality assessment using the mobile app rating scale (MARS). JMIR Mhealth Uhealth 2020 Sep 16;8(9):e18339 [FREE Full text] [doi: 10.2196/18339] [Medline: 32936078]
53. Romero RL, Kates F, Hart M, Ojeda A, Meirom I, Hardy S. Quality of deaf and hard-of-hearing mobile apps: evaluation using the mobile app rating scale (MARS) with additional criteria from a content expert. JMIR Mhealth Uhealth 2019 Oct 30;7(10):e14198 [FREE Full text] [doi: 10.2196/14198] [Medline: 31670695]
54. Escriche-Escuder A, De-Torres I, Roldán-Jiménez C, Martín-Martín J, Muro-Culebras A, González-Sánchez M, et al. Assessment of the quality of mobile applications (Apps) for management of low back pain using the mobile app rating scale (MARS). Int J Environ Res Public Health 2020 Dec 09;17(24):9209 [FREE Full text] [doi: 10.3390/ijerph17249209] [Medline: 33317134]
55. Salehinejad S, Niakan Kalhori SR, Hajesmaeel Gohari S, Bahaadinbeigy K, Fatehi F. A review and content analysis of national apps for COVID-19 management using Mobile Application Rating Scale (MARS). Inform Health Soc Care 2021 Mar 02;46(1):42-55. [doi: 10.1080/17538157.2020.1837838] [Medline: 33164594]
56. Kalhori SR, Hemmat M, Noori T, Heydarian S, Katigari MR. Quality evaluation of english mobile applications for gestational diabetes: app review using mobile application rating scale (MARS). Curr Diabetes Rev 2021;17(2):161-168. [doi: 10.2174/1573399816666200703181438] [Medline: 32619173]
57. Davalbhakta S, Advani S, Kumar S, Agarwal V, Bhoyar S, Fedirko E, et al. A systematic review of the smartphone applications available for coronavirus disease 2019 (COVID19) and their assessment using the mobile app rating scale (MARS). medRxiv 2020 Jul 04 (forthcoming) [FREE Full text] [doi: 10.1101/2020.07.02.20144964] [Medline: 32637969]
58. Siddiqui NR, Hodges SJ, Sharif MO. Orthodontic apps: an assessment of quality (using the Mobile App Rating Scale (MARS)) and behaviour change techniques (BCTs). Prog Orthod 2021 Sep 13;22(1):25 [FREE Full text] [doi: 10.1186/s40510-021-00373-5] [Medline: 34514529]
59. Martín-Martín J, Muro-Culebras A, Roldán-Jiménez C, Escriche-Escuder A, De-Torres I, González-Sánchez M, et al. Evaluation of android and apple store depression applications based on mobile application rating scale. Int J Environ Res Public Health 2021 Nov 27;18(23):12505 [FREE Full text] [doi: 10.3390/ijerph182312505] [Medline: 34886232]
60. Évaluation des Applications dans le champ de la santé mobile (mHealth) - État des lieux et critères de qualité du contenu médical pour le référencement des services numériques dans l'espace numérique de santé et le bouquet de services des professionnels. Haute Autorité de la Santé. 2021 Jun 24. URL: https://www.has-sante.fr/upload/docs/application/pdf/2021- [accessed 2022-09-22]
61. Baumel A, Faber K, Mathur N, Kane JM, Muench F. Enlight: a comprehensive quality and therapeutic potential evaluation tool for mobile and web-based eHealth interventions. J Med Internet Res 2017 Mar 21;19(3):e82 [FREE Full text] [doi: 10.2196/jmir.7270] [Medline: 28325712]
62. Naydenova G. An overview of existing methods for evaluating mobile applications to change user behavior. Int Scientific J Scientific Technical Union Mechanical Eng 2022;7(1):21-25.
63. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and validation of the User Version of the Mobile Application Rating Scale (uMARS). JMIR Mhealth Uhealth 2016 Jun 10;4(2):e72 [FREE Full text] [doi: 10.2196/mhealth.5849] [Medline: 27287964]
64. Kane H, Gourret Baumgart J, El-Hage W, Deloyer J, Maes C, Lebas M, et al. Opportunities and challenges for professionals in psychiatry and mental health care using digital technologies during the COVID-19 pandemic: systematic review. JMIR Hum Factors 2022 Feb 04;9(1):e30359 [FREE Full text] [doi: 10.2196/30359] [Medline: 34736224]
65. Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017 Oct;16(3):287-298 [FREE Full text] [doi: 10.1002/wps.20472] [Medline: 28941113]
66. Zelmer J, van Hoof K, Notarianni M, van Mierlo T, Schellenberg M, Tannenbaum C. An assessment framework for e-mental health apps in Canada: results of a modified Delphi process. JMIR Mhealth Uhealth 2018 Jul 09;6(7):e10016 [FREE Full text] [doi: 10.2196/10016] [Medline: 29986846]
67. Nicholas J, Larsen ME, Proudfoot J, Christensen H. Mobile apps for bipolar disorder: a systematic review of features and content quality. J Med Internet Res 2015 Aug 17;17(8):e198 [FREE Full text] [doi: 10.2196/jmir.4581] [Medline: 26283290]

XSL•FO
RenderX

68. Larsen ME, Nicholas J, Christensen H. A systematic assessment of smartphone tools for suicide prevention. PLoS One 2016;11(4):e0152285 [FREE Full text] [doi: 10.1371/journal.pone.0152285] [Medline: 27073900]

Abbreviations

CBT: cognitive behavioral therapy
ICC: intraclass correlation
iOS: iPhone Operating System
MARS: Mobile App Rating Scale
MARS-F: Mobile App Rating Scale–French
uMARS: user version of the Mobile App Rating Scale

Edited by L Buis; submitted 21.07.22; peer-reviewed by F Denis, A Finny, D Faye, J Grosjean; comments to author 02.09.22; revised version received 03.09.22; accepted 09.09.22; published 12.10.22.

Please cite as:
Carrouel F, du Sartz de Vigneulles B, Bourgeois D, Kabuth B, Baltenneck N, Nusbaum F, Burge V, Roy S, Buchheit S, Carrion-Martinaud ML, Massoubre C, Fraticelli L, Dussart C
Mental Health Mobile Apps in the French App Store: Assessment Study of Functionality and Quality
JMIR Mhealth Uhealth 2022;10(10):e41282
URL: https://mhealth.jmir.org/2022/10/e41282
doi:10.2196/41282
PMID:36223178

©Florence Carrouel, Benjamin du Sartz de Vigneulles, Denis Bourgeois, Bernard Kabuth, Nicolas Baltenneck, Fanny Nusbaum, Valérie Burge, Sylvain Roy, Sophie Buchheit, Marie-Line Carrion-Martinaud, Catherine Massoubre, Laurie Fraticelli, Claude Dussart. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 12.10.2022. 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.

XSL•FO
RenderX

Original Paper

Long-term Effects of the Use of a Step Count–Specific Smartphone App on Physical Activity and Weight Loss: Randomized Controlled Clinical Trial

Eiichi Yoshimura[1], PhD; Eri Tajiri[2], PhD; Ryota Michiwaki[2], MSc; Naoyuki Matsumoto[2], PhD; Yoichi Hatamoto[1], PhD; Shigeho Tanaka[3], PhD

[1]Department of Nutrition and Metabolism, National Institutes of Biomedical Innovation, Health and Nutrition, Tokyo, Japan

[2]Graduate School of Environmental & Symbiotic Sciences, Prefectural University of Kumamoto, Kumamoto, Japan

[3]Faculty of Nutrition, Kagawa Nutrition University, Saitama, Japan

Corresponding Author:
Eiichi Yoshimura, PhD
Department of Nutrition and Metabolism
National Institutes of Biomedical Innovation, Health and Nutrition
1-23-1 Toyama, Shinjuku-ku, Tokyo
Tokyo, 162-8636
Japan
Phone: 81 0332035725
Fax: 81 0332023278
Email: eyoshi@nibiohn.go.jp

Abstract

Background: Some studies on weight loss promotion using smartphone apps have shown a weight loss effect but not an increase in physical activity. However, the long-term effects of smartphone apps on weight loss and increasing physical activity have not been rigorously examined to date.

Objective: The aim of this study was to assess whether the use of a smartphone app will increase physical activity and reduce body weight.

Methods: In this parallel randomized clinical trial, participants recruited between April 2018 and June 2019 were randomized in equal proportions to a smartphone app group (n=55) or a control group (n=54). The intention-to-treat approach was used to analyze the data from December 2019 through November 2021. Before the intervention, an hour-long lecture on weight loss instruction and increasing physical activity was conducted once for both groups. Participants in both groups were instructed to weigh themselves immediately after waking up at least once daily from the start of the intervention. Monthly emails were sent advising the participants in both groups on how to lose weight and increase physical activity in order to maintain or increase motivation. Participants in the smartphone app group were instructed to open the app at least once a day to check their step count and rank. The primary outcome was daily accelerometer-measured physical activity (step count) and the secondary outcome was body weight. Since there was a significant difference in the wear time of the accelerometer depending on the intervention period (P<.001), the number of steps and moderate-to-vigorous physical activity were also evaluated per wear time.

Results: The mean age of the 109 participants in this study was 47 (SD 8) years. At baseline, the mean daily total steps were 7259 (SD 3256) steps per day for the smartphone app group and 8243 (SD 2815) steps per day for the control group. The difference in the step count per wear time between preintervention and postintervention was significantly different between the app group and the control group (average difference [95% CI], 65 [30 to 101] steps per hour vs –9 [–56 to 39] steps per hour; P=.042). The weight loss was –2.2 kg (SD –3.1%) in the smartphone app group and –2.2 kg (SD –3.1%) in the control group, with no significant difference between the groups. In addition, when divided into weekdays (Monday through Friday) and weekends (Saturday and Sunday), there was a significant interaction between step counts (P=.004) and MVPA (P=.003) during the intervention, with the app group showing higher interaction on weekends than the control group.

Conclusions: In this trial, the group with the smartphone app intervention showed increased physical activity, especially on weekends. However, this increased physical activity did not lead to increased weight loss.

XSL•FO
RenderX

Trial Registration: University Hospital Medical Information Network UMIN000033397; https://center6.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000037956

(JMIR Mhealth Uhealth 2022;10(10):e35628) doi:10.2196/35628

KEYWORDS

step counts; weight loss; smartphone app; step count–specific mobile app; physical activity; moderate-to-vigorous intensity physical activity; lifestyle intervention; mHealth; mobile app: mobile phone

Introduction

Background

Physical inactivity–related deaths contribute to US $13.7 billion in productivity losses, and physical inactivity is responsible for 13.4 million disability-adjusted life-years worldwide [1]. Conversely, higher levels of total physical activity at any intensity and less time spent sedentary are associated with a substantially reduced risk of premature mortality [2]. Increasing the amount of physical activity is also effective for weight loss [3,4]. Despite the many health benefits of physical activity, most Japanese adults do not meet the current recommendations for physical activity when assessed using objective measures [5]. Wearing a pedometer encourages increased step count in both the short and long terms [6,7]. A study [8] has shown that those who increased their daily steps over the monitoring period had a substantial reduction in mortality risk. A pedometer-based walking program resulted in moderate weight loss, with participants losing an average of 0.05 kg per week during the intervention period [9].

Recently, the use of mobile apps has led to notable success in increasing physical activity [10,11] and in weight control or weight reduction [12]. In particular, interventions with text messaging and personalization features seem to be more effective [10]. A meta-analysis showed that the use of mobile apps had a positive effect on physical activity measures corresponding to 1850 steps/day [10]. However, a meta-analysis of interventions that promote weight loss by using mobile apps showed a weight loss effect but not an increase in physical activity [12], and several factors might limit the effectiveness of mobile app–based physical activity interventions. First, the duration of most intervention studies using the app is short term (less than 6 months) and the long-term effects of such interventions are still unclear. Second, the 7 papers reviewed in the meta-analysis assessed physical activity by using questionnaires, which may have been unable to assess the change in physical activity. If physical activity can be increased by utilizing mobile apps during weight loss support, it will lead to an increase in energy expenditure, and further weight loss is expected. In addition, to understand the effectiveness of mobile apps in promoting increased physical activity during weight loss support, it is necessary to utilize mobile apps that focus solely on physical activity. Furthermore, body weight and physical activity fluctuate not only seasonally but also within the week [13,14]. Although the effect of app use on physical activity may differ between weekends and weekdays, the difference in the app effects has not yet been elucidated.

Objective

The purpose of this study was to determine whether using a mobile app would promote increased physical activity and weight loss after 32 weeks of the intervention. In addition, we aimed to assess the intraweek variability of physical activity during the intervention period and to evaluate the impact of using or not using the app.

Methods

Study Design and Participants

The participants in this study were recruited using a web portal for municipal employees according to the following inclusion criteria: (1) age, 30-60 years, (2) gained more weight than the weight at 20 years of age, (3) BMI>20 kg/m^2, and (4) possession of a smartphone. Participants with any disease and who could not obtain permission from their physicians were excluded from the study. The calculated sample size of 102 participants was determined based on a previous study [15] investigating the effect of using a smartphone app on increasing the number of steps taken (effect size=0.564586, α error=.05, power=.80). However, we recruited 110 participants based on the assumption that approximately 10% would drop out, and 109 participants who were finally considered for the analysis were randomized in equal proportions to the smartphone app group (n=55) or control group (n=54) by EY with a random number generator. The participants were first classified by sex and then randomly divided into 2 groups. Recruitment of the target population occurred between April 2018 and June 2019. The 32-week interventions were conducted twice during the same period between June 2018 and January 2020. The last evaluation date was set as 224 days (32 weeks) after the start of the intervention. Assessments of the impact of the intervention on physical activity and body weight were conducted at 10-12 weeks and 30-32 weeks from the start of the intervention. All data analyses were performed at the Kumamoto Prefectural University. The primary and secondary outcomes of this study were step count (physical activity) and body weight.

Ethics Approval

This study was conducted in accordance with the guidelines of the Consolidated Standards of Reporting Trials (CONSORT). This study followed the guidelines of the Declaration of Helsinki and was approved by the ethics committee for Clinical Research of the Prefectural University of Kumamoto (approval 30-30,01-20) and the ethics committee of the National Institutes of Biomedical Innovation, Health and Nutrition (approval 122-01). Informed consent was obtained from all the participants in this study. The protocol was registered in the University Hospital Medical Information Network (UMIN000033397).

XSL•FO
RenderX

Intervention

Before the intervention, EY gave both groups 1-hour group-based lectures on weight loss and increasing physical activity. The in-person lecture sessions consisted of 7 domains that focused on the following: (1) the benefits and barriers to engaging in health behaviors, (2) the health benefits of increased physical activity and weight loss, (3) how to calculate energy expenditure by activity intensity, (4) the amount of energy contained in cooked foods and seasonings, (5) how to set a goal of +1000 steps/day (increase walking time by approximately 10 minutes) of increased step count from participants' current (preintervention period) daily step counts, (6) how to set a weight loss goal of –5% from the participant's current body weight, and (7) healthy diet and weight maintenance. The participants' body weight and physical activity were measured for 3 consecutive weeks before and during the intervention (10-12 weeks and 30-32 weeks, respectively). Preintervention evaluations were assessed during the 3 weeks before the intervention began. In addition, dietary intake was assessed during the evaluation period using a questionnaire to assess the average amount of food consumed in the past month. Participants in both groups were instructed to weigh themselves immediately after waking up at least once daily from the start of the intervention. Monthly emails were sent advising the participants in both groups on how to lose weight and to increase physical activity to maintain or increase motivation. In the app group, a smartphone app (present in both Apple and Android smartphones) capable of managing the tracking steps [16] was downloaded by the participants before the intervention. The number of steps taken was displayed in the app in conjunction with the smartphone's built-in function to evaluate the number of steps taken. The app also displayed the information on walking distance, energy expenditure, and vegetable intake; however, participants were instructed to check only the number and ranking of steps taken. Within the app, the number of steps taken and the rank in the group could be tracked and this information was shared with the smartphone app group. Participants in the app group were instructed to open the app at least once a day to check their step count and rank. Participants in both groups wore accelerometers for 3 weeks that were only set to display the number of steps taken immediately after the intervention began so that they knew how many steps they were taking each day. Participants were asked to compare their daily activity to the feedback results to realize the amount of physical activity that had increased by more than 1000 steps before the intervention. In addition, physical activity and weight data measured prior to the intervention were fed at this time. The results of body weight and physical activity assessments at weeks 10-12 of the intervention were fed back within 1 month; in December, weight change results since the start of the intervention were fed back. The target for the step counts was to increase by at least +1000 steps/day from the preintervention rating. No maximum goal was set and the goal was to reach +1000 steps/day from the preintervention level, and those who reached the goal were encouraged to at least maintain that number of steps. For example, if the average step rating before the intervention was 6000 steps, they were advised to reach at least 7000 steps daily and encouraged to maintain at least 7000 steps even when the goal was reached.

Anthropometric Measurement

Height was measured using digital scales with a stadiometer to the nearest 0.1 cm (BW-306, Yamato scale) before the intervention period. The body weights of the participants were measured with a body composition monitor to the nearest 50 g (BC-308, Tanita). The participants were instructed to weigh themselves at least once a day, and the time of weighing was to be measured every day within an hour of waking up in the morning, wearing as similar clothes as possible and under fasting conditions. The measured weight and time data were recorded on the Secure Digital card built into the body composition monitor. BMI was calculated as weight (kg) divided by height (in m^2).

Physical Activity

Physical activity was measured for 3 weeks [17] at 1-minute epochs by using a triaxial accelerometer (Active Style Pro HJA-750C) [18] before and during the intervention (10-12 weeks and 30-32 weeks, respectively). The accelerometer was worn on their waist, except while sleeping or bathing. Physical activity was assessed using step count and activity time based on the intensity levels. The obtained physical activity intensity level in each minute was classified as sedentary behavior (≤1.5 metabolic equivalents [METs]), light physical activity (1.6-2.9 METs), and moderate-to-vigorous physical activity (MVPA, ≥3.0 METs). A consecutive zero count of ≥60 minutes was defined as nonwear time. Assessment of accelerometer data was adopted if there was more than 600 minutes of wear time per day. Overall participant wearing time (min/day) of the accelerometer decreased from 940 (SD 102) min/day at baseline to 919 (SD 118) min/day at 12 weeks and 905 (SD 125) min/day at 32 weeks (P<.001). Thus, to account for the differences in the wearing time, the average number of steps and MVPA per wearing time (per hour) were calculated with reference to previous studies [11,19]. During the measurement period (before the intervention, 10-12 weeks, and 30-32 weeks), the accelerometer display was set to not be able to see the amount of physical activity for the day.

Dietary Intake

Food intake was assessed using a validated brief self-administered diet history questionnaire [20,21]. Participants reported the foods they consumed in the past month by selecting the frequency of each food group and the average intake per week. The brief self-administered diet history questionnaire consists of questions on frequency of food and beverage consumption, frequency of rice and miso soup consumption per day, frequency of alcohol consumption and the percentage of alcohol in each of the 5 alcoholic beverages, cooking methods for the dishes eaten most often, and general eating habits. After the participants completed the questionnaire, a dietician checked the completed questionnaire with the participant.

Statistical Analysis

Changes in the body weight during the intervention are shown as raw data and as a moving average over a week. The demographic variables were assessed using independent sample two-sided t tests for continuous data and the chi-square test for categorical data to compare between groups. All analyses were

XSL•FO
RenderX

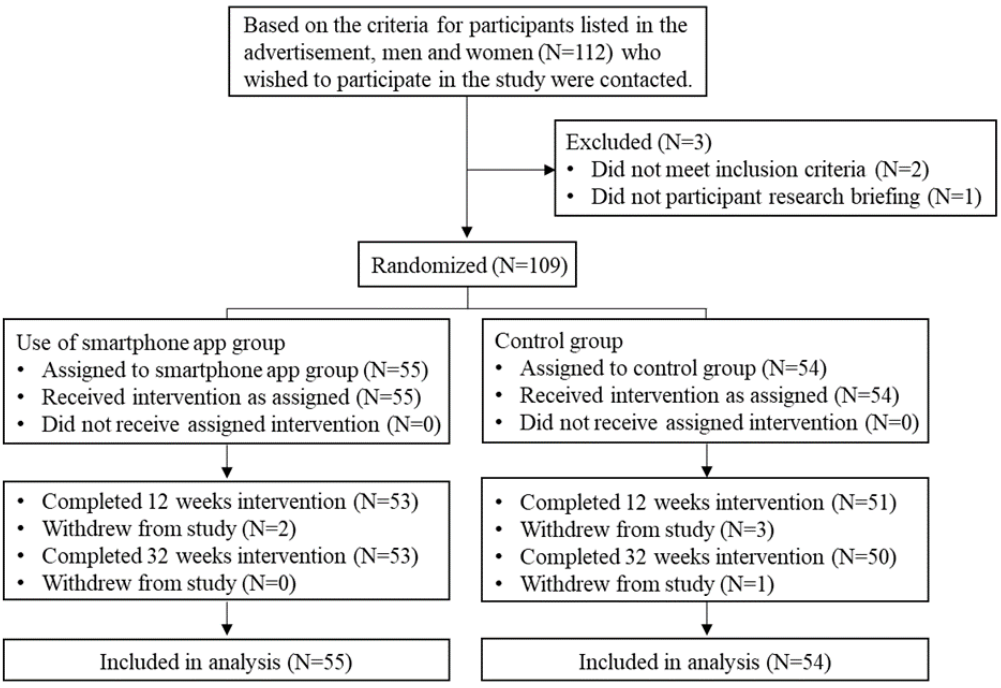
performed using an intention-to-treat approach. To evaluate the effects of using the smartphone app on physical activity and body weight, this study used a mixed-design analysis of variance between the participant groups (smartphone app group vs control group) and within participant groups based on repeated measurements (preintervention vs postintervention at 12 and 32 weeks, respectively). To assess intraweek variability in body weight and physical activity before and during intervention, secondary analyses included comparisons of intraweek changes in body weight and physical activity (step count and MVPA) by using mixed-design analysis of variance. We also estimated the effect of the fixed model and tested its significance for the day of the week with reference to a previous study [17]. The data for body weight are presented in the change based on Friday since the standard error was larger than the intraweek variation and it is difficult to see the variation. Statistical significance was set at $P<.05$. Statistical analyses were performed using the SPSS version 22.0 software (IBM Japan Ltd).

Results

Participants and Intervention Adherence

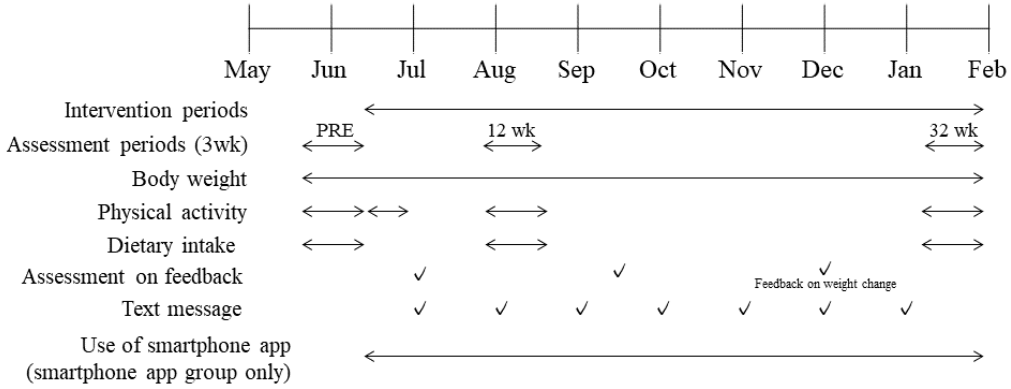
The flowchart of the participants included in this study is shown in Figure 1. The study protocol is illustrated in Figure 2. Two and 3 participants dropped out from the smartphone app and control groups, respectively, between the preintervention and 12 weeks and 1 participant in the control group dropped out at 12-32 weeks. The final dropout rate at 32 weeks was 4% (2/55) in the smartphone app group and 7% (4/54) in the control group, with no significant difference between the 2 groups ($P=.44$). Moreover, before and after the intervention period, 3 participants in the smartphone app group (completion rate, 52/55, 95%) and 6 participants in the control group (completion rate, 48/54, 89%) could not be assessed because the weight and physical activity data were only provided for a few days or due to errors in the analysis process. Thus, the number of participants with complete weight and physical activity data before and after the intervention was 50 in the smartphone app group and 44 in the control group. The proportion of those in the smartphone app group who checked the app at least once a day during the intervention period was 73.4% (overall: 80.1% at the start of intervention to 12 weeks and 73.6% at 12 to 32 weeks). During the 32-week intervention period, the average number of days with at least one weight measurement per day was 187 in the smartphone app group and 185 in the control group. The trends in the weight, BMI, fasting blood glucose level, and hemoglobin A_{1c} level of the 102 participants with the results of physical examinations for the past 10 years and the intervention year are shown in the Multimedia Appendix 1. The year in which the intervention was implemented is shown as zero. The participants tended to gain weight and BMI over the years, and their glycemic status worsened.

Figure 1. Study flow diagram.



XSL•FO
RenderX

Figure 2. Study protocol. Wk: weeks.



Intervention Effects

The baseline data of the 109 participants by their group are shown in Table 1. Figure 3 shows the change in the body weight during the intervention period from June to the end of January (32 weeks) in the years 2018-2020. After 32 weeks of intervention, the weight loss was –2.2 kg (SD –3.1%) in the smartphone app group and –2.2 kg (SD –3.1%) in the control group, with no significant difference between the groups. Intraweek variations in the body weight before the intervention are shown in Figure 4. Before the intervention, there was no significant interaction (group × day of the week) in body weight (Figure 4A), step count (Figure 4B), and MVPA (Figure 4C) between the groups (P>.05).

The effects of the intervention on body weight and physical activity before and after the intervention are shown in Table 2. The step count per wear time and MVPA per wear time showed a significant interaction between the groups (P=.04 and P=.03, respectively). Similar results were obtained for the per-protocol set. The intraweek variations in body weight, step count, and MVPA with or without wear time during the intervention period are shown in Figure 5. There was a significant interaction (group × day of the week) in the step count and MVPA per wear time between the groups (P=.01 and P=.007, respectively), and the step counts on Saturdays and Sundays in the smartphone app group were higher than those in the control group (P<.05). In addition, when divided into weekdays (Monday through Friday) and weekends (Saturday and Sunday), there was a significant interaction between step counts (P=.004) and MVPA (P=.003) during the intervention, with the app group showing higher interaction on weekends than the control group. In the analysis including both groups, the correlations between change in step counts and change in energy intake (r=–0.141) and weight loss (r=–0.026) were not statistically significant; however, there was a significant correlation between dietary intake and weight loss (r=0.198, P=.048).

XSL•FO
RenderX

Table 1. Baseline information of the participants (N=109).

	Smartphone app group (n=55)	Control group (n=54)	*P* value
General characteristics			
Women, n (%)	26 (47)	24 (44)	.85
Age (years), mean (SD)	47 (8)	47 (8)	.93
Body weight (kg), mean (SD)	71.0 (13.9)	70.0 (13.0)	.69
BMI (kg/m^2), mean (SD)	25.9 (4.1)	25.3 (3.6)	.46
Step counts (steps/day), mean (SD)	7259 (3256)	8243 (2815)	.09
Activity time of 1.5 METs[a] (min/day), mean (SD)	571.5 (99.6)	603.5 (109.5)	.11
Activity time of 1.6-2.9 METs (min/day), mean (SD)	302.2 (79.9)	289.3 (69.6)	.37
Activity time of over 3.0 METs (min/day), mean (SD)	55.2 (25.2)	61.5 (23.8)	.18
Activity time of 1.5 METs (%), mean (SD)	61.4 (8.7)	62.9 (6.6)	.31
Activity time of 1.6-2.9 METs (%), mean (SD)	32.7 (8.0)	30.5 (6.2)	.13
Activity time of over 3.0 METs (%), mean (SD)	6.0 (2.7)	6.6 (2.6)	.25
Energy intake (kcal/day), mean (SD)	2005 (649)	1857 (624)	.23
Protein (%), mean (SD)	14.6 (2.9)	15.5 (2.6)	.08
Fat (%), mean (SD)	27.8 (5.1)	28.9 (5.6)	.29
Carbohydrate (%), mean (SD)	57.6 (7.1)	55.5 (7.6)	.16
Self-reported cardiovascular risk factors, n (%)			
Overweight/obese[b]	25 (45)	22 (41)	.70
Obese[c]	8 (15)	7 (13)	>.99
Hypertension	5 (9)	6 (11)	.76
Dyslipidemia	1 (2)	4 (7)	.21
Diabetes	1 (2)	0 (0)	>.99
Current smoker	2 (4)	5 (9)	.27

[a]MET: metabolic equivalent.
[b]BMI≥25 kg/m^2.
[c]BMI≥30 kg/m^2.

XSL•FO
RenderX

Figure 3. Changes in body weight during the intervention period (32 weeks). The change in the body weight of the participants is shown by the solid line for the smartphone app group and the dotted line for the control group. (A) Values are shown as average and (B) as the moving average over 1 week. The change in body weight before and after the intervention (preintervention, 12 weeks, and 32 weeks) was not significantly different between the groups.

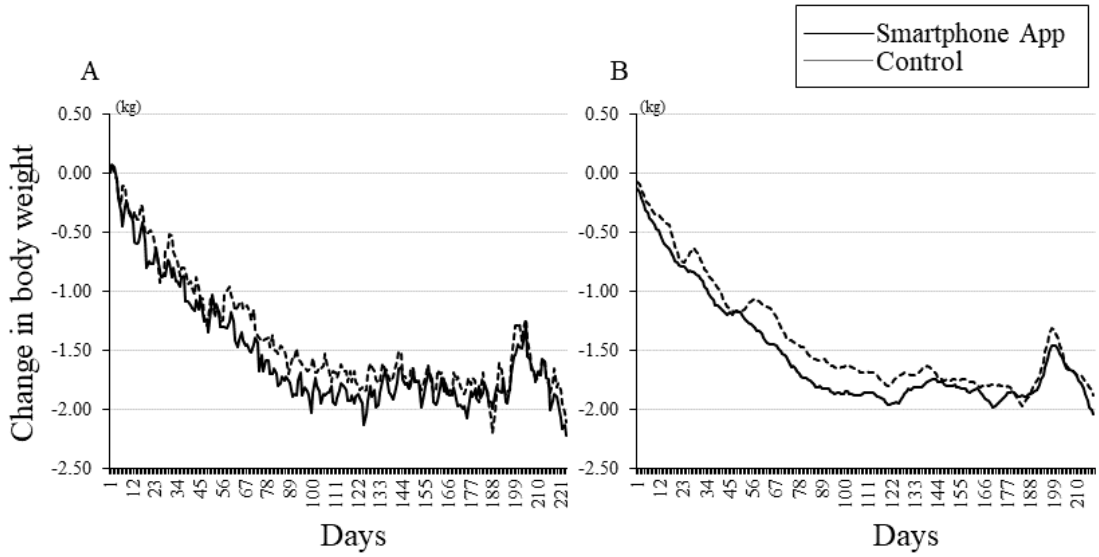
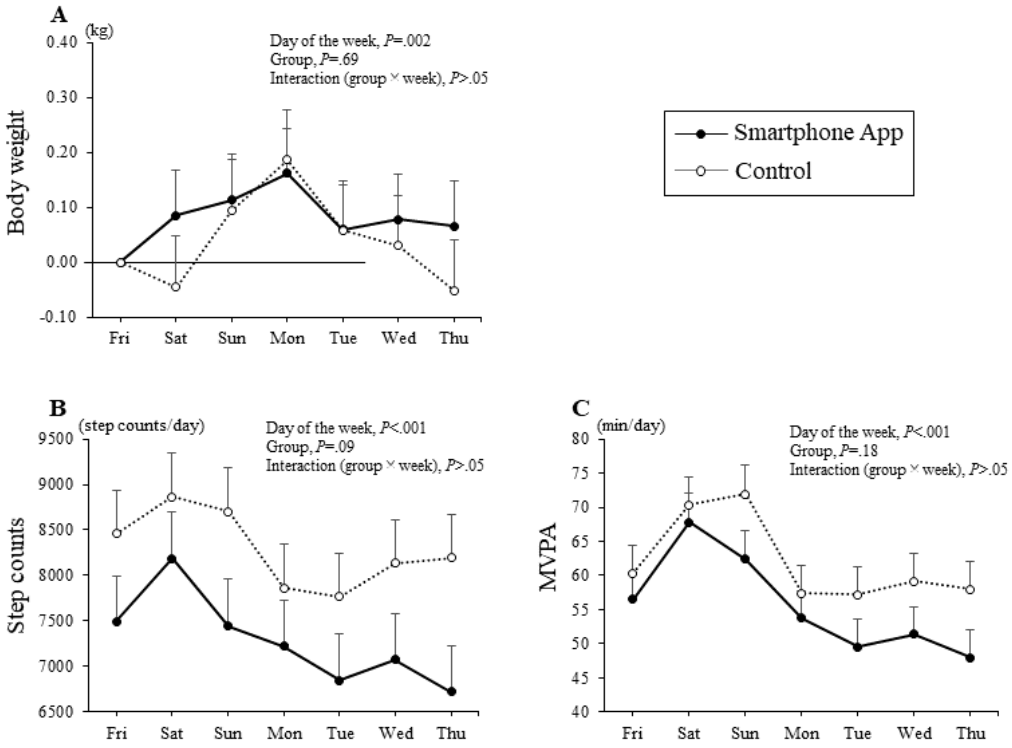


Figure 4. Intraweek variation in (A) body weight, (B) step counts, and (C) moderate-to-vigorous physical activity before the intervention. Missing data were taken into account and analyzed using two-way repeated measures mixed analysis of variance to examine the interaction between the groups. Values are presented as means and standard errors. MVPA: moderate-to-vigorous physical activity.



XSL•FO
RenderX

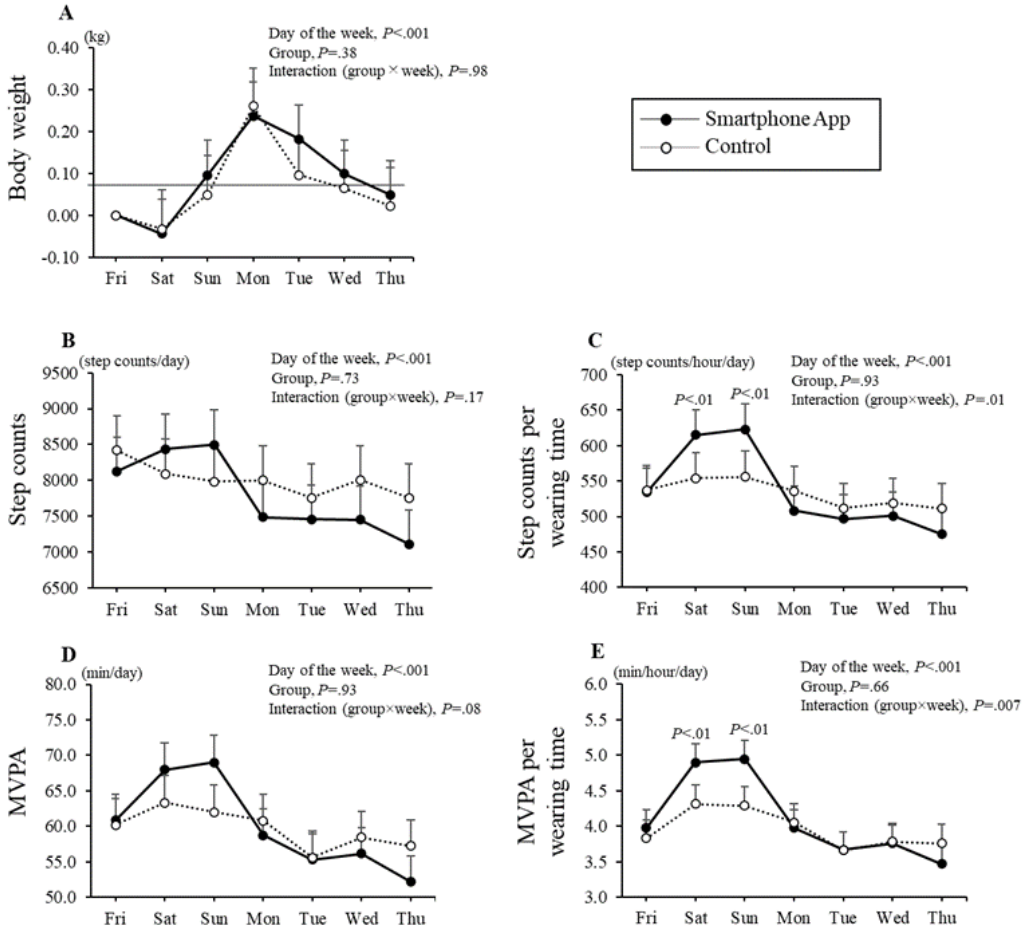
Table 2. Intervention effects on body weight, physical activity, and dietary intake before and after intervention.

	Smartphone app group			Control group			Group × interaction, *P* value
	Preintervention, mean (95% CI)	12 weeks, mean (95% CI)	32 weeks, mean (95% CI)	Preintervention, mean (95% CI)	12 weeks, mean (95% CI)	32 weeks, mean (95% CI)	
Intention-to-treat							
Body weight (kg)	70.9 (67.4-74.5)	68.8 (65.3-72.4)	68.8 (65.2-72.3)	70.0 (66.4-73.5)	68.0 (64.4-71.6)	67.8 (64.2-71.4)	.94
Step count (steps/day)	7259 (6335-8183)	7850 (6921-8779)	7846 (6910-8782)	8243 (7465-9021)	8143 (7346-8940)	7806 (6998-8615)	.06
Step counts per wear time (steps/h/day)	473 (408-538)	527 (462-592)	538 (473-604)	525 (472-579)	528 (4737-583)	517 (461-572)	.04
MVPA[a] (min/day)	55 (48-62)	58 (51-65)	62 (55-69)	62 (55-68)	60 (54-67)	59 (52-65)	.05
MVPA per wear time (min/h/day)	4 (3-4)	4 (3-4)	4 (4-5)	4 (3-4)	4 (3-4)	4 (3-4)	.03
Energy intake (kcal/day)	2005 (1844-2166)	1783 (1621-1946)	1836 (1673-2000)	1857 (1694-2020)	1718 (1553-1883)	1758 (1593-1923)	.66
Protein (%)	14.6 (13.8-15.4)	15.6 (14.8-16.4)	15.5 (14.7-16.3)	15.5 (14.7-16.3)	16.1 (15.3-16.9)	16.1 (15.3-16.9)	.65
Fat (%)	27.8 (26.3-29.4)	28.7 (27.1-30.2)	27.5 (25.9-29.1)	28.9 (27.3-30.5)	29.5 (27.9-31.1)	28.3 (26.7-29.9)	.96
Carbohydrate (%)	57.6 (55.4-59.7)	55.7 (53.5-57.9)	57 (54.8-59.2)	55.5 (53.4-57.7)	54.4 (52.2-56.6)	55.7 (53.4-57.9)	.87
Per protocol							
Body weight (kg)	70.6 (67.0-74.2)	68.4 (64.9-71.9)	68.4 (64.9-71.9)	68.3 (64.7-72.0)	66.4 (62.8-69.9)	66.2 (62.6-69.7)	.90
Step count (steps/day)	7130 (6287-7972)	7833 (6939-8728)	7779 (6874-8685)	8231 (7333-9129)	8149 (7196-9103)	7826 (6861-8791)	.05
Step counts per wear time (steps/h/day)	465 (408-522)	527 (465-588)	535 (467-602)	526 (465-587)	527 (462-592)	520 (448-592)	.047
MVPA (min/day)	54 (48-61)	58 (51-65)	61 (54-68)	61 (54-68)	60 (53-67)	59 (51-66)	.04
MVPA per wear time (min/h/day)	4 (3-4)	4 (3-4)	4 (4-5)	4 (3-4)	4 (3-4)	4 (3-4)	.03
Energy intake (kcal/day)	1966 (1800-2133)	1759 (1600-1919)	1811 (1677-1945)	1851 (1680-2022)	1704 (1539-1870)	1752 (1576-1927)	.79
Protein (%)	14.8 (14.0-15.5)	15.7 (14.9-16.6)	15.6 (14.8-16.4)	15.7 (15.0-16.4)	16.2 (15.4-17.1)	16.2 (15.3-17.0)	.68
Fat (%)	27.8 (26.4-29.2)	28.7 (27.1-30.3)	27.5 (25.8-29.2)	28.8 (27.2-30.3)	29.3 (27.6-30.9)	28.0 (26.3-29.7)	.90
Carbohydrate (%)	57.4 (55.4-59.4)	55.5 (53.2-57.9)	56.9 (54.5-59.2)	55.5 (53.4-57.6)	54.5 (52.2-56.8)	55.9 (53.5-58.2)	.81

[a]MVPA: moderate-to-vigorous physical activity.

XSL•FO
RenderX

Figure 5. Intraweek variation in (A) body weight, (B and C) step counts, and (D and E) moderate-to-vigorous physical activity during the intervention. Data were combined from 12 and 32 weeks to analyze the relationship between physical activity and intraweek variability. Missing data were taken into account and analyzed using two-way repeated-measures mixed analysis of variance to examine the interaction between the groups. Values are presented as means and standard errors. MVPA: moderate-to-vigorous physical activity.



Discussion

The aim of this study was to determine the effects of using a mobile app on increasing physical activity and weight loss by assessing accelerometer data and weight loss data after 32 weeks of the app intervention. Our findings showed that the use of a step count–specific mobile app for the assessment of physical activity for weight loss might be effective in increasing the step count, although it may not affect the amount of weight loss. In addition, we found that the effects of using the mobile app on physical activity differed between weekends and weekdays and that the mobile app showed data of higher physical activity on weekends.

Flores Mateo et al [22] and Islam et al [12] reported meta-analyses that were related to our study. Their meta-analyses indicated that the app intervention group showed more weight loss than the control group, but there was no statistically significant change in physical activity. Most previous studies assessed habitual physical activity through questionnaires or self-reports [15,23-28]. To our knowledge, our study is the first to assess physical activity by using accelerometers for weight loss and to examine the long-term effects of physical activity by using step count–specific mobile apps. In this study, we found that the use of a step count–specific mobile app for weight loss led to an increase in physical activity. Previous studies have demonstrated that increased physical activity reduces the risk of mortality and has various physiological benefits [29-31]. The impact on physical activity from using a step-counting app for weight loss is not as significant as expected; however, continued use might provide physiological benefits. In the future, the long-term impact of using health indicators should be examined.

Inconsistent with that reported in a meta-analysis [22], the change in weight loss was not significantly different between the 2 groups in our study. This may be because the control group in our study demonstrated a certain weight loss effect after 1 lecture and a few text messages even without using the app. A study has shown that a single motivational lecture can promote moderate weight loss in the short term [32]. Moreover, our study and the previous study [32] consisted of populations that were motivated to lose weight, which may have influenced the results. In addition, many intervention studies using mobile apps included several support tools in the apps, such as texts, emails, internet, interactive chatbots, and voice agents. Although the effectiveness of an intervention may be increased by the inclusion of many support features in mobile apps, it is difficult to understand the exact factors that affect the observed changes. Pedometers and accelerometers affect the number of steps just

XSL•FO
RenderX

by wearing them, but it is unclear whether the use of mobile apps that focus only on step count affects the step count. If the number of steps taken by a 70-kg person increases by 1000 and the intensity is 3 METs, the person should lose more than 1 kg in 32 weeks, but our study showed no such effect. Wu et al [33] indicated in their meta-analysis that the pooled weight loss was 1.14 kg or 0.50 kg/m^2 greater for the diet-plus-exercise group than for the diet-only group. However, results with an intervention period of less than 1 year showed no intervention effect. Furthermore, most studies did not show the expected weight loss effect from the energy expenditure generated by the set exercise [33-35]. Interventions that combine dietary restriction and physical activity may attenuate the effects of physical activity interventions. The weight loss after 32 weeks of the intervention in our study was approximately 2 kg. Further improvements in lecture content and support tools are needed to increase the effectiveness of additional weight loss. In addition, participants in the app group were instructed to use the app daily to check their steps and rank. Although such an approach is intended to motivate participants, we could not assess the motivational impact of app use. Further research is needed in this direction in the future.

Notably, we found that the effects of the mobile app on physical activity differed between weekends and weekdays, that is, the mobile app data showed higher physical activity on weekends. In a study on Japanese white-collar workers, the sedentary behavior time was significantly longer on weekdays than on weekends (598 min/day vs 479 min/day, respectively; $P<.001$) [36]. However, among blue-collar workers, there was no significant difference in the sedentary behavior time between weekdays and weekends (462 min/day vs 485 min/day; $P=.43$) [36]. The proportion of workers who achieved the recommended sedentary behavior levels (≤8 hours) was only 4.8% for white-collar workers on weekdays and 54.8% on weekends ($P=.04$) [36]. All the participants in our study were prefectural employees, and the weekends were Saturday and Sunday. Although change is required regarding work or after-work behavior on weekdays, it may be difficult to increase the number of steps taken on weekdays. Our study showed that the use of smartphone apps could increase physical activity by increasing the step count, especially during leisure time on nonworking days such as Saturday and Sunday. These results can be obtained using an accelerometer. The originality of this study is in the use of accelerometers to assess physical activity and the use of up to 6 weeks of physical activity data assessed at weeks 12 and 32 of the intervention to examine the within-week variability.

The dropout rate in our study was less than 10%, despite the long intervention period of 32 weeks, and there was no difference between the 2 groups. The reasons for this cannot be ascertained; however, intervention content such as monthly emails and feedback on the results may have had an impact. Continued participation in the intervention is an essential factor affecting the validation of intervention effectiveness and should be investigated in future studies.

This study has several limitations. First, all the participants in this study were prefectural employees, which limits the generalizability of the study. Second, although this study calculated the sample size and conducted an intervention, it is possible that sample size estimation was inadequate due to the lack of appropriate studies utilizing apps focused on step count and using physical activity and weight loss as outcomes. Future large-scale intervention studies are needed. Third, although the duration of the intervention in this study was longer compared to that in previous studies, it is necessary to examine the impact of the intervention for more than 1 year, and the influence of the seasons needs to be considered. Further, the use of the app may have affected the motivation for the intervention but could not be assessed in this study. Future studies should also evaluate the motivational impact of app use. Finally, it is necessary to develop support tools to increase not only the amount of physical activity but also the weight loss effect. However, the major strength of our study was the parallel randomized controlled trial design, which indicates that our findings are reliable.

In conclusion, this study shows that the use of a step count–specific mobile app during weight loss support might be effective in increasing the step count, although it might not affect the amount of weight loss. Moreover, we found that the effects of the mobile app on physical activity differed between weekends and weekdays, with the mobile app data showing higher physical activity on weekends. Future studies need to focus on the development of methods for increasing the effectiveness of physical activity and weight loss by using mobile apps.

Acknowledgments

This work was supported by the JST-Mirai Program (grant JP0000012) and a grant-in-aid for Scientific Research (C) (KAKENHI 20K02410) and (A) (KAKENHI 16H01877).

Authors' Contributions

EY and ST conceptualized this study and the methodology. EY, ET, RM, NM, and YH performed the formal analysis. EY performed investigation, visualization, supervision, and project administration. EY, RM, and ET performed data curation. EY and YH performed the writing and original draft preparation, EY, ST, RM, NM, and YH reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

XSL•FO
RenderX

Multimedia Appendix 1
Changes in body composition and blood glucose levels in physical examination findings. Missing data were considered and analyzed using one-way repeated-measures mixed analysis of variance. Values are presented as means and standard errors.
[PNG File , 18 KB - mhealth_v10i10e35628_app1.png]

Multimedia Appendix 2
CONSORT-eHEALTH (V 1.6.1) checklist.
[PDF File (Adobe PDF File), 7579 KB - mhealth_v10i10e35628_app2.pdf]

References

1. Ding D, Lawson KD, Kolbe-Alexander TL, Finkelstein EA, Katzmarzyk PT, van Mechelen W, Lancet Physical Activity Series 2 Executive Committee. The economic burden of physical inactivity: a global analysis of major non-communicable diseases. Lancet 2016 Sep 24;388(10051):1311-1324. [doi: 10.1016/S0140-6736(16)30383-X] [Medline: 27475266]
2. Ekelund U, Tarp J, Steene-Johannessen J, Hansen BH, Jefferis B, Fagerland MW, et al. Dose-response associations between accelerometry measured physical activity and sedentary time and all cause mortality: systematic review and harmonised meta-analysis. BMJ 2019 Aug 21;366:l4570 [FREE Full text] [doi: 10.1136/bmj.l4570] [Medline: 31434697]
3. Ross R, Janssen I, Dawson J, Kungl A, Kuk JL, Wong SL, et al. Exercise-induced reduction in obesity and insulin resistance in women: a randomized controlled trial. Obes Res 2004 May;12(5):789-798 [FREE Full text] [doi: 10.1038/oby.2004.95] [Medline: 15166299]
4. Ross R, Dagnone D, Jones PJ, Smith H, Paddags A, Hudson R, et al. Reduction in obesity and related comorbid conditions after diet-induced weight loss or exercise-induced weight loss in men. A randomized, controlled trial. Ann Intern Med 2000 Jul 18;133(2):92-103. [doi: 10.7326/0003-4819-133-2-200007180-00008] [Medline: 10896648]
5. Annual report of the national health and nutrition survey in Japan, 2019. Ministry of Health, Labour and Welfare. 2019. URL: https://www.mhlw.go.jp/bunya/kenkou/eiyou/h28-houkoku.html [accessed 2022-09-27]
6. Clemes SA, Parker RAA. Increasing our understanding of reactivity to pedometers in adults. Med Sci Sports Exerc 2009 Mar;41(3):674-680. [doi: 10.1249/MSS.0b013e31818cae32] [Medline: 19204581]
7. Harris T, Kerry SM, Limb ES, Furness C, Wahlich C, Victor CR, et al. Physical activity levels in adults and older adults 3-4 years after pedometer-based walking interventions: Long-term follow-up of participants from two randomised controlled trials in UK primary care. PLoS Med 2018 Mar;15(3):e1002526 [FREE Full text] [doi: 10.1371/journal.pmed.1002526] [Medline: 29522529]
8. Dwyer T, Pezic A, Sun C, Cochrane J, Venn A, Srikanth V, et al. Objectively Measured Daily Steps and Subsequent Long Term All-Cause Mortality: The Tasped Prospective Cohort Study. PLoS One 2015;10(11):e0141274 [FREE Full text] [doi: 10.1371/journal.pone.0141274] [Medline: 26536618]
9. Richardson CR, Newton TL, Abraham JJ, Sen A, Jimbo M, Swartz AM. A meta-analysis of pedometer-based walking interventions and weight loss. Ann Fam Med 2008;6(1):69-77 [FREE Full text] [doi: 10.1370/afm.761] [Medline: 18195317]
10. Laranjo L, Ding D, Heleno B, Kocaballi B, Quiroz JC, Tong HL, et al. Do smartphone applications and activity trackers increase physical activity in adults? Systematic review, meta-analysis and metaregression. Br J Sports Med 2021 Apr;55(8):422-432. [doi: 10.1136/bjsports-2020-102892] [Medline: 33355160]
11. Fukuoka Y, Haskell W, Lin F, Vittinghoff E. Short- and Long-term Effects of a Mobile Phone App in Conjunction With Brief In-Person Counseling on Physical Activity Among Physically Inactive Women: The mPED Randomized Clinical Trial. JAMA Netw Open 2019 May 03;2(5):e194281 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.4281] [Medline: 31125101]
12. Islam MM, Poly TN, Walther BA, Jack Li Y. Use of Mobile Phone App Interventions to Promote Weight Loss: Meta-Analysis. JMIR Mhealth Uhealth 2020 Jul 22;8(7):e17039. [doi: 10.2196/17039] [Medline: 32706724]
13. Racette SB, Weiss EP, Schechtman KB, Steger-May K, Villareal DT, Obert KA, et al. Influence of weekend lifestyle patterns on body weight. Obesity (Silver Spring) 2008 Aug;16(8):1826-1830 [FREE Full text] [doi: 10.1038/oby.2008.320] [Medline: 18551108]
14. Turicchi J, O'Driscoll R, Horgan G, Duarte C, Palmeira AL, Larsen SC, et al. Weekly, seasonal and holiday body weight fluctuation patterns among individuals engaged in a European multi-centre behavioural weight loss maintenance intervention. PLoS One 2020;15(4):e0232152 [FREE Full text] [doi: 10.1371/journal.pone.0232152] [Medline: 32353079]
15. Glynn LG, Hayes PS, Casey M, Glynn F, Alvarez-Iglesias A, Newell J, et al. Effectiveness of a smartphone application to promote physical activity in primary care: the SMART MOVE randomised controlled trial. Br J Gen Pract 2014 Jul;64(624):e384-e391 [FREE Full text] [doi: 10.3399/bjgp14X680461] [Medline: 24982490]
16. Kumamoto Prefecture. URL: https://www.pref.kumamoto.jp/soshiki/44/5078.html [accessed 2022-09-27]
17. Yoshimura E, Tajiri E, Hatamoto Y, Tanaka S. Changes in Season Affect Body Weight, Physical Activity, Food Intake, and Sleep in Female College Students: A Preliminary Study. Int J Environ Res Public Health 2020 Nov 24;17(23):8713 [FREE Full text] [doi: 10.3390/ijerph17238713] [Medline: 33255205]

XSL•FO
RenderX

18. Ohkawara K, Oshima Y, Hikihara Y, Ishikawa-Takata K, Tabata I, Tanaka S. Real-time estimation of daily physical activity intensity by a triaxial accelerometer and a gravity-removal classification algorithm. Br J Nutr 2011 Jun;105(11):1681-1691. [doi: 10.1017/S0007114510005441] [Medline: 21262061]
19. Davis MG, Fox KR, Stathi A, Trayers T, Thompson JL, Cooper AR. Objectively measured sedentary time and its association with physical function in older adults. J Aging Phys Act 2014 Oct;22(4):474-481. [doi: 10.1123/japa.2013-0042] [Medline: 24085473]
20. Kobayashi S, Murakami K, Sasaki S, Okubo H, Hirota N, Notsu A, et al. Comparison of relative validity of food group intakes estimated by comprehensive and brief-type self-administered diet history questionnaires against 16 d dietary records in Japanese adults. Public Health Nutr 2011 Jul;14(7):1200-1211. [doi: 10.1017/S1368980011000504] [Medline: 21477414]
21. Kobayashi S, Honda S, Murakami K, Sasaki S, Okubo H, Hirota N, et al. Both comprehensive and brief self-administered diet history questionnaires satisfactorily rank nutrient intakes in Japanese adults. J Epidemiol 2012;22(2):151-159 [FREE Full text] [doi: 10.2188/jea.je20110075] [Medline: 22343326]
22. Flores Mateo G, Granado-Font E, Ferré-Grau C, Montaña-Carreras X. Mobile Phone Apps to Promote Weight Loss and Increase Physical Activity: A Systematic Review and Meta-Analysis. J Med Internet Res 2015 Nov 10;17(11):e253 [FREE Full text] [doi: 10.2196/jmir.4836] [Medline: 26554314]
23. Smith JJ, Morgan PJ, Plotnikoff RC, Dally KA, Salmon J, Okely AD, et al. Smart-phone obesity prevention trial for adolescent boys in low-income communities: the ATLAS RCT. Pediatrics 2014 Sep;134(3):e723-e731. [doi: 10.1542/peds.2014-1012] [Medline: 25157000]
24. Allen JK, Stephens J, Dennison Himmelfarb CR, Stewart KJ, Hauck S. Randomized controlled pilot study testing use of smartphone technology for obesity treatment. J Obes 2013;2013:151597 [FREE Full text] [doi: 10.1155/2013/151597] [Medline: 24392223]
25. Partridge SR, McGeechan K, Hebden L, Balestracci K, Wong AT, Denney-Wilson E, et al. Effectiveness of a mHealth Lifestyle Program With Telephone Support (TXT2BFiT) to Prevent Unhealthy Weight Gain in Young Adults: Randomized Controlled Trial. JMIR Mhealth Uhealth 2015 Jun 15;3(2):e66 [FREE Full text] [doi: 10.2196/mhealth.4530] [Medline: 26076688]
26. Allman-Farinelli M, Partridge SR, McGeechan K, Balestracci K, Hebden L, Wong A, et al. A Mobile Health Lifestyle Program for Prevention of Weight Gain in Young Adults (TXT2BFiT): Nine-Month Outcomes of a Randomized Controlled Trial. JMIR Mhealth Uhealth 2016 Jun 22;4(2):e78 [FREE Full text] [doi: 10.2196/mhealth.5768] [Medline: 27335237]
27. Turner-McGrievy G, Tate D. Tweets, Apps, and Pods: Results of the 6-month Mobile Pounds Off Digitally (Mobile POD) randomized weight-loss intervention among adults. J Med Internet Res 2011 Dec 20;13(4):e120 [FREE Full text] [doi: 10.2196/jmir.1841] [Medline: 22186428]
28. Laing BY, Mangione CM, Tseng C, Leng M, Vaisberg E, Mahida M, et al. Effectiveness of a smartphone application for weight loss compared with usual care in overweight primary care patients: a randomized, controlled trial. Ann Intern Med 2014 Nov 18;161(10 Suppl):S5-12 [FREE Full text] [doi: 10.7326/M13-3005] [Medline: 25402403]
29. Camps SGJA, Verhoef SPM, Westerterp KR. Physical activity and weight loss are independent predictors of improved insulin sensitivity following energy restriction. Obesity (Silver Spring) 2016 Feb;24(2):291-296 [FREE Full text] [doi: 10.1002/oby.21325] [Medline: 26813523]
30. Inoue M, Iso H, Yamamoto S, Kurahashi N, Iwasaki M, Sasazuki S, Japan Public Health Center-Based Prospective Study Group. Daily total physical activity level and premature death in men and women: results from a large-scale population-based cohort study in Japan (JPHC study). Ann Epidemiol 2008 Jul;18(7):522-530. [doi: 10.1016/j.annepidem.2008.03.008] [Medline: 18504139]
31. Samitz G, Egger M, Zwahlen M. Domains of physical activity and all-cause mortality: systematic review and dose-response meta-analysis of cohort studies. Int J Epidemiol 2011 Oct;40(5):1382-1400. [doi: 10.1093/ije/dyr112] [Medline: 22039197]
32. Nakata Y, Sasai H, Tsujimoto T, Hashimoto K, Kobayashi H. A Single Motivational Lecture Can Promote Modest Weight Loss: A Randomized Controlled Trial. Obes Facts 2020;13(2):267-278 [FREE Full text] [doi: 10.1159/000506813] [Medline: 32289804]
33. Wu T, Gao X, Chen M, van Dam RM. Long-term effectiveness of diet-plus-exercise interventions vs. diet-only interventions for weight loss: a meta-analysis. Obes Rev 2009 May;10(3):313-323. [doi: 10.1111/j.1467-789X.2008.00547.x] [Medline: 19175510]
34. Yoshimura E, Kumahara H, Tobina T, Matsuda T, Ayabe M, Kiyonaga A, et al. Lifestyle intervention involving calorie restriction with or without aerobic exercise training improves liver fat in adults with visceral adiposity. J Obes 2014;2014:197216 [FREE Full text] [doi: 10.1155/2014/197216] [Medline: 24864199]
35. Yoshimura E, Kumahara H, Tobina T, Matsuda T, Watabe K, Matono S, et al. Aerobic exercise attenuates the loss of skeletal muscle during energy restriction in adults with visceral adiposity. Obes Facts 2014;7(1):26-35 [FREE Full text] [doi: 10.1159/000358576] [Medline: 24457527]
36. Fukushima N, Kikuchi H, Amagasa S, Machida M, Kitabayashi M, Hayashi T, et al. Exposure to prolonged sedentary behavior on weekdays rather than weekends in white-collar workers in comparison with blue-collar workers. J Occup Health 2021 Jan;63(1):e12246 [FREE Full text] [doi: 10.1002/1348-9585.12246] [Medline: 34275174]

XSL•FO
RenderX

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
MET: metabolic equivalent
MVPA: moderate-to-vigorous physical activity

Edited by L Buis; submitted 11.12.21; peer-reviewed by C Friel, S De La Torre, N Maglaveras; comments to author 08.04.22; revised version received 31.05.22; accepted 16.09.22; published 24.10.22.

Please cite as:
Yoshimura E, Tajiri E, Michiwaki R, Matsumoto N, Hatamoto Y, Tanaka S
Long-term Effects of the Use of a Step Count–Specific Smartphone App on Physical Activity and Weight Loss: Randomized Controlled Clinical Trial
JMIR Mhealth Uhealth 2022;10(10):e35628
URL: https://mhealth.jmir.org/2022/10/e35628
doi:10.2196/35628
PMID:36279159

©Eiichi Yoshimura, Eri Tajiri, Ryota Michiwaki, Naoyuki Matsumoto, Yoichi Hatamoto, Shigeho Tanaka. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 24.10.2022. 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.

XSL•FO
RenderX

Corrigenda and Addenda

Correction: Smartphone Apps for Patients With Hematologic Malignancies: Systematic Review and Evaluation of Content

Nerea Báez Gutiérrez[1], PharmD; Héctor Rodríguez Ramallo[2], PharmD; Marcos Fernández González[2], PharmD; Laila Abdel-Kader Martín[2], PhD

[1]Hospital Universitario Reina Sofía, Córdoba, Spain

[2]Hospital Universitario Virgen del Rocío, Sevilla, Spain

Corresponding Author:
Héctor Rodríguez Ramallo, PharmD
Hospital Universitario Virgen del Rocío
Avenida Manuel Siurot, s/n
Sevilla, 41013
Spain
Phone: 34 688 933 454
Email: hector.rodriguez.sspa@juntadeandalucia.es

Related Article:

Correction of: https://mhealth.jmir.org/2022/9/e35851

(JMIR Mhealth Uhealth 2022;10(10):e43412) doi:10.2196/43412

In "Smartphone Apps for Patients With Hematologic Malignancies: Systematic Review and Evaluation of Content" (JMIR Mhealth Uhealth 2022;10(9):e35851), the authors noted one error.

In the originally published article, the article type appeared inadvertently as "Review."

In the corrected version, the article type appears correctly as "Original Paper."

The correction will appear in the online version of the paper on the JMIR Publications website on October 13, 2022, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 11.10.22; this is a non–peer-reviewed article;accepted 11.10.22; published 13.10.22.

Please cite as:
Báez Gutiérrez N, Rodríguez Ramallo H, Fernández González M, Abdel-Kader Martín L
Correction: Smartphone Apps for Patients With Hematologic Malignancies: Systematic Review and Evaluation of Content
JMIR Mhealth Uhealth 2022;10(10):e43412
URL: https://mhealth.jmir.org/2022/10/e43412
doi:10.2196/43412
PMID:36229033

©Nerea Báez Gutiérrez, Héctor Rodríguez Ramallo, Marcos Fernández González, Laila Abdel-Kader Martín. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 13.10.2022. 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.

XSL•FO
RenderX

Original Paper

Associations Between Depression Symptom Severity and Daily-Life Gait Characteristics Derived From Long-Term Acceleration Signals in Real-World Settings: Retrospective Analysis

Yuezhou Zhang[1], MSc; Amos A Folarin[1,2,3,4,5], PhD; Shaoxiong Sun[1], PhD; Nicholas Cummins[1], PhD; Srinivasan Vairavan[6], PhD; Linglong Qian[1], MSc; Yatharth Ranjan[1], MSc; Zulqarnain Rashid[1], PhD; Pauline Conde[1], BSc; Callum Stewart[1], MSc; Petroula Laiou[1], PhD; Heet Sankesara[1], BSc; Faith Matcham[7,8], PhD; Katie M White[7], BSc; Carolin Oetzmann[7], MSc; Alina Ivan[7], MSc; Femke Lamers[9,10], PhD; Sara Siddi[11,12,13], PhD; Sara Simblett[14], PhD; Aki Rintala[15,16], PhD; David C Mohr[17], PhD; Inez Myin-Germeys[15], PhD; Til Wykes[14,18], PhD; Josep Maria Haro[11,12,13], PhD; Brenda W J H Penninx[9,10], PhD; Vaibhav A Narayan[6], PhD; Peter Annas[19], PhD; Matthew Hotopf[3,7,18], PhD; Richard J B Dobson[1,2,3,4,5], PhD; RADAR-CNS Consortium

[1]Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[2]Institute of Health Informatics, University College London, London, United Kingdom

[3]NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust, London, United Kingdom

[4]Health Data Research UK London, University College London, London, United Kingdom

[5]NIHR Biomedical Research Centre at University College London Hospitals, NHS Foundation Trust, London, United Kingdom

[6]Janssen Research and Development LLC, Titusville, NJ, United States

[7]Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[8]School of Psychology, University of Sussex, Falmer, United Kingdom

[9]Department of Psychiatry, Amsterdam UMC location Vrije Universiteit, Amsterdam, Netherlands

[10]Mental Health Program, Amsterdam Public Health Research Institute, Amsterdam, Netherlands

[11]Teaching Research and Innovation Unit, Parc Sanitari Sant Joan de Déu, Fundació Sant Joan de Déu, Barcelona, Spain

[12]Centro de Investigación Biomédica en Red de Salud Mental, Madrid, Spain

[13]Faculty of Medicine and Health Sciences, Universitat de Barcelona, Barcelona, Spain

[14]Department of Psychology, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[15]Department of Neurosciences, Center for Contextual Psychiatry, Katholieke Universiteit Leuven, Leuven, Belgium

[16]Faculty of Social Services and Health Care, LAB University of Applied Sciences, Lahti, Finland

[17]Center for Behavioral Intervention Technologies, Department of Preventive Medicine, Northwestern University, Chicago, IL, United States

[18]South London and Maudsley NHS Foundation Trust, London, United Kingdom

[19]H Lundbeck A/S, Copenhagen, Denmark

Corresponding Author:
Yuezhou Zhang, MSc
Department of Biostatistics & Health Informatics
Institute of Psychiatry, Psychology and Neuroscience
King's College London
SGDP Centre, IoPPN
PO Box 80, De Crespigny Park, Denmark Hill
London, SE5 8AF
United Kingdom
Phone: 44 75 7985 6617
Email: yuezhou.zhang@kcl.ac.uk

Abstract

XSL•FO
RenderX

Background: Gait is an essential manifestation of depression. However, the gait characteristics of daily walking and their relationships with depression have yet to be fully explored.

Objective: The aim of this study was to explore associations between depression symptom severity and daily-life gait characteristics derived from acceleration signals in real-world settings.

Methods: We used two ambulatory data sets (N=71 and N=215) with acceleration signals collected by wearable devices and mobile phones, respectively. We extracted 12 daily-life gait features to describe the distribution and variance of gait cadence and force over a long-term period. Spearman coefficients and linear mixed-effects models were used to explore the associations between daily-life gait features and depression symptom severity measured by the 15-item Geriatric Depression Scale (GDS-15) and 8-item Patient Health Questionnaire (PHQ-8) self-reported questionnaires. The likelihood-ratio (LR) test was used to test whether daily-life gait features could provide additional information relative to the laboratory gait features.

Results: Higher depression symptom severity was significantly associated with lower gait cadence of high-performance walking (segments with faster walking speed) over a long-term period in both data sets. The linear regression model with long-term daily-life gait features (R^2=0.30) fitted depression scores significantly better (LR test P=.001) than the model with only laboratory gait features (R^2=0.06).

Conclusions: This study indicated that the significant links between daily-life walking characteristics and depression symptom severity could be captured by both wearable devices and mobile phones. The daily-life gait patterns could provide additional information for predicting depression symptom severity relative to laboratory walking. These findings may contribute to developing clinical tools to remotely monitor mental health in real-world settings.

(JMIR Mhealth Uhealth 2022;10(10):e40667) doi:10.2196/40667

KEYWORDS

depression; gait; mobile health; mHealth; acceleration signals; monitoring; wearable devices; mobile phones; mental health

Introduction

Depression affects the lives of over 300 million people worldwide [1] and is associated with many adverse outcomes, including decreased quality of life, loss of occupational function, disability, premature mortality, and suicide [2-5]. While early treatment can be effective and prevent more serious adverse outcomes [6], more than half of depressed people do not receive timely treatment [7,8]. Current questionnaire-based depression assessments may be affected by recall bias and may not be able to collect dynamic information [9,10]. Therefore, several recent studies have attempted to explore the associations between depression and changes in individuals' behaviors using mobile technologies [11].

Changes in gait are essential manifestations of depression [12,13]. The main hypothesis linking gait with depression is a bidirectional interaction between the brain motor system and cortical and subcortical structures, which are related to emotions and cognitive functions [14-16]. Many studies have explored the relationships between depression and gait characteristics based on "gold-standard" laboratory walking tests. Longer gait cycles, reduced stride length, and slower gait cadence were observed in participants with depression compared with healthy controls, which have been consistently shown in several studies [17-25]. Other gait abnormalities such as reduced gait force [21], increased double support time [22], reduced swing time variability [23], slumped postures [24], and increased body sway [25] have been reported, but with less consistency across studies.

Laboratory gait tests are hard to be applied in real-world settings because of the need for expensive equipment (eg, video camera and force plates), specialized laboratories, and the inconvenience of wearing sensors on the knees and ankles, for example [14,26]. Some researchers have suggested that people's daily-life activity characteristics should have stronger links to their health conditions than laboratory tests [27-29]. Therefore, it is necessary to monitor and evaluate daily-life walking using efficient methods.

In recent years, several studies have used mobile technologies to measure daily-life walking patterns and explored their associations with depression. However, most of these studies only measured the number of cumulative steps of daily-life walking [30-32], which is more related to individuals' mobility and physical activity than to gait patterns (eg, gait cadence and gait force). To our knowledge, there have been only a few studies exploring the associations between daily-life gait patterns and depression directly. Adolph et al [33] found that depressed participants had reduced walking speed, reduced vertical up-and-down movements, and more slumped postures compared with controls by placing two accelerometers on the participant's trunk and right leg for 2 days [33]. However, wearing multiple sensors on the body may not be suitable for long-term monitoring. With the development of sensors, the mobile phone provides a cost-effective, continuous, and unobtrusive means to measure individuals' behaviors, including daily walking. Therefore, the mobile phone may be a potential tool for long-term gait monitoring.

The aim of this study was to explore the value of daily-walking monitoring for improving the evaluation of depression symptom severity. Our first objective was to design and extract gait features from raw acceleration signals to describe the characteristics of daily walking. The second objective was to explore the associations between gait features and depression symptom severity, and to test whether these associations could be captured by different acceleration devices. The third objective

XSL•FO
RenderX

was to test whether daily-life walking could provide additional information for predicting depression relative to laboratory walking. To achieve the second and third objectives, we performed our analyses on two ambulatory data sets, the Long Term Movement Monitoring (LTMM) and Remote Assessment of Disease and Relapse–Major Depressive Disorder (RADAR-MDD) data sets [34,35], with acceleration signals collected by a wearable device and mobile phone, respectively. Importantly, the LTMM data set contains data related to both laboratory and daily walking, which could address the third study objective.

Methods

Data Sets

LTMM Data Set

The LTMM data set includes demographics (age and gender), depression scores (15-item Geriatric Depression Scale [GDS-15] [36]), and raw acceleration signals (100 Hz) of laboratory walking tests and 3-day activities for 71 elderly adults [34], which can be downloaded at PhysioNet [37]. Participants were included if they did not have any cognitive or gait/balance disorders [34]. Participants were asked to walk at a self-selected and comfortable speed for 1 minute in the laboratory while wearing a 3-axis accelerometer on their lower back [34]. The GDS-15 questionnaire contains 15 easy-to-understand, yes/no format questions, which is suitable for depression screening in the older population [38,39]. After the laboratory walking test, all participants were asked to wear the accelerometer for the next 3 consecutive days to record daily activities [34].

Ethics Considerations

RADAR-MDD was conducted per the Declaration of Helsinki and Good Clinical Practice, adhering to principles outlined in the National Health Service (NHS) Research Governance Framework for Health and Social Care (2nd edition). Ethical approval has been obtained in London from the Camberwell St Giles Research Ethics Committee (REC reference 17/LO/1154), in Spain from the CEIC Fundació Sant Joan de Deu (CI PIC-128-17), and in the Netherlands from the Medische Ethische Toetsingscommissie VUms (METc VUmc registratienummer 2018.012–NL63557.029.17).

RADAR-MDD Data Set

The EU research program RADAR-MDD aimed to investigate the utility of mobile technologies for the long-term monitoring of participants with depression in real-world settings [35,40]. Adult participants with a depression history were included in the study if they did not meet the following criteria: (1) have other psychiatric disorders (eg, bipolar disorder, schizophrenia, and dementia), (2) have received treatment for drug or alcohol use in the past 6 months, (3) a major medical diagnosis that affects daily activities, and (4) pregnancy [35]. A detailed study protocol was published previously [35]. In this study, we used a subset of RADAR-MDD data collected from a study site in the United Kingdom (King's College London [KCL]) between November 2017 and April 2021, because the KCL site was the only site to acquire ethical approval for collecting the phone's acceleration signals. We hereafter denote this subset as the RADAR-MDD-KCL data set for convenience. The phone's acceleration signals were collected at 50 Hz and uploaded to an open-source platform, RADAR-base [41]. The participants' depression symptom severity was assessed by the 8-item Patient Health Questionnaire (PHQ-8) [42] self-reported through mobile phones every 2 weeks. A patient advisory board comprising service users co-developed the study. They were involved in the choice of measures, timing, and issues of engagement, and have also been involved in developing the analysis plan.

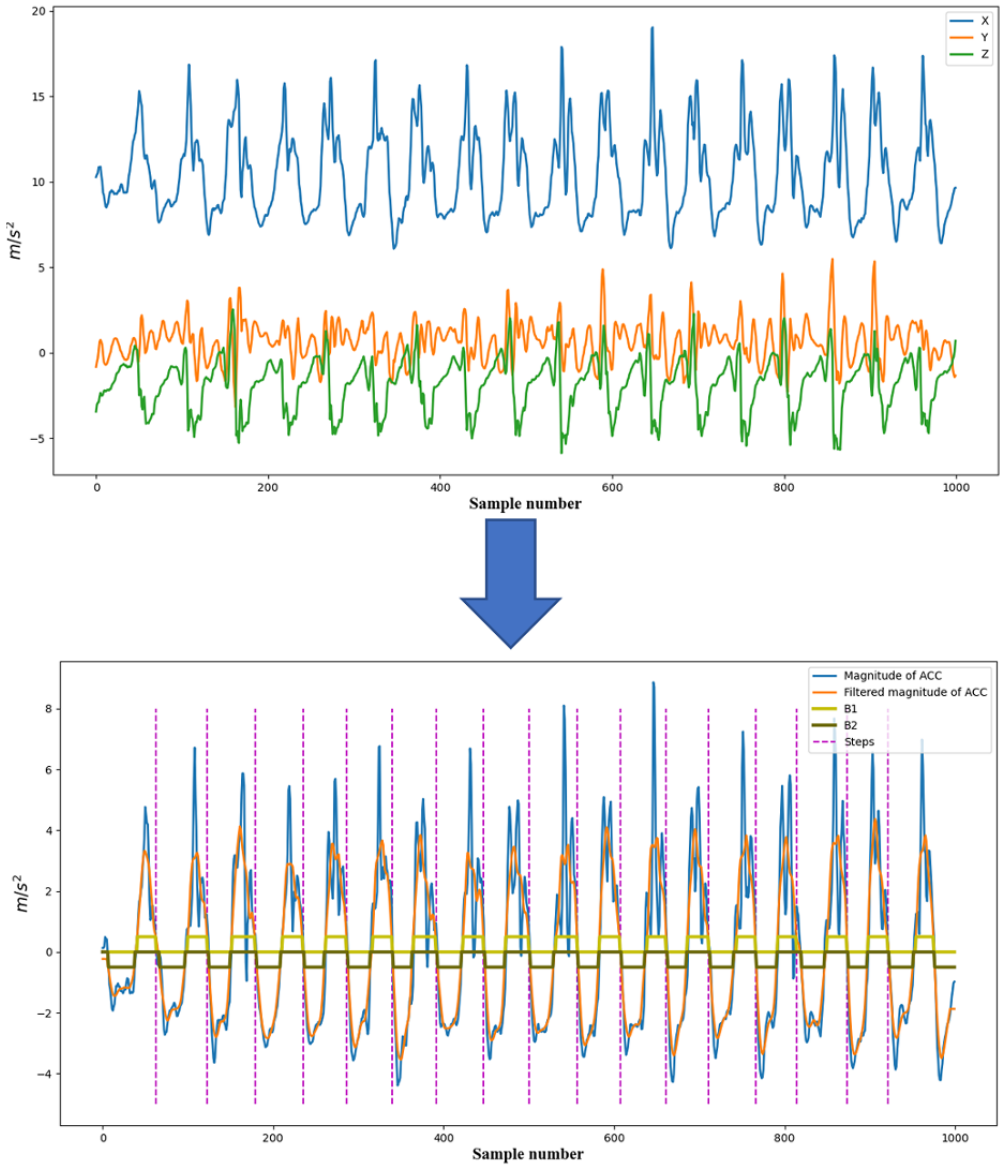
Step Detection Algorithm

Since we needed to respectively detect steps on the acceleration signals collected by wearable devices and mobile phones, we chose to use the step detection algorithm [43], which was based on mobile phones (Figure 1). Given a segment of 3-axis acceleration signals (x_i, y_i, z_i), the magnitude of the acceleration of the segment of acceleration signals was calculated to combine 3D signals to a single series, r_i, where  . The magnitude of the acceleration signals does not depend on the orientation and tilt of the mobile phone during walking [43]. Subsequently, r_i was filtered by a weighted moving-average filter to remove noise (Equation 1, w=150 milliseconds). Next, the filtered r_i was subtracted by the mean of $\bar{r}_i$ to make $\bar{r}_i$ symmetric to the x-axis. We calculated two new series, $B1_i$ and $B2_i$, based on two thresholds to detect the walking swing phase and stance phase, respectively (see Equations 2 and 3). If a swing phase ends and a stance phase starts, we can identify a step that occurred. The formal detection rule of a step S_i at sample i is that the following two conditions must be satisfied: (1) a change from –0.5 to 0 in B1 ($B1_i$=0 and $B1_{i-1}$=0.5); (2) there is at least one detection of $B2$=–0.5 in a window of size w=150 milliseconds in sample i ($Min(B2_{i:i+w})$=–0.5).

Then, the gait cycle series could be derived by calculating time intervals between consecutive steps, which was denoted as *Cycles*. During each gait cycle, the amplitude from the peak to the valley of the magnitude of the acceleration signals was used to reflect the gait force of each step. The force of all steps in the given acceleration signal was denoted as the series *Force*.

XSL•FO
RenderX

Figure 1. Step detection algorithm. ACC is the 3-axis acceleration signal; B1 and B2 are two series calculated by thresholds to detect walking swing and stance phase, respectively; and pink dashed lines represent the detected steps.



Feature Extraction

Feature Window Size

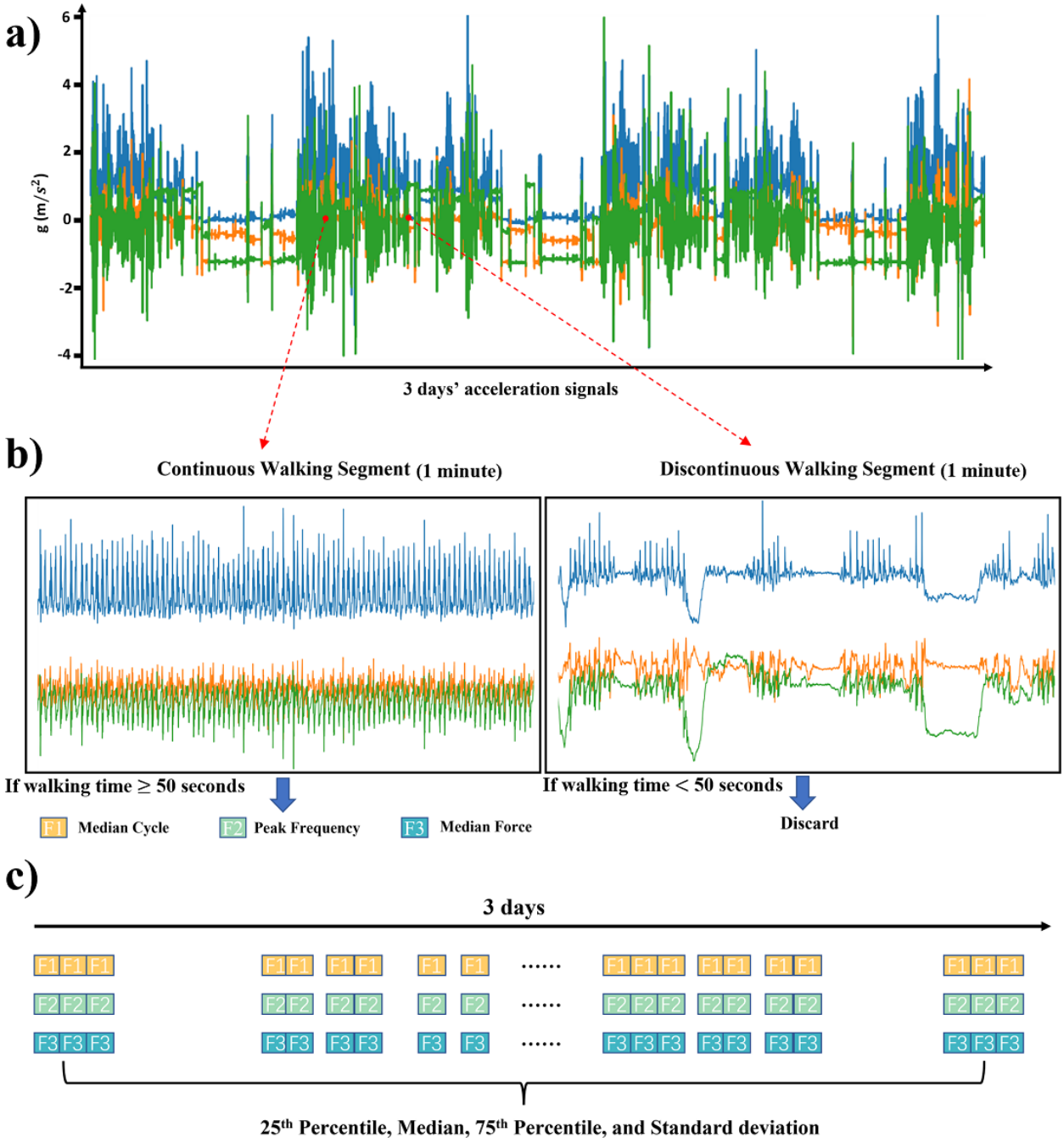
Since the PHQ-8 score is used to estimate depression symptom severity for the past 2 weeks [42], we extracted gait features from a 14-day time window prior to each PHQ-8 record from the RADAR-MDD-KCL data set. For the LTMM data set, we extracted gait features from 3-day activities to link daily-life walking with the GDS-15 score.

Step Detection Window and the Continuous Walking Segment

Daily-life walking in real-world settings is complex and contains some intermittent walking segments (such as walking in a crowded environment or a walking-rest transition status). These intermittent walking segments may not fully reflect a participant's normal walking patterns. Therefore, to distinguish between continuous and intermittent walking, we used a 1-minute sliding window [44] to detect steps from the long-term raw acceleration signals. If the participant was walking most of the time in this minute, we considered this minute as the continuous walking segment. Based on our experience, we set 50 seconds as the threshold for selecting the continuous walking segment; that is, the segment with more than 50 seconds of walking time (sum of all gait cycles in the minute) was selected for further analysis (Figure 2b).

XSL•FO
RenderX

Figure 2. Schematic diagram of long-term gait feature extraction for the Long-Term Movement Monitoring data set. (a) Three-axis acceleration signals of 3 consecutive days; (b) examples of continuous and discontinuous walking segments and three short-term gait features (definitions in Table 1) were extracted from each continuous walking segment; (c) long-term gait feature extraction: 25th percentile, median, 75th percentile, and standard deviation of short-term gait feature values of all continuous walking segments over 3 days for each participant.



Gait Features

Overview

The performance of walking varies over time due to several factors such as mood, energy, and environment. Therefore, the long-term gait features need to represent the distribution and variance of walking patterns over the feature window. We first extracted three short-term gait features from every detected continuous walking segment in the feature window. Then, for each short-term gait feature, we calculated four statistical second-order features (long-term features) across all values of continuous walking segments. In total, 12 long-term gait features were extracted in this study, and a summary of these features is shown in Table 1. A schematic diagram of long-term gait feature extraction is shown in Figure 2.

XSL•FO
RenderX

Table 1. Short-term and long-term gait features extracted and their short descriptions.

Gait feature	Description
Short-term gait features	
Median cycle (seconds)	Median of gait cycles in the 1-minute walking segment
Peak frequency (Hz)	Peak frequency in the PSD[a] of the magnitude of 1-minute acceleration signals
Median force (m/s^2)	Median of gait force in the 1-minute walking segment
Long-term gait features	
25th percentile of median cycle	25th percentile of median gait cycle values of all walking segments[b]
50th percentile of median cycle	Median of median gait cycle values of all walking segments
75th percentile of median cycle	75th percentile of median gait cycle values of all walking segments
SD of median cycle	Standard deviation of median gait cycle values of all walking segments
25th percentile of peak frequency	25th percentile of peak frequency values of all walking segments
50th percentile of peak frequency	Median of peak frequency values of all walking segments
75th percentile of peak frequency	75th percentile of peak frequency values of all walking segments
SD of peak frequency	Standard deviation of peak frequency values of all walking segments
25th percentile of median force	25th percentile of median gait force values of all walking segments
50th percentile of median force	Median of median gait force values of all walking segments
75th percentile of median force	75th percentile of median gait force values of all walking segments
SD of median force	Standard deviation of median gait force values of all walking segments

[a]PSD: power spectral density (from 0.5 Hz to 3 Hz).

[b]All detected continuous walking segments (defined in the Methods section) in a feature window (3 days for the Long Term Movement Monitoring data set and 14 days for the Remote Assessment of Disease and Relapse–Major Depressive Disorder data set).

Short-Term Gait Features From the 1-Minute Continuous Walking Segment

Gait cadence and gait force are essential characteristics of walking. Gait cadence is the rate at which the individual feet contact the ground [45]. Gait force reflects the ground reaction force during walking [46]. For every continuous walking segment, the median of the gait cycle series (*Cycles*) was calculated to reflect the gait cadence of this minute from the time domain, which was denoted as *median cycle*. To assess the gait cadence from the frequency domain, the power spectral density (PSD) of walking was obtained by applying the fast Fourier transformation to the filtered magnitude ($\bar{r}_i$) of the acceleration signals of every continuous walking segment. The peak frequency [47] of the 0.5-3–Hz band (reflecting walking) [34] of the PSD was used to reflect the main rhythm of steps from the frequency domain, which was denoted as *peak frequency*. For gait force, we calculated the median of the *Force* series (*median force*) to represent the average power of all steps in the minute.

Long-Term Gait Features

For each of the short-term gait features (*median cycle, peak frequency,* and *median force*), we calculated four statistical second-order features (25th percentile, median, 75th percentile, and SD) from all detected continuous walking segments during a feature window.

Previous studies suggested that the extreme values of gait characteristics over the long term could reflect the optimal or worst walking performance of the participant, which could in turn reflect physical or mental conditions better than the median value [29]. Therefore, we used *25th percentile*, *median*, and *75th percentile* second-order statistics to represent three levels of walking performance (low, medium, and high) during a feature window. For example, faster walking during a feature window could represent *high-performance walking*, which may not be affected by other factors such as fatigue and the crowded environment. *High-performance walking* could be represented by the *75th percentile of peak frequency* and the *25th percentile of median cycle* in a feature window, which is expected to be closely associated with depression status. The variance of daily-life walking in a feature window was measured by the SD.

Laboratory Gait Features Extracted From Laboratory Walking Tests in the LTMM Data Set

We also extracted *median cycle, peak frequency,* and *median force* from the 1-minute acceleration signals of laboratory walking tests in the LTMM data set. For reading convenience, we denoted these as laboratory gait features.

Inclusive Criteria for Data Missingness in the RADAR-MDD-KCL Data Set

The raw acceleration signals were remotely collected by mobile phones in the RADAR-MDD-KCL study. Possibly due to the high battery consumption and network traffic for uploading the raw signal, the missing rate of acceleration signals was relatively high. To reduce the impact of missingness, a PHQ-8 period (14

XSL•FO
RenderX

days) included in this study should have at least 3 days (aligned with the LTMM data set) with more than 50% acceleration signals [48,49].

Statistical Analyses

For the LTMM data set, Spearman coefficients [50] were calculated to assess associations between the GDS-15 score and gait features (3 laboratory gait features and 12 long-term gait features). As the data in the RADAR-MDD-KCL data set are longitudinal (repeated PHQ-8 measurements for each participant), a series of pairwise linear mixed-effects regression models [51] with random participant intercepts were performed to explore the association between the PHQ-8 score and each of the 12 long-term gait features (no laboratory tests were included in the RADAR-MDD-KCL data set). Age, gender, and the number of comorbidities (see Table S1 in Multimedia Appendix 1) were considered as covariates. The Benjamini-Hochberg method was used for multiple-comparison corrections in both data sets [52].

To test whether long-term gait features could explain additional data variance in depression scores relative to laboratory gait features, we built two nested multivariate linear regression models without and with long-term gait features for the GDS-15 score (denoted as Model A and Model B; Equations 4 and 5) in the LTMM data set. Specifically, predictors of Model A are age, gender, and the 3 laboratory gait features, while predictors of Model B are age, gender, the 3 laboratory gait features, and the 12 long-term gait features. The coefficient of determination (R^2) was calculated for both models to estimate how much data variance was explained by predictors. Then, the likelihood ratio test [53] was used to test whether Model B fit the GDS-15 score better than Model A. Since the laboratory walking test was not included in the RADAR-MDD-KCL data set, the likelihood ratio test was only performed in the LTMM data set.

Model A: GDS-15=Age+Gender+3 laboratory gait features **(4)**

Model B: GDS-15=Age+Gender+3 laboratory gait features+12 long-term gait features **(5)**

Results

Data Summary

The 71 participants in the LTMM data set have a mean age of 78.36 (SD 4.71) years with 18 (25%) participants having potential depressive disorders (GDS-15≥5) and 69.82 (SD 9.65) hours of acceleration signals per participant. The RADAR-MDD-KCL data set, according to the data inclusion criteria, contains 659 PHQ-8 records collected from 215 participants and corresponding 99,445 hours (average 463 hours per participant). The cohort in the RADAR-MDD-KCL data set has a mean age of 43.36 (SD 15.12) years with the majority being women (75%), and half of the PHQ-8 records indicated potential depression symptoms (PHQ-8≥10). The average missing rate of acceleration signals collected by phones in the RADAR-MDD-KCL data set (70.60%) was significantly higher than that of the acceleration signals collected by the wearable device in the LTMM data set (3.03%). A summary of the demographics, and distributions of depression scores and available acceleration signals for participants in the LTMM and the RADAR-MDD-KCL data sets is shown in Table 2. The heatmaps of correlations between the 12 long-term gait features of the LTMM and RADAR-MDD-KCL data sets are presented in Figure 3.

XSL•FO
RenderX

Table 2. Demographics and distributions of depression scores and available acceleration signals of participants in the two data sets.

Characteristic	LTMM[a] (N=71)	RADAR-MDD-KCL[b] (N=215)
Age (years), mean (SD)	78.36 (4.71)	43.36 (15.12)
Female, n (%)	46 (65%)	162 (75%)
Depression score, mean (SD)	GDS-15[c]: 3.18 (2.81)	PHQ-8[d]: 9.67 (5.84)
Potential depressive episode (GDS-15≥5) and PHQ-8≥10), n (%)[e]	18 (25%)	330 (50%)
Number of completed depression questionnaires[f]	71	659
Number of completed depression questionnaires per participant, mean (SD)	1 (0)	3.09 (2.76)
Length of total available acceleration signals (hours)	4817	99,445
Length of available acceleration signals (hours) for each GDS-15/PHQ-8 record[g], mean (SD)	69.82 (9.65)	98.77 (105.20)
Average missing rate of acceleration signals (%)	3.03	70.60
Number of continuous walking segments[h] detected from each GDS-15/PHQ-8 record, mean (SD)	73.48 (66.98)	113.24 (170.48)

[a]LTMM: Long Term Movement Monitoring.

[b]RADAR-MDD-KCL: subset of the Remote Assessment of Disease and Relapse–Major Depressive Disorder data set collected from King's College London, United Kingdom.

[c]GDS-15: 15-item Geriatric Depression Scale.

[d]PHQ-8: 8-item Patient Health Questionnaire.

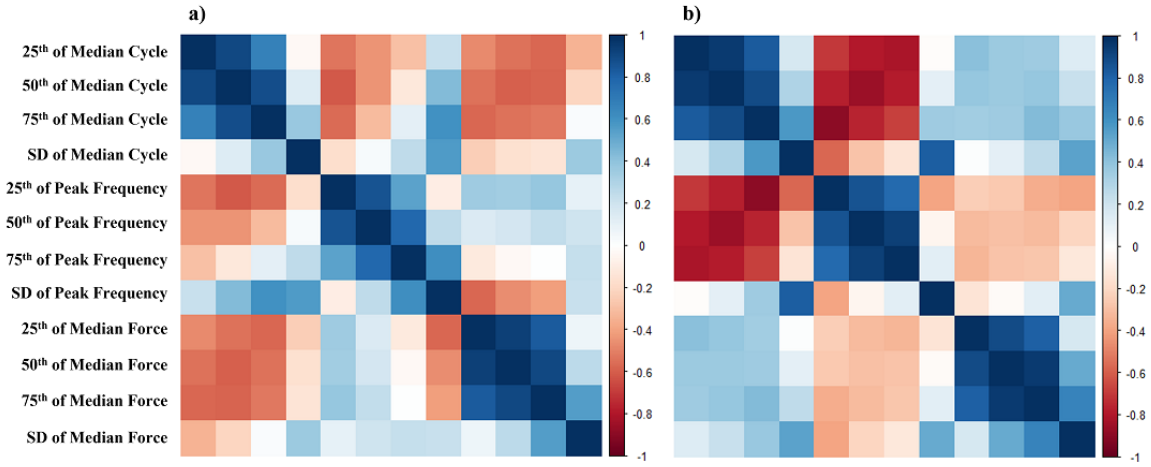
[e]Based on the total number of completed questionnaires.

[f]The RADAR-MDD-KCL data set has multiple PHQ-8 records for each participant, which was conducted every 2 weeks.

[g]We regarded acceleration signals in the 14 days before a PHQ-8 record. For the GDS-15 record, we considered acceleration signals of all 3-day activities after enrollment.

[h]Continuous walking segment was defined as 1-minute acceleration signals with at least 50 seconds of walking (see Methods section).

Figure 3. Heatmaps of correlations between 12 long-term gait features of the Long-term Movement Monitoring data set (a) and Remote Assessment of Disease and Relapse–Major Depressive Disorder King's College London data set (b).



Associations Between Gait Features and the GDS-15 Score in the LTMM Data Set

The Spearman correlations between the GDS-15 score and gait features (both laboratory and long-term gait features) in the LTMM data set are shown in Table 3. We found that a higher GDS-15 score was significantly correlated with a larger median of gait cycles, lower peak frequency, and smaller median gait force in the 1-minute laboratory walking test. For the long-term period, a higher GDS-15 score was significantly correlated with lower variance of gait force and slower cadence of high-performance walking and 75th percentile of peak frequency during 3-day activities.

XSL•FO
RenderX

Table 3. Spearman correlations between the 15-item Geriatric Depression Scale score and gait features, including laboratory and long-term gait features, in the Long-Term Movement Monitoring data set.

Feature[a]	ρ	*P* value[b]
Laboratory gait features extracted from the 1-minute laboratory walking test		
Median cycle	0.39	.001
Peak frequency	–0.32	.01
Median force	–0.25	.04
Long-term gait feature extracted from 3-day activities		
25th percentile of median cycle	0.31	.01
50th percentile of median cycle	0.13	.29
75th percentile of median cycle	0.02	.86
SD of median cycle	–0.24	.06
25th percentile of peak frequency	–0.02	.85
50th percentile of peak frequency	–0.09	.45
75th percentile of peak frequency	–0.27	.03
SD of peak frequency	–0.12	.33
25th percentile of median force	0.02	.85
50th percentile of median force	–0.01	.98
75th percentile of median force	–0.10	.41
SD of median force	–0.30	.02

[a]Definitions of gait features in this table are provided in Table 1 and the Methods section.

[b]*P* values were adjusted by the Benjamini-Hochberg method for correction of multiple comparisons.

Associations Between Long-Term Gait Features and the PHQ-8 Score in the RADAR-MDD-KCL Data Set

The pairwise linear mixed-effects models performed in the RADAR-MDD-KCL data set revealed a significant and negative link between the PHQ-8 score and the gait cadence of *high-performance walking* during the 14 days before submitting PHQ-8 records. Specifically, the *25th percentile of median cycle* was positively associated with the PHQ-8 score; that is, for every increase of 0.1 seconds in the median gait cycle of *high-performance walking*, the PHQ-8 score increased by 0.606 points. Likewise, the *75th percentile of peak frequency* was negatively associated with the PHQ-8 score, indicating that a reduction of 0.1 Hz in the peak frequency of *high-performance walking* was associated with an increase of 0.26 PHQ-8 points. Other long-term gait features were not found to be significantly associated with the PHQ-8 score in the RADAR-MDD-KCL data set. A summary of all 12 linear mixed-effects regression models is provided in Table 4.

XSL•FO
RenderX

Table 4. Twelve pairwise linear mixed-effects models for exploring associations between long-term gait features and depression symptom severity (8-item Patient Health Questionnaire) in the RADAR-MDD-KCL data set.[a]

Long-term gait feature[b]	Estimate	SE	*df*	*t* value	*P* value[c]
25th percentile of median cycle	6.06	2.72	648.75	2.23	.03
50th percentile of median cycle	3.98	2.51	639.41	1.59	.11
75th percentile of median cycle	2.49	2.08	653.72	1.20	.23
STD of median cycle	2.87	4.41	631.11	0.65	.52
25th percentile of peak frequency	–1.50	1.02	656.44	–1.46	.15
50th percentile of peak frequency	–1.93	1.05	650.76	–1.83	.07
75th percentile of peak frequency	–2.62	1.01	634.70	–2.60	.01
SD of peak frequency	0.21	1.86	600.50	0.12	.91
25th percentile of median force	–0.57	2.24	637.46	–0.25	.80
50th percentile of median force	0.88	1.79	655.77	0.49	.62
75th percentile of median force	0.44	1.66	656.37	0.26	.79
SD of median force	2.05	3.78	602.90	0.54	.59

[a]RADAR-MDD-KCL: Subset of Remote Assessment of Disease and Relapse–Major Depressive Disorder collected from King's College London.

[b]Definitions of daily-life gait features are provided in Table 1 and the Methods section.

[c]*P* values were adjusted by the Benjamini-Hochberg method for correction of multiple comparisons.

Results of the Likelihood Ratio Test in the LTMM Data Set

The regression model with long-term gait features (Model B) achieved better performance (R^2=0.30) than the model without long-term gait features (Model A) (R^2=0.06). We found that the 12 long-term gait features extracted from 3-day activities could explain an extra 24% data variance (an increase of 0.24 in R^2) of GDS-15 scores relative to the laboratory gait features and participants' demographics. The likelihood ratio test showed that Model B fitted GDS-15 scores significantly better than Model A (χ^2=32.91>$\chi^2_{0.05}$(12), *P*=.001). The detailed results of the two nested regression models are shown in Table S2 of Multimedia Appendix 1.

Discussion

Principal Findings

This study retrospectively used two ambulatory data sets for exploring the associations between depression symptom severity and daily-life gait characteristics. We extracted 12 long-term gait features to describe the distribution and variance of gait cadence and force over a long-term period and link daily-life gait patterns with a self-reported depression score. The main findings of this study are (1) higher depression symptom severity is significantly associated with lower gait cadence of *high-performance walking* (faster walking in all continuous walking segments) over a long-term period; (2) long-term daily-life walking has the potential to provide additional information for predicting depression symptom severity relative to laboratory gait characteristics and demographics; and (3) wearable devices and mobile phones both have potential to capture the associations between daily gait and depression.

The results of Spearman correlations between laboratory gait features and the GDS-15 score in the LTMM data set are consistent with previous studies [17-25]; that is, the participants with more severe depression symptoms were more likely to have slower gait cadence (longer median of gait cycles and lower gait frequency) and smaller gait force in laboratory walking tests.

For daily-life walking, this study used the faster walking (*75th percentile of peak frequency* and *25th percentile of median cycle*) in all detected continuous walking segments to represent *high-performance walking* during a feature window (3 days for LTMM and 14 days for RADAR-MDD-KCL). Only gait cadence of *high-performance walking* was found to be significantly and negatively associated with depression symptom severity, whereas gait patterns under *medium/low-performance walking* were not significantly associated with the depression score. This finding was consistent in both the LTMM and RADAR-MDD-KCL data sets. A potential reason is that the walking performance in real-world scenarios may be affected by multiple factors (such as walking during the day or at night, walking under fatigue or walking after rest, and walking to a destination or navigating a crowded supermarket) [29]; therefore, the lower walking performance may not fully reflect the participant's physical or mental conditions. Therefore, from the main finding of this study, we speculated that faster steps over a long-term period could represent the high performance of participants' walking, which could be closely associated with their depression status.

In the LTMM data set, we found that the variance of gait force (*SD of median force*) in 3-day activities was significantly and negatively associated with the depression symptom severity, indicating that participants with higher depression symptom severity were likely to have relatively monotonous walking over 3 days. However, the feature was not significantly associated

XSL•FO
RenderX

with the PHQ-8 score in the RADAR-MDD-KCL data set. One reason is that the magnitude (r_i) (explained in the Step Detection Algorithm section) of the acceleration signals depends on the location of the accelerometers attached to the body [54]. As acceleration signals in the RADAR-MDD-KCL data set were collected by mobile phones, the variable locations of phones when attached to participants' bodies (such as in the hand, handbag, and pocket) affected the magnitude of acceleration signals. Therefore, the magnitude of phone-collected acceleration signals cannot fully reflect the gait force.

Results of regression models and the likelihood test in the LTMM data set illustrated the importance of monitoring daily-life gait in real-world settings. Laboratory gait features and demographics in LTMM data only explained a small proportion of data variance of the GDS-15 score (R^2=0.06), whereas long-term gait features extracted from 3-day activities could explain an extra 24% of data variance (R^2=0.30). This finding supported that long-term daily-life walking has the potential to provide additional information for predicting depression symptom severity relative to laboratory gait characteristics and demographics. Further, this finding also indicated that the laboratory walking test may be affected by several factors such as subjective psychological factors and laboratory-controlled conditions, which may not fully reflect the condition of a participant's mental health [27,29]. Since there were no laboratory tests in the RADAR-MDD-KCL data set, the comparison between laboratory gait features and long-term daily-life gait features was not performed in the RADAR-MDD-KCL. We will consider adding laboratory tests at enrollment in future digital depression studies.

Limitations

Although we found that wearables and mobile phones have the potential to capture the associations between depression and daily-life gait patterns, both devices have some limitations. Wearables could collect relatively complete walking data; however, wearing sensors may not be suitable for long-term monitoring. Mobile phones could be used for long-term monitoring without user burden, but the missing rate of mobile phone acceleration signals is relatively high. The findings of this study support that the links between gait and depression could still be revealed from the limited and sparse daily-life walking acceleration signals. Missingness is a common challenge in remote digital studies [55], which may be caused by high battery consumption, network traffic for uploading the raw acceleration signals, and the Android operating system moderation of resources. According to the findings of this study, a possible solution to reduce missingness is uploading gait cycles instead of uploading raw acceleration signals in future long-term monitoring research. This is not difficult to implement, as most current smartphones have real-time step detection functions or apps [56,57]. Furthermore, the self-reported PHQ-8 data may be subject to recall bias. We may consider implementing ecological momentary assessments with passive gait data collection in future research.

The hyperparameters in step detection and feature extraction need further investigation. We considered using a 1-minute window size for step detection and 50 seconds for continuous walking segment selection based on previous studies [34,44] and our experience. The feature window sizes for the two data sets are different due to the different study designs. However, the optimal hyperparameters are still unclear and will be investigated in future research.

Gait features extracted in this study were simple and statistically based, which were used to illustrate the importance of daily walking in our initial analysis. More features such as nonlinear features will be considered in future research.

Gait characteristics could be affected by some physical diseases, neurological disorders, and age [58-60]. Although none of the participants had any cognitive or gait/balance disorders in the LTMM data set and the number of comorbidities and age were considered as covariates in the RADAR-MDD-KCL data set, physical comorbidities and other comorbidities may have different impacts on the gait characteristics. We will consider a wider range of comorbidities and investigate them further in future research.

Conclusion

In summary, the findings of this study showed that significant links between depression symptom severity and daily-life gait characteristics could be captured in different data sets and by different accelerometer devices. Long-term daily-life walking patterns could provide additional value for understanding depression manifestations relative to gait patterns in laboratory walking tests, which illustrated the importance of long-term gait monitoring. The gait cadence of high-performance walking in daily life has the potential to be an indicator for monitoring depression severity, which may contribute to developing clinical tools to remotely monitor mental health in real-world settings.

Acknowledgments

The Remote Assessment of Disease and Relapse–Central Nervous System (RADAR-CNS) project has received funding from the Innovative Medicines Initiative (IMI) 2 Joint Undertaking under grant agreement number 115902. This Joint Undertaking receives support from the European Union's Horizon 2020 Research and Innovation Program and the European Federation of Pharmaceutical Industries and Associations (EFPIA). This communication reflects the views of the RADAR-CNS consortium and neither IMI nor the European Union and EFPIA are liable for any use that may be made of the information contained herein. The funding bodies have not been involved in the design of the study, the collection or analysis of data, or the interpretation of data. This study represents independent research partly funded by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre at South London, and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health and Social

XSL•FO
RenderX

Care. We thank all the members of the RADAR-CNS patient advisory board for their contribution to the device selection procedures, and their invaluable advice throughout the study protocol design. This research was reviewed by a team with experience of mental health problems and their careers, who have been specially trained to advise on research proposals and documentation through Feasibility and Acceptability Support Team for Researchers (FAST-R), a free, confidential service in England provided by the NIHR Maudsley Biomedical Research Centre via King's College London and South London and Maudsley NHS Foundation Trust. We thank all GLAD Study volunteers for their participation, and gratefully acknowledge the NIHR BioResource, NIHR BioResource centers, NHS Trusts and staff for their contribution. We also acknowledge NIHR BRC, King's College London, South London and Maudsley NHS Trust and King's Health Partners. We thank the NIHR, NHS Blood and Transplant, and Health Data Research UK as part of the Digital Innovation Hub Program.

CO is supported by the UK Medical Research Council (MR/N013700/1) and King's College London member of the MRC Doctoral Training Partnership in Biomedical Sciences.

RJBD is supported by the following: (1) NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, United Kingdom; (2) Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation, and Wellcome Trust; (3) The BigData@Heart consortium, funded by the IMI-2 Joint Undertaking under grant agreement number 116074. This Joint Undertaking receives support from the European Union's Horizon 2020 Research and Innovation Program and EFPIA; it is chaired by DE Grobbee and SD Anker, partnering with 20 academic and industry partners and ESC; (4) the NIHR University College London Hospitals Biomedical Research Centre; (5) the NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London; (6) the UK Research and Innovation London Medical Imaging and Artificial Intelligence Centre for Value Based Healthcare; and (7) the NIHR Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust.

Conflicts of Interest

SV and VAN are employees of Janssen Research and Development LLC. PA is employed by the pharmaceutical company H Lundbeck A/S. DCM has accepted honoraria and consulting fees from Otsuka Pharmaceuticals, Optum Behavioral Health, Centerstone Research Institute, and the One Mind Foundation; has received royalties from Oxford Press; and has an ownership interest in Adaptive Health, Inc. MH is the principal investigator of RADAR-CNS, a private public precompetitive consortium that receives funding from Janssen, UCB, Lundbeck, MSD, and Biogen. The other authors have no conflicts of interest to declare.

Multimedia Appendix 1
Tables S1-S2.
[DOCX File , 22 KB - mhealth_v10i10e40667_app1.docx]

References

1. Friedrich M. Depression Is the Leading Cause of Disability Around the World. JAMA 2017 Apr 18;317(15):1517. [doi: 10.1001/jama.2017.3826] [Medline: 28418490]
2. Lenox-Smith A, Macdonald MTB, Reed C, Tylee A, Peveler R, Quail D, et al. Quality of life in depressed patients in UK primary care: The FINDER Study. Neurol Ther 2013 Dec 13;2(1-2):25-42 [FREE Full text] [doi: 10.1007/s40120-013-0006-1] [Medline: 26000214]
3. Lerner D, Adler D, Chang H, Berndt ER, Irish JT, Lapitsky L, et al. The clinical and occupational correlates of work productivity loss among employed patients with depression. J Occup Environ Med 2004 Jun;46(6 Suppl):S46-S55 [FREE Full text] [doi: 10.1097/01.jom.0000126684.82825.0a] [Medline: 15194895]
4. Lewinsohn PM, Solomon A, Seeley JR, Zeiss A. Clinical implications of "subthreshold" depressive symptoms. J Abnorm Psychol 2000 May;109(2):345-351. [Medline: 10895574]
5. Hawton K, Casañas I Comabella C, Haw C, Saunders K. Risk factors for suicide in individuals with depression: a systematic review. J Affect Disord 2013 May;147(1-3):17-28. [doi: 10.1016/j.jad.2013.01.004] [Medline: 23411024]
6. Kamphuis MH, Stegenga BT, Zuithoff NPA, King M, Nazareth I, de Wit NJ, et al. Does recognition of depression in primary care affect outcome? The PREDICT-NL study. Fam Pract 2012 Feb 22;29(1):16-23. [doi: 10.1093/fampra/cmr049] [Medline: 21859837]
7. Harman JS, Edlund MJ, Fortney JC. Disparities in the adequacy of depression treatment in the United States. Psychiatr Serv 2004 Dec;55(12):1379-1385. [doi: 10.1176/appi.ps.55.12.1379] [Medline: 15572565]
8. Young AS, Klap R, Sherbourne CD, Wells KB. The quality of care for depressive and anxiety disorders in the United States. Arch Gen Psychiatry 2001 Jan 01;58(1):55-61. [doi: 10.1001/archpsyc.58.1.55] [Medline: 11146758]
9. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc 2016 May;9:211-217. [doi: 10.2147/JMDH.S104807] [Medline: 27217764]

XSL•FO
RenderX

10. Devaux M, Sassi F. Social disparities in hazardous alcohol use: self-report bias may lead to incorrect estimates. Eur J Public Health 2016 Feb 19;26(1):129-134 [FREE Full text] [doi: 10.1093/eurpub/ckv190] [Medline: 26585784]
11. Rohani DA, Faurholt-Jepsen M, Kessing LV, Bardram JE. Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: systematic review. JMIR Mhealth Uhealth 2018 Aug 13;6(8):e165 [FREE Full text] [doi: 10.2196/mhealth.9691] [Medline: 30104184]
12. Sobin C, Sackeim HA. Psychomotor symptoms of depression. Am J Psychiatry 1997 Jan;154(1):4-17. [doi: 10.1176/ajp.154.1.4] [Medline: 8988952]
13. Schrijvers D, Hulstijn W, Sabbe BG. Psychomotor symptoms in depression: a diagnostic, pathophysiological and therapeutic tool. J Affect Disord 2008 Jul;109(1-2):1-20. [doi: 10.1016/j.jad.2007.10.019] [Medline: 18082896]
14. Deligianni F, Guo Y, Yang G. From emotions to mood disorders: a survey on gait analysis methodology. IEEE J Biomed Health Inform 2019 Nov;23(6):2302-2316. [doi: 10.1109/jbhi.2019.2938111]
15. Sanders RD, Gillig PM. Gait and its assessment in psychiatry. Psychiatry 2010 Jul;7(7):38-43 [FREE Full text] [Medline: 20805918]
16. Yogev-Seligmann G, Hausdorff JM, Giladi N. The role of executive function and attention in gait. Mov Disord 2008 Feb 15;23(3):329-42; quiz 472 [FREE Full text] [doi: 10.1002/mds.21720] [Medline: 18058946]
17. Lemke MR, Wendorff T, Mieth B, Buhl K, Linnemann M. Spatiotemporal gait patterns during over ground locomotion in major depression compared with healthy controls. J Psychiatr Res 2000;34(4-5):277-283. [doi: 10.1016/s0022-3956(00)00017-0] [Medline: 11104839]
18. Michalak J, Troje N, Heidenreich T. The effects of mindfulness-based cognitive therapy on depressive gait patterns. J Cogn Behav Psychother 2011;11(1):13-27.
19. van Iersel MB, Haitsma A, Olde Rikkert MGM, Benraad C. Quantitative gait analysis to detect gait disorders in geriatric patients with depression. J Am Geriatr Soc 2005 Aug;53(8):1441-1442. [doi: 10.1111/j.1532-5415.2005.53433_4.x] [Medline: 16078982]
20. Sanders JB, Bremmer MA, Comijs HC, Deeg DJ, Beekman AT. Gait speed and the natural course of depressive symptoms in late life; an independent association with chronicity? J Am Med Dir Assoc 2016 Apr 01;17(4):331-335. [doi: 10.1016/j.jamda.2015.11.016] [Medline: 26775581]
21. Sloman L, Pierrynowski M, Berridge M, Tupling S, Flowers J. Mood, depressive illness and gait patterns. Can J Psychiatry 1987 Apr 01;32(3):190-193. [doi: 10.1177/070674378703200306] [Medline: 3567834]
22. Radovanović S, Jovičić M, Marić NP, Kostić V. Gait characteristics in patients with major depression performing cognitive and motor tasks while walking. Psychiatry Res 2014 Jun 30;217(1-2):39-46. [doi: 10.1016/j.psychres.2014.02.001] [Medline: 24613201]
23. Brandler TC, Wang C, Oh-Park M, Holtzer R, Verghese J. Depressive symptoms and gait dysfunction in the elderly. Am J Geriatr Psychiatry 2012 May;20(5):425-432 [FREE Full text] [doi: 10.1097/JGP.0b013e31821181c6] [Medline: 21422907]
24. Michalak J, Troje N, Fischer J, Vollmar P, Heidenreich T, Schulte D. Embodiment of sadness and depression--gait patterns associated with dysphoric mood. Psychosom Med 2009 Jun;71(5):580-587. [doi: 10.1097/PSY.0b013e3181a2515c] [Medline: 19414617]
25. Pieruccini-Faria F, Muir-Hunter SW, Montero-Odasso M. Do depressive symptoms affect balance in older adults with mild cognitive impairment? Results from the "gait and brain study". Exp Gerontol 2018 Jul 15;108:106-111. [doi: 10.1016/j.exger.2018.04.004] [Medline: 29653157]
26. Wang Y, Wang J, Liu X, Zhu T. Detecting depression through gait data: examining the contribution of gait features in recognizing depression. Front Psychiatry 2021 May 7;12:661213. [doi: 10.3389/fpsyt.2021.661213] [Medline: 34025483]
27. Notthoff N, Drewelies J, Kazanecka P, Steinhagen-Thiessen E, Norman K, Düzel S, et al. Feeling older, walking slower-but only if someone's watching. Subjective age is associated with walking speed in the laboratory, but not in real life. Eur J Ageing 2018 Dec 1;15(4):425-433 [FREE Full text] [doi: 10.1007/s10433-017-0450-3] [Medline: 30532679]
28. Atrsaei A, Corrà MF, Dadashi F, Vila-Chã N, Maia L, Mariani B, et al. Gait speed in clinical and daily living assessments in Parkinson's disease patients: performance versus capacity. NPJ Parkinsons Dis 2021 Mar 05;7(1):24. [doi: 10.1038/s41531-021-00171-0] [Medline: 33674597]
29. Rispens SM, van Schooten KS, Pijnappels M, Daffertshofer A, Beek PJ, van Dieën JH. Do extreme values of daily-life gait characteristics provide more information about fall risk than median values? JMIR Res Protoc 2015 Jan 05;4(1):e4 [FREE Full text] [doi: 10.2196/resprot.3931] [Medline: 25560937]
30. McKercher CM, Schmidt MD, Sanderson KA, Patton GC, Dwyer T, Venn AJ. Physical activity and depression in young adults. Am J Prev Med 2009 Feb;36(2):161-164. [doi: 10.1016/j.amepre.2008.09.036] [Medline: 19062235]
31. Abedi P, Nikkhah P, Najar S. Effect of pedometer-based walking on depression, anxiety and insomnia among postmenopausal women. Climacteric 2015 Sep 23;18(6):841-845. [doi: 10.3109/13697137.2015.1065246] [Medline: 26100101]
32. Große J, Petzold MB, Brand R, Ströhle A. Step away from depression-Study protocol for a multicenter randomized clinical trial for a pedometer intervention during and after in-patient treatment of depression. Int J Methods Psychiatr Res 2021 Mar 12;30(1):e1862 [FREE Full text] [doi: 10.1002/mpr.1862] [Medline: 33180994]
33. Adolph D, Tschacher W, Niemeyer H, Michalak J. Gait patterns and mood in everyday life: a comparison between depressed patients and non-depressed controls. Cogn Ther Res 2021 Feb 25;45(6):1128-1140. [doi: 10.1007/s10608-021-10215-7]

XSL•FO
RenderX

34. Weiss A, Brozgol M, Dorfman M, Herman T, Shema S, Giladi N, et al. Does the evaluation of gait quality during daily life provide insight into fall risk? A novel approach using 3-day accelerometer recordings. Neurorehabil Neural Repair 2013 Oct 17;27(8):742-752. [doi: 10.1177/1545968313491004] [Medline: 23774124]
35. Matcham F, Barattieri di San Pietro C, Bulgari V, de Girolamo G, Dobson R, Eriksson H, RADAR-CNS consortium. Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-centre prospective cohort study protocol. BMC Psychiatry 2019 Feb 18;19(1):72 [FREE Full text] [doi: 10.1186/s12888-019-2049-z] [Medline: 30777041]
36. D'Ath P, Katona P, Mullan E, Evans S, Katona C. Screening, detection and management of depression in elderly primary care attenders. I: The acceptability and performance of the 15 item Geriatric Depression Scale (GDS15) and the development of short versions. Fam Pract 1994 Sep;11(3):260-266. [doi: 10.1093/fampra/11.3.260] [Medline: 7843514]
37. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000 Jun 13;101(23):E215-E220. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
38. Yesavage JA, Sheikh JI. 9/Geriatric Depression Scale (GDS). Clinical Gerontologist 2008 Oct 25;5(1-2):165-173. [doi: 10.1300/J018v05n01_09]
39. Williams EI, Wallace P. Health checks for people aged 75 and over. Occas Pap R Coll Gen Pract 1993 Apr(59):1-30 [FREE Full text] [Medline: 8199581]
40. Matcham F, Leightley D, Siddi S, Lamers F, White KM, Annas P, RADAR-CNS consortium. Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): recruitment, retention, and data availability in a longitudinal remote measurement study. BMC Psychiatry 2022 Feb 21;22(1):136 [FREE Full text] [doi: 10.1186/s12888-022-03753-1] [Medline: 35189842]
41. Ranjan Y, Rashid Z, Stewart C, Conde P, Begale M, Verbeeck D, Hyve, RADAR-CNS Consortium. RADAR-Base: open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices. JMIR Mhealth Uhealth 2019 Aug 01;7(8):e11734 [FREE Full text] [doi: 10.2196/11734] [Medline: 31373275]
42. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
43. Marron J, Labrador M, Valle A, Lanvin D, Rodriguez M. Multi sensor system for pedestrian tracking and activity recognition in indoor environments. IJAHUC 2016;23(1/2):3-23. [doi: 10.1504/IJAHUC.2016.078480]
44. Ihlen EAF, Weiss A, Helbostad JL, Hausdorff JM. The discriminant value of phase-dependent local dynamic stability of daily life walking in older adult community-dwelling fallers and nonfallers. Biomed Res Int 2015;2015:402596-402511. [doi: 10.1155/2015/402596] [Medline: 26491669]
45. Whittle M. Gait analysis. In: McLatchie GR, Lennox CME, editors. The soft tissues. Amsterdam: Elsevier; 1993:187-199.
46. Herzog W, Nigg BM, Read LJ, Olsson E. Asymmetries in ground reaction force patterns in normal human gait. Med Sci Sports Exerc 1989 Feb;21(1):110-114. [doi: 10.1249/00005768-198902000-00020] [Medline: 2927295]
47. Sun L, Zheng X, Shou H, Li J, Li Y. Quantitative prediction of channel sand bodies based on seismic peak attributes in the frequency domain and its application. Appl. Geophys 2010 Jun 10;7(1):10-17. [doi: 10.1007/s11770-010-0009-y]
48. Zhang Y, Folarin AA, Sun S, Cummins N, Ranjan Y, Rashid Z, et al. Predicting depressive symptom severity through individuals' nearby bluetooth device count data collected by mobile phones: preliminary longitudinal study. JMIR Mhealth Uhealth 2021 Jul 30;9(7):e29840 [FREE Full text] [doi: 10.2196/29840] [Medline: 34328441]
49. Saeb S, Zhang M, Kwasny MM, Karr CJ, Kording K, Mohr DC. The relationship between clinical, momentary, and sensor-based assessment of depression. Int Conf Pervasive Comput Technol Healthc 2015 Aug;2015:229-232 [FREE Full text] [doi: 10.4108/icst.pervasivehealth.2015.259034] [Medline: 26640739]
50. Spearman C. The proof and measurement of association between two things. Am J Psychol 1987;100(3/4):441. [doi: 10.2307/1422689]
51. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982 Dec;38(4):963. [doi: 10.2307/2529876]
52. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 2018 Dec 05;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]
53. Glover S, Dixon P. Likelihood ratios: a simple and flexible statistic for empirical psychologists. Psychon Bull Rev 2004 Oct;11(5):791-806. [doi: 10.3758/bf03196706] [Medline: 15732688]
54. Derawi M, Bours P. Gait and activity recognition using commercial phones. Comput Secur 2013 Nov;39:137-144. [doi: 10.1016/j.cose.2013.07.004]
55. Onnela J. Opportunities and challenges in the collection and analysis of digital phenotyping data. Neuropsychopharmacology 2021 Jan 17;46(1):45-54 [FREE Full text] [doi: 10.1038/s41386-020-0771-3] [Medline: 32679583]
56. Silva AG, Simões P, Queirós A, Rodrigues M, Rocha NP. Mobile apps to quantify aspects of physical activity: a systematic review on its reliability and validity. J Med Syst 2020 Jan 08;44(2):51. [doi: 10.1007/s10916-019-1506-z] [Medline: 31915935]
57. Stavropoulos T, Andreadis S, Mpaltadoros L, Nikolopoulos S, Kompatsiaris I. Wearable sensors and smartphone apps as pedometers in eHealth: a comparative accuracy, reliability and user evaluation. 2020 Presented at: 2020 IEEE International Conference on Human-Machine Systems (ICHMS); September 7-9, 2020; online. [doi: 10.1109/ichms49158.2020.9209441]

XSL•FO
RenderX

58. Del Din S, Godfrey A, Galna B, Lord S, Rochester L. Free-living gait characteristics in ageing and Parkinson's disease: impact of environment and ambulatory bout length. J Neuroeng Rehabil 2016 May 12;13(1):46 [FREE Full text] [doi: 10.1186/s12984-016-0154-5] [Medline: 27175731]
59. Rodgers MM, Mulcare JA, King DL, Mathews T, Gupta SC, Glaser RM. Gait characteristics of individuals with multiple sclerosis before and after a 6-month aerobic training program. J Rehabil Res Dev 1999 Jul;36(3):183-188. [Medline: 10659801]
60. Helbostad JL, Leirfall S, Moe-Nilssen R, Sletvold O. Physical fatigue affects gait characteristics in older persons. J Gerontol A Biol Sci Med Sci 2007 Sep 01;62(9):1010-1015. [doi: 10.1093/gerona/62.9.1010] [Medline: 17895440]

Abbreviations

EFPIA: European Federation of Pharmaceutical Industries and Associations
FAST-R: Feasibility and Acceptability Support Team for Researchers
GDS-15: 15-item Geriatric Depression Scale
IMI: Innovative Medicines Initiative
LR: likelihood ratio
KCL: King's College London
LTMM: Long Term Movement Monitoring
NHS: National Health Service
NIHR: National Institute of Health Research
PHQ-8: 8-item Patient Health Questionnaire
PSD: power spectral density
RADAR-CNS: Remote Assessment of Disease and Relapse–Central Nervous System
RADAR-MDD: Remote Assessment of Disease and Relapse–Major Depressive Disorder

Edited by L Buis; submitted 30.06.22; peer-reviewed by Z Yang, D Lekkas; comments to author 02.08.22; revised version received 11.08.22; accepted 26.08.22; published 04.10.22.

Please cite as:
Zhang Y, Folarin AA, Sun S, Cummins N, Vairavan S, Qian L, Ranjan Y, Rashid Z, Conde P, Stewart C, Laiou P, Sankesara H, Matcham F, White KM, Oetzmann C, Ivan A, Lamers F, Siddi S, Simblett S, Rintala A, Mohr DC, Myin-Germeys I, Wykes T, Haro JM, Penninx BWJH, Narayan VA, Annas P, Hotopf M, Dobson RJB, RADAR-CNS Consortium
Associations Between Depression Symptom Severity and Daily-Life Gait Characteristics Derived From Long-Term Acceleration Signals in Real-World Settings: Retrospective Analysis
JMIR Mhealth Uhealth 2022;10(10):e40667
URL: https://mhealth.jmir.org/2022/10/e40667
doi:10.2196/40667
PMID:36194451

©Yuezhou Zhang, Amos A Folarin, Shaoxiong Sun, Nicholas Cummins, Srinivasan Vairavan, Linglong Qian, Yatharth Ranjan, Zulqarnain Rashid, Pauline Conde, Callum Stewart, Petroula Laiou, Heet Sankesara, Faith Matcham, Katie M White, Carolin Oetzmann, Alina Ivan, Femke Lamers, Sara Siddi, Sara Simblett, Aki Rintala, David C Mohr, Inez Myin-Germeys, Til Wykes, Josep Maria Haro, Brenda W J H Penninx, Vaibhav A Narayan, Peter Annas, Matthew Hotopf, Richard J B Dobson, RADAR-CNS Consortium. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 04.10.2022. 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.

XSL•FO
RenderX

Original Paper

The Effects of Objective Push-Type Sleep Feedback on Habitual Sleep Behavior and Momentary Symptoms in Daily Life: mHealth Intervention Trial Using a Health Care Internet of Things System

Hiroki Takeuchi[1], MSc; Kaori Suwa[1], MSc; Akifumi Kishi[2,3], PhD; Toru Nakamura[4], PhD; Kazuhiro Yoshiuchi[2], MD, PhD; Yoshiharu Yamamoto[1], PhD

[1]Graduate School of Education, The University of Tokyo, Tokyo, Japan

[2]Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

[3]Precursory Research for Embryonic Science and Technology, Japan Science and Technology Agency, Saitama, Japan

[4]Graduate School of Engineering Science, Osaka University, Osaka, Japan

Corresponding Author:
Hiroki Takeuchi, MSc
Graduate School of Education
The University of Tokyo
7-3-1, Hongo
Bunkyo-ku
Tokyo, 113-8654
Japan
Phone: 81 03 5841 3968
Email: takeuchi@p.u-tokyo.ac.jp

Abstract

Background: Sleep is beneficial for physical and mental health. Several mobile and wearable sleep-tracking devices have been developed, and personalized sleep feedback is the most common functionality among these devices. To date, no study has implemented an objective push-type feedback message and investigated the characteristics of habitual sleep behavior and diurnal symptoms when receiving sleep feedback.

Objective: We conducted a mobile health intervention trial to examine whether sending objective push-type sleep feedback changes the self-reported mood, physical symptoms, and sleep behavior of Japanese office workers.

Methods: In total, 31 office workers (mean age 42.3, SD 7.9 years; male-to-female ratio 21:10) participated in a 2-arm intervention trial from November 30 to December 19, 2020. The participants were instructed to indicate their momentary mood and physical symptoms (depressive mood, anxiety, stress, sleepiness, fatigue, and neck and shoulder stiffness) 5 times a day using a smartphone app. In addition, daily work performance was rated once a day after work. They were randomly assigned to either a feedback or control group, wherein they did or did not receive messages about their sleep status on the app every morning, respectively. All participants wore activity monitors on their nondominant wrists, through which objective sleep data were registered on the web on a server. On the basis of the estimated sleep data on the server, personalized sleep feedback messages were generated and sent to the participants in the feedback group using the app. These processes were fully automated.

Results: Using hierarchical statistical models, we examined the differences in the statistical properties of sleep variables (sleep duration and midpoint of sleep) and daily work performance over the trial period. Group differences in the diurnal slopes for mood and physical symptoms were examined using a linear mixed effect model. We found a significant group difference among within-individual residuals at the midpoint of sleep (expected a posteriori for the difference: −15, 95% credible interval −26 to −4 min), suggesting more stable sleep timing in the feedback group. However, there were no significant group differences in daily work performance. We also found significant group differences in the diurnal slopes for sleepiness (P<.001), fatigue (P=.002), and neck and shoulder stiffness (P<.001), which was largely due to better scores in the feedback group at wake-up time relative to those in the control group.

Conclusions: This is the first mobile health study to demonstrate that objective push-type sleep feedback improves sleep timing of and physical symptoms in healthy office workers. Future research should incorporate specific behavioral instructions intended to improve sleep habits and examine the effectiveness of these instructions.

XSL•FO
RenderX

(JMIR Mhealth Uhealth 2022;10(10):e39150) doi:10.2196/39150

KEYWORDS

wearable activity monitor; smartphone app; sleep feedback; ecological momentary assessment; stabilized sleep timing; mood and physical symptoms

Introduction

Development of Mobile Health Technologies

Recent developments in internet and communication technologies, mobile sensing devices, and the Internet of Things (IoT) have enabled the acquisition of longitudinal multidimensional information, including real-time physiological, behavioral, and environmental data. For example, consumer-grade wearable fitness trackers (eg, Fitbit, Garmin, and Jawbone) can record health-related information such as physical activity, sleep, and heart rate objectively and repeatedly.

By using smartphones and their SMS text messaging, it is possible to provide effective support at the optimal time to promote awareness of one's health condition and improve compliance with intervention trials. Mobile devices can be useful for changing health-related behaviors as a part of daily living. Following these technological advancements and their applicability, medical and public health practices supported by mobile devices, such as mobile health (mHealth) [1,2], have been attracting attention in recent years. Since 2017, more than 325,000 mHealth apps have become available in commercial app stores, and the number of available apps continues to grow [3]. mHealth apps have been used in several intervention trials targeting physical activity [4-6], smoking [7,8], and suicidal ideation [9], and their usefulness and effectiveness have been proven.

Health Effects of Habitual Sleep Behavior

Sleep is a significant aspect of recovery. In fact, adequate and good-quality sleep is associated with better physical and mental health [10-12] and improved daytime functioning, including decreased physical symptoms [13,14], better work performance [15], and improved quality of life [16].

Although research interest has focused on the importance of sleep duration and quality, recent studies indicate that sleep timing and stability of habitual sleep behavior also play important roles in health. In fact, delayed sleep timing is associated with obesity [17], congestive heart failure [18], poor glycemic control [19], and increased severity of depressive symptoms [20,21]. Additional studies have indicated that variability in day-to-day sleep behavior (ie, duration, timing, and quality), referred to as intraindividual variability [22], is linked to physiological dysfunction [23,24], adverse medical and mental health conditions [25], and poor psychological well-being [26]. Therefore, the multifaceted monitoring and regulation of habitual sleep behavior can contribute to the prevention and support for physical and mental health problems.

Applications of Mobile and Wearable Technologies to Improve Sleep

Sleep data, along with other health-related behavioral and physiological data, can be gathered in real time using mobile devices. The widespread use of mobile or wearable sensing devices (eg, bed sensors, smartphone apps, and activity monitors) makes it easier and more commonplace to monitor sleep behavior in real time [27]. Although most consumer devices do not have Food and Drug Administration clearance as medical devices, they are expected to provide opportunities to track habitual sleep behavior longitudinally in large-scale populations [28]. For instance, Crowley et al [29] attempted to incorporate consumer wearable devices into health-promoting trials. They investigated the efficacy of these devices in improving physical activity and sleep among 565 employees over a 12-month period and found that sleep duration increased steadily throughout the study period.

In addition, mHealth apps for treating sleep disturbance have been developed rapidly [30,31], and more than 2000 mHealth apps targeting sleep are presently available in commercial app stores [32]. Pulantara et al [33,34] developed the interactive Resilience Enhancing Sleep Tactics app and examined its clinical feasibility as a treatment for sleep behavior. They reported that using the app improved insomnia severity and overall sleep quality, and the app was not inferior to traditional in-person sleep treatment. Furthermore, Hoersch et al [35] and Kuhn et al [36] conducted randomized controlled trials and reported that participants who received mHealth interventions had improved insomnia severity and sleep quality compared with waitlisted control participants.

Remaining Issues

Despite the rapid growth of mHealth apps and mobile sensing technologies, recent reviews have indicated that scientific trials examining the usefulness of mHealth apps are limited [32], and further research is required to test whether objective data enhance sleep outcomes [37]. This research investigating the usefulness of mHealth apps in enhancing sleep has several limitations. First, most mHealth trials have not assessed sleep behavior objectively [31], and the feedback provided depended on self-report assessments by the participants. While some studies incorporated wearable devices into the trials [38,39], they used the measurements only to assess the efficacy of the trial but not to objectivize the feedback. Given the importance of self-management for habitual sleep behavior to prevent future health problems, it is beneficial to implement objective sleep feedback into the apps. Second, previous mHealth studies focused on improvement in limited aspects of habitual sleep behavior, such as sleep quality and sleep duration [31], and the dynamic aspects of sleep, including intraindividual variability in sleep measurements over the trial period, have tended to be ignored. To the best of our knowledge, only 1 mHealth study by Murawski et al [40] reported the dynamic aspects of sleep behavior and found that variability in sleep timing, as assessed by a self-report questionnaire, was improved after their intervention. Studying the dynamic features in sleep behavior may provide insight into the typical properties of sleep

XSL•FO
RenderX

self-regulation processes when responding to feedback messages or other interventions. Finally, in view of the wide range of effects of sleep, including psychological well-being, the covariant relationships of sleep with daytime functions, such as mood, physical symptoms, and work performance, should be examined. The covariant relationships may contain important information about subordinate effects and facilitate a comprehensive understanding of sleep self-management.

Objective

The objective of this study was to conduct an mHealth trial sending objective push-type sleep feedback to healthy participants using a smartphone app and a wearable activity monitor. Specifically, we examined whether sending daily sleep feedback messages changed sleep behavior and self-reported symptoms of the participants, particularly depressive mood, anxiety, stress, sleepiness, fatigue, and neck and shoulder stiffness. We used exploratory analysis of the statistical properties of objectively measured sleep variables and the characteristics of momentary symptoms recorded during the day using ecological momentary assessment (EMA).

Methods

Study Design

In this study, we conducted a 2-arm intervention trial by performing random convenience sampling of office workers at an insurance company and stratifying them into control and feedback groups. By comparing the groups in terms of the characteristics of habitual sleep behaviors (sleep duration and midpoint of sleep), momentary symptoms (depressive mood, anxiety, stress, sleepiness, fatigue, and neck and shoulder stiffness), and daily work performance, we examined the effects of personalized sleep feedback. To minimize the memory distortion caused by retrospective recall, momentary symptoms were recorded on a smartphone app in real time. Habitual sleep behaviors were measured objectively using a wearable device. Possible extraneous variables, including pretrial psychological symptoms, habitual sleep behaviors, and work performance, were assessed before the trial.

EMA Method

We used the EMA method to acquire momentary mood and physical symptom data (ie, depressive mood, anxiety, stress, fatigue, sleepiness, and neck and shoulder stiffness) in real time. EMA is a method for recording participants' behavior, psychological state, and physical symptoms in real time and at multiple time points, allowing the collection of self-report and objective data with reliability and ecological validity. Thus, EMA avoids potential distortions of retrospective recall in self-reported data [41,42].

Health Care Internet of Things System

We developed a cloud-based health care Internet of Things (HIT) system that can continuously acquire health-related information, including momentary symptoms, biological signals, and surrounding environmental information, recorded as part of daily living. The HIT system consists of a cloud server and a smartphone app for data collection (HIT server and HIT app, respectively). The HIT app is equipped with an EMA and users can record their momentary symptoms in daily life (Multimedia Appendix 1). In addition, the HIT app can connect with various IoT devices, including a proprietary activity monitor (Sciencenet device, Sciencenet Inc) used in this study, using Bluetooth Low Energy (BLE). Data are transferred from the IoT devices to the HIT server. The app is compatible with both Android and iOS operating systems. The HIT server can store, integrate, and manage data uploaded from the app and send personalized messages (push-type feedback messages) to the HIT app users. HIT systems have been used to assess self-reported symptoms in real time [43].

Participants

A convenience sample of 31 office workers working at an insurance company participated in this study. The mean age of the participants was 42.3 (SD 7.9) years, and the male-to-female ratio was 21:10. All participants were working from home during the trial period at the request of their employer to prevent the spread of COVID-19.

Participants were randomly assigned to a control or feedback group using the "sample" function in the R statistical software (version 4.0.2; R Foundation for Statistical Computing) so that the ratio of the sample size was 1:1. No stratification by age or sex was observed. Coauthor KS conducted this randomization, independent of the primary researcher HT. Although author HT was also informed of who was assigned to which group after the random assignment, he was not allowed to contact the participants during the trial period.

During the trial period, the participants in the feedback group received personalized messages regarding their current sleep status every morning, whereas the participants in the control group did not receive any messages. The control group consisted of 16 participants, including 9 males and 7 females, with a mean age of 44.1 (SD 8.3) years. The feedback group consisted of 15 participants, including 12 males and 3 females, with a mean age of 40.5 (SD 7.2) years.

Instruments

Baseline Questionnaire

Before the trial, the participants completed a baseline questionnaire, including their demographic information (age, sex, and BMI), psychological symptoms (depressive and anxiety symptoms), habitual sleep behaviors (habitual sleep duration and self-reported sleep quality), and self-reported work performance. Items included in the baseline questionnaire are listed in subsequent sections.

Psychological Symptoms

Depressive symptoms were assessed using the Japanese version of the Beck Depression Inventory second edition (BDI-II) [44,45]. The BDI-II is a 21-item self-report inventory for measuring the presence and severity of depression (score range 0-63). A high level of internal consistency (Cronbach α=.87) and item homogeneity have been confirmed for the Japanese version of the BDI-II [45]. The BDI-II classifies individuals into 4 categories based on an overall score: minimal or no depression, 0 to 13; mild depression, 14 to 19; moderate

XSL•FO
RenderX

depression, 20 to 28; and severe depression, 29 to 63. A score of ≥14 points was used as the clinical cutoff point for depression.

Anxiety symptoms were assessed using the Japanese version of the State-Trait Anxiety Inventory (STAI) Form Y [46]. The STAI is a standardized self-report inventory for measuring state and trait anxiety with 20 items (STAI Y-1 and STAI Y-2, respectively). The STAI Y-1 measures the intensity of the anxiety felt by an individual in the present, whereas the STAI Y-2 measures how often an individual feels anxious. Scores range from 20 to 80 for each subscale, with higher scores indicating higher levels of anxiety.

Habitual Sleep Behaviors

Habitual sleep duration on workdays and free days (SL_w and SL_f, respectively) was assessed using a single question for each ("How long do you sleep on weekdays?" and "How long do you sleep if tomorrow is a holiday?"). These measurements were used to compute an index representing participants' sleep status during the trial period. Sleep quality was assessed using the Pittsburgh Sleep Quality Index (PSQI). The PSQI is a self-report inventory used to assess sleep quality over the preceding month [47,48]. The PSQI consists of 19 items on self-reported sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, use of sleeping medication, and daytime dysfunction. Scores range from 0 to 21, with a higher total score indicating poorer sleep quality. The strong reliability and validity of this questionnaire have been confirmed in a recent meta-analysis [49].

Self-reported Work Performance

Self-reported work performance was measured using the World Health Organization Health and Work Performance Questionnaire (HPQ) [50,51]. The HPQ asks participants to rate their overall work performance over the preceding 4 weeks on a self-anchoring scale from 0 to 10: "On a scale of 0 to 10, how would you rate your usual work performance over the past four weeks?" The score was converted to a 100-point scale by multiplying the raw score by 10, with a higher score indicating better work performance.

EMA Questionnaire

The participants answered an EMA questionnaire 5 times per day using the HIT app. The EMA included the following measurements:

1. Depressive mood and anxiety were scored using the Depression and Anxiety Mood Scale [52]. This scale comprises the following 9 adjectives representing mood states: "vigorous," "gloomy," "concerned," "happy," "unpleasant," "anxious," "cheerful," "depressed," and "worried." On the basis of the 9 items, anxious (the sum of "concerned," "anxious," and "worried" scores), positive (the sum of "vigorous," "happy," and "cheerful" scores), and negative (the sum of "gloomy," "unpleasant," and "depressed" scores) moods were calculated. Depressive mood scores were obtained by combining the last 2 mood scores as follows: (300-positive mood score) +negative mood score. The resulting depressive mood scores were rescaled to range from 0 to 100.
2. Physical symptoms, including stress, sleepiness, fatigue, and neck and shoulder stiffness were rated according to the participant's response to being asked if they felt "stressed," "sleepy," or "fatigued" and "if their neck and shoulders were stiff."
3. Daily work performance was rated after work in response to the question, "How would you rate your work performance of today?"

These measurements were rated using a visual analog scale from 0 to 100 displayed on the screen. All scores were transferred to the HIT server immediately after the completion of each EMA questionnaire.

Sleep Monitoring

The participants were instructed to wear a wristband-type activity monitor on their nondominant wrist during the trial period, except while bathing, showering, performing rigorous exercise, or any other activities likely to damage the device. The device is equipped with triaxial piezoelectric accelerometers capable of detecting small changes in bodily acceleration (≥0.01 G/rad/s). We confirmed that the device performs at a level equivalent to research-grade actigraphy (Ambulatory Monitors Inc), which is widely used in clinical settings. Results of the comparative analysis are presented in Multimedia Appendix 2.

The device was configured to transfer physical activity data to the HIT server using the HIT app whenever a participant launched the app. We used zero-crossing counts, which counts the number of times per epoch that the acceleration signal level crosses 0 [53], accumulated per minute, to compute objective sleep variables.

To estimate sleep variables, we adopted the Cole-Kripke algorithm with Webster's rescoring rules [54,55] for zero-crossing count data to identify whether the recorded 1-minute epoch was sleep or wake. Next, we introduced the sleep probability function (SPF) $\theta(t)$ as follows:

where t denotes the clock time converted to a numerical value (eg, 3 AM is transformed to 3.00, and 6:30 PM is converted to 18.50). Thus, the SPF is a value ranging from 0 to 1 and represents circadian oscillations.

Then, we estimated the effective SPF parameters (β_0, β_1, and β_2) to fit the actual Cole-Kripke identification using a Bernoulli logistic regression model.

$$CK(t) \sim bernoulli(\theta(t)) \quad \textbf{(2)}$$

where $CK(t)$ represents the result of Cole-Kripke identification when the clock time is t; $CK(t) = 0$ and $CK(t) = 1$ denote that the epoch at time t was labeled as wake and sleep, respectively.

Finally, we estimated the square waveform function $\theta'(t)$ from $\theta(t)$ by introducing onset and offset (t_{on} and t_{off}, respectively).

Effective t_{on} and t_{off} were computed to maximize the R^2 value between $\theta'(t)$ and $CK(t)$ using the Nelder-Mead method. We

XSL•FO
RenderX

assumed that t_{on} and t_{off} represent bedtime and wake-up time, respectively; thus, sleep duration and midpoint of sleep were determined using their interval and midpoint. We confirmed that the algorithm was performed at a level equivalent to Action-W version 2 software (AW2 software, Ambulatory Monitors Inc), which was used to analyze the research-grade actigraphy data (Multimedia Appendix 2).

The average sleep duration per day was calculated based on the habitual sleep duration on workdays and free days (S_w and S_f, respectively) in the baseline questionnaire using the following equation:

which represents the expected sleep duration per day because the participants worked 5 days per week. Sleep debt and cumulative sleep debt were defined as follows:

sleep debt = Average sleep duration − Estimated sleep duration...**(5)**

cumulative sleep debt = Σsleep debt...**(6)**

Thus, sleep debt represents sleep insufficiency per day, and cumulative sleep debt represents its cumulative value over time. During the trial period, these values were automatically sent to the participants in the feedback group using the HIT app.

Data Collection Protocol

Trained researchers provided participants with a comprehensive explanation of the purpose and potential risks of the study. Subsequently, they signed an informed consent form and completed the baseline questionnaire. In addition, they received a sleep hygiene guide that listed daytime activities to improve their habitual sleep behaviors or health conditions (Multimedia Appendix 3). They were then asked to install the HIT app on their smartphones and wear an activity monitor on their nondominant wrist. All participants were instructed on the use of the app and the activity monitor. Using these instruments, we repeatedly measured their momentary symptoms and physical activity data in real time. The setup and operating procedures were presented on the web as much as possible using a communication service and videoconferencing system.

An overview of the trial is shown in Figure 1. This trial was carried out for almost 3 weeks (from November 30 to December 19, 2020). During the trial period, the participants were asked to complete the EMA questionnaires (see the section EMA Questionnaire) at randomly selected times within +10 minutes to −10 minutes of predetermined times (10 AM and 2 PM). In addition, they were asked to complete the EMA when they woke up, finished their work, and went to bed (wake-up time, after work, and bedtime, respectively).

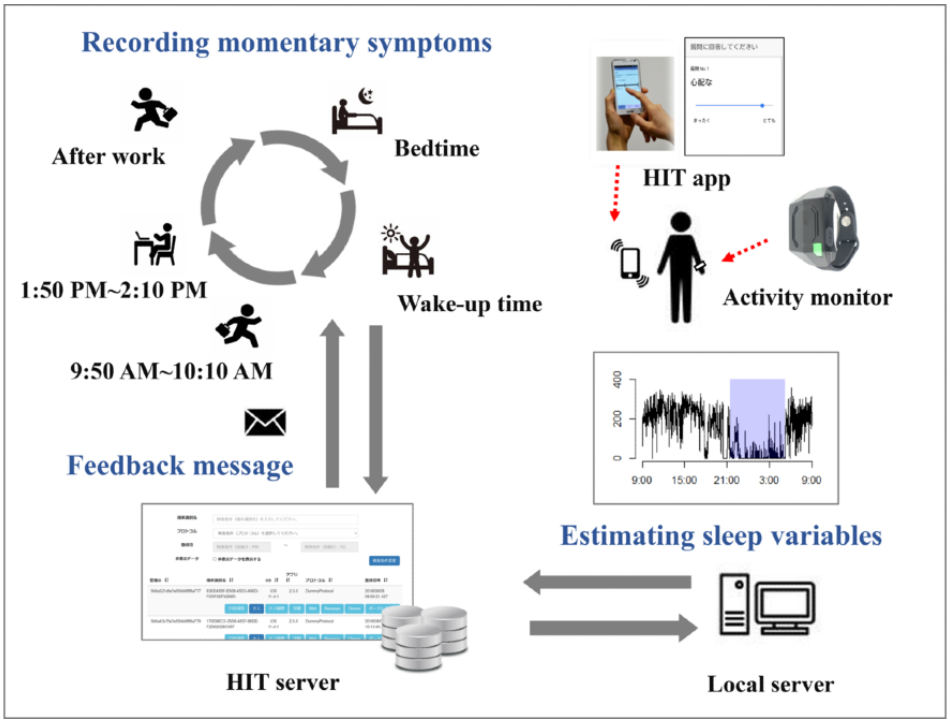
At 9 AM every day, the physical activity data on the HIT server were collated and analyzed to estimate sleep duration, sleep debt, and cumulative sleep debt using a local data analysis server. On the basis of the estimated sleep data, personalized sleep feedback messages were generated and sent to participants in the feedback group. They were informed about their sleep status (estimated sleep debt and cumulative sleep debt) and requested to plan and adjust their daytime activities with reference to guidelines for reducing their sleep debt. The message read as follows: “You accumulated XX minutes of sleep debt yesterday. Your current overall debt is XX minutes. Sleep debt has adverse effects on physical and mental health. Adjust your daytime behavior to cancel your debt.”

If sufficient physical activity data (<720 records/day) had not been uploaded to the server by 9 AM, an alternative message was sent to the participant requesting them to confirm the BLE pairing of the activity monitor with their app: “It seems that your data have not been uploaded successfully. We will analyze the data again at 1 PM. Please check the BLE connection of your HIT app before then.” The same computation process was executed at 1 PM for the relevant participants. These processes were executed by the local servers and were fully automated.

After the trial period, the participants rated their overall work performance in the preceding 2 weeks on a self-anchoring scale from 1 to 10. The question was as follows: “On a scale from 0 to 10, how would you rate your work performance over the past two weeks?” While the HPQ was originally developed to evaluate work performance over the past month, the scale used in this study was modified to evaluate performance over the past 2 weeks, corresponding to the survey period.

XSL•FO
RenderX

Figure 1. Overview of the trial. HIT: health care Internet of Things.



Statistical Analysis

Sleep variables, including sleep duration and midpoint of sleep, and daily work performance were recorded once a day, while mood and physical symptoms were recorded several times a day using smartphone-based EMA. Owing to this difference in the frequency of data recording, we used different models to analyze the group differences in the obtained data.

For the sleep variables and daily work performance, we constructed a hierarchical Bayesian model to capture the daily trend, baseline level, and within-individual stability, as follows:

$$y_{ijk} = \beta_{ij}^{0} + \beta_{i}^{1} Day_{ijk} + e_{ijk}$$

$$\beta_{ij}^{0} = \beta_{i}^{0} + r_{ij}^{0}$$

$$e_{ijk} \sim N(0, \sigma_{i}^{y})$$

$$r_{ij}^{0} \sim N(0, \sigma^{0})$$

where y_{ijk} indicates the dependent variable (sleep duration, midpoint of sleep, or daily work performance) at the k-th recording for the j-th participant in the i-th group; Day_{ijk} indicates the day on which the corresponding dependent variable was measured to record the trend (improving or worsening); β_{ij}^{0} is the intercept of the j-th participant in the i-th group; β_{i}^{1} is the slope for the day of the i-th group; the random terms r_{ij}^{0} are the between-individual residuals; and e_{ijk} are the within-individual residuals. In particular, the variance component σ_{i}^{y} can be interpreted as a measurement of how stable or fluctuating the sleep variables within an individual are at the group level. Therefore, σ_{i}^{y} was assumed to be affected by whether the intervention was provided and estimated for each group. All random terms were assumed to follow a normal distribution.

For EMA-recorded mood and physical symptoms, we used linear mixed effect models to examine their diurnal slopes and group differences.

$$y_{ijk} = \beta_{j}^{0} + \beta^{1} Time_{ijk} + \beta^{2} Group_{ijk} + \beta^{3} Time_{ijk} * Group_{ijk} + e_{ijk}$$

$$\beta_{j}^{0} = \beta^{0} + r_{j}^{0}$$

$$e_{ijk} \sim N(0, \sigma^{y})$$

$$r_{j}^{0} \sim N(0, \sigma^{0})$$

where y_{ijk} indicates the moods or physical symptoms at the k-th recording for the j-th participant in the i-th group; $Time_{ijk}$ indicates the categorical variable representing the timing when the corresponding dependent variables were recorded; thus, $Time_{ijk} = 0$ and $Time_{ijk} = 1$ denote that the dependent variables were recorded at wake-up time and bedtime, respectively; $Group_{ijk}$ indicates the categorical variable representing the group in which the j-th participant was classed; thus, $Group_{ijk} = 0$ and $Group_{ijk} = 1$ denote that the dependent variable was obtained from the participant classed as the control and feedback groups, respectively; β_{j}^{0} is the intercept of the j-th participant; β^{1}, β^{2}, and β^{3} are the coefficients for $Time_{ijk}$, $Group_{ijk}$, and their interaction term, respectively; the random terms r_{j}^{0} are the between-individual residuals and e_{ijk} are the within-individual residuals. All random terms were assumed to follow a normal distribution. When a significant interaction effect was observed, we performed multiple comparison tests with Tukey correction.

In addition, we performed two-tailed Welch t test and Fisher exact test for the baseline data to confirm that there were no significant group differences. As the HPQ scores were recorded in the baseline and follow-up questionnaires, we performed 2-way repeated measures ANOVA to examine the main effects

XSL•FO
RenderX

of group (control vs feedback) and time (pre- vs postintervention) and their interaction effect. For the cumulative sleep debt, we performed the Welch 2-tailed *t* test and Levene test for equality of variances for the final observation data per participant to examine the group differences in terms of mean and variance.

All analyses were performed using R statistical software (version 4.0.2). In particular, the parameters of the statistical models were computed using the rstan [56] and lmerTest [57] packages. The emmeans package (also known as the lsmeans package) [58] was used for multiple comparison tests with Tukey correction. Statistical significance was defined as when the 95% credible interval (CI) did not include the null value or when a *P* value <.05.

Ethics Approval

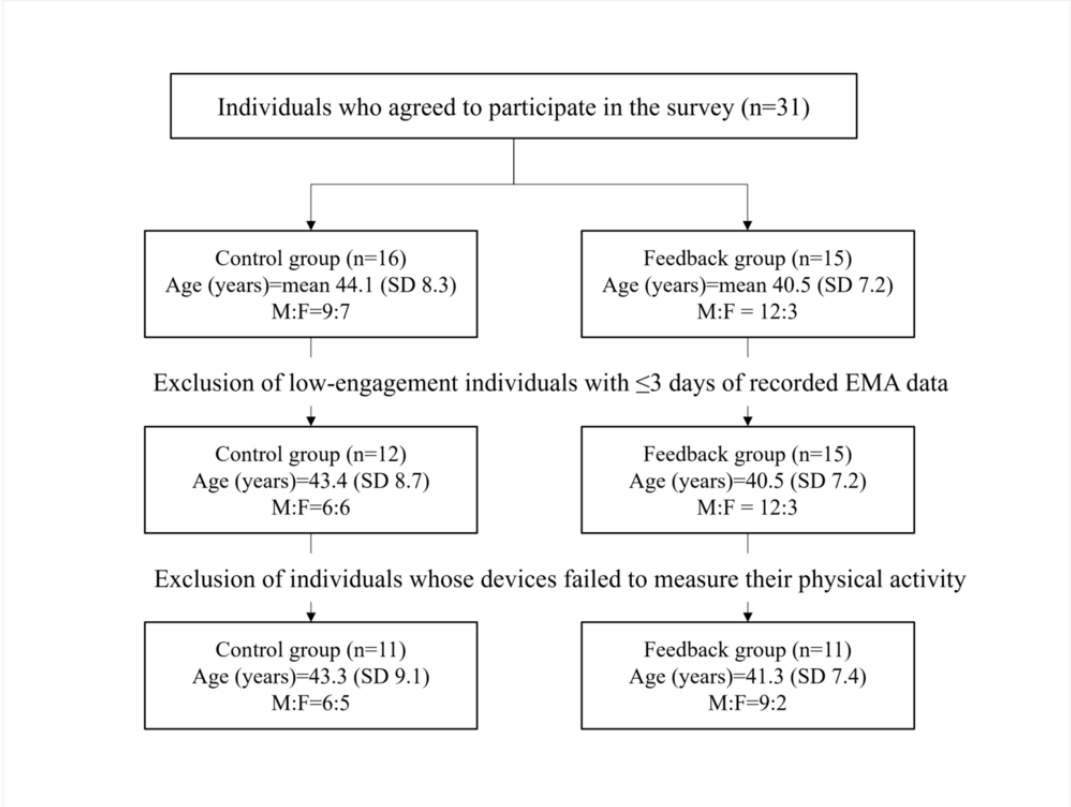
The Ethics Committee of the University of Tokyo approved this study and the informed consent form (approval number 20-20).

Results

Demographic Characteristics

In total, 31 individuals agreed to participate in the study. However, the data of 4 participants were excluded from the statistical analyses because of their low response rates for wake-up time and bedtime on the EMA questionnaire (ie, <3 days; Figure 2). The demographic characteristics of the participants, including the control group (n=12) and the feedback group (n=15), are presented in Table 1. The mean age of the participants was 41.8 (SD 7.9) years, and 33% (9/27) were females. The Welch 2-tailed *t* test and Fisher exact test showed that there were no significant differences in demographic characteristics between the groups. The total number of EMA records was 1839, and the overall response rate was 64.88% (1839/2835).

Figure 2. Flowchart of participant data selection for analyses. EMA: ecological momentary assessment; F: female; M: male.



XSL•FO
RenderX

Table 1. Baseline demographic characteristics of participants by group.

	Overall (N=27)	Control group (n=12)	Feedback group (n=15)	*P* value
Age (years), mean (SD)	41.78 (7.90)	43.42 (8.72)	40.47 (7.21)	.35[a]
Female, n (%)	9 (33)	6 (50)	3 (20)	.22[b]
BMI (kg/m^2), mean (SD)	21.65 (2.60)	21.18 (2.54)	22.03 (2.67)	.40[a]
BDI-II[c], mean (SD)	6.51 (5.13)	6.92 (4.94)	6.20 (5.42)	.72[a]
BDI-II >13, n (%)	4 (15)	2 (17)	2 (13)	.99[b]
STAI Y-1[d], mean (SD)	39.78 (11.05)	38.75 (11.31)	40.60 (11.17)	.67[a]
STAI Y-2[e], mean (SD)	40.74 (10.28)	38.67 (8.50)	41.20 (11.69)	.52[a]
PSQI[f], mean (SD)	4.73 (2.16)	4.75 (2.05)	4.71 (2.33)	.97[a]
S_w[g], mean (SD)	6 h 18 min (1 h 4 min)	6 h 20 min (1 h 18 min)	6 h 16 min (53 min)	.88[a]
S_f[h], mean (SD)	6 h 49 min (1 h 23 min)	6 h 35 min (1 h 23 min)	7 h 00 min (1 h 25 min)	.45[a]

[a]Welch *t* test.

[b]Fisher exact test.

[c]BDI-II: Beck Depression Inventory second edition.

[d]STAI Y-1: State Anxiety Scale.

[e]STAI Y-2: Trait Anxiety Scale.

[f]PSQI: Pittsburgh Sleep Quality Index.

[g]S_w: sleep duration on work days.

[h]S_f: sleep duration on free days.

Statistical Properties of Sleep Variables

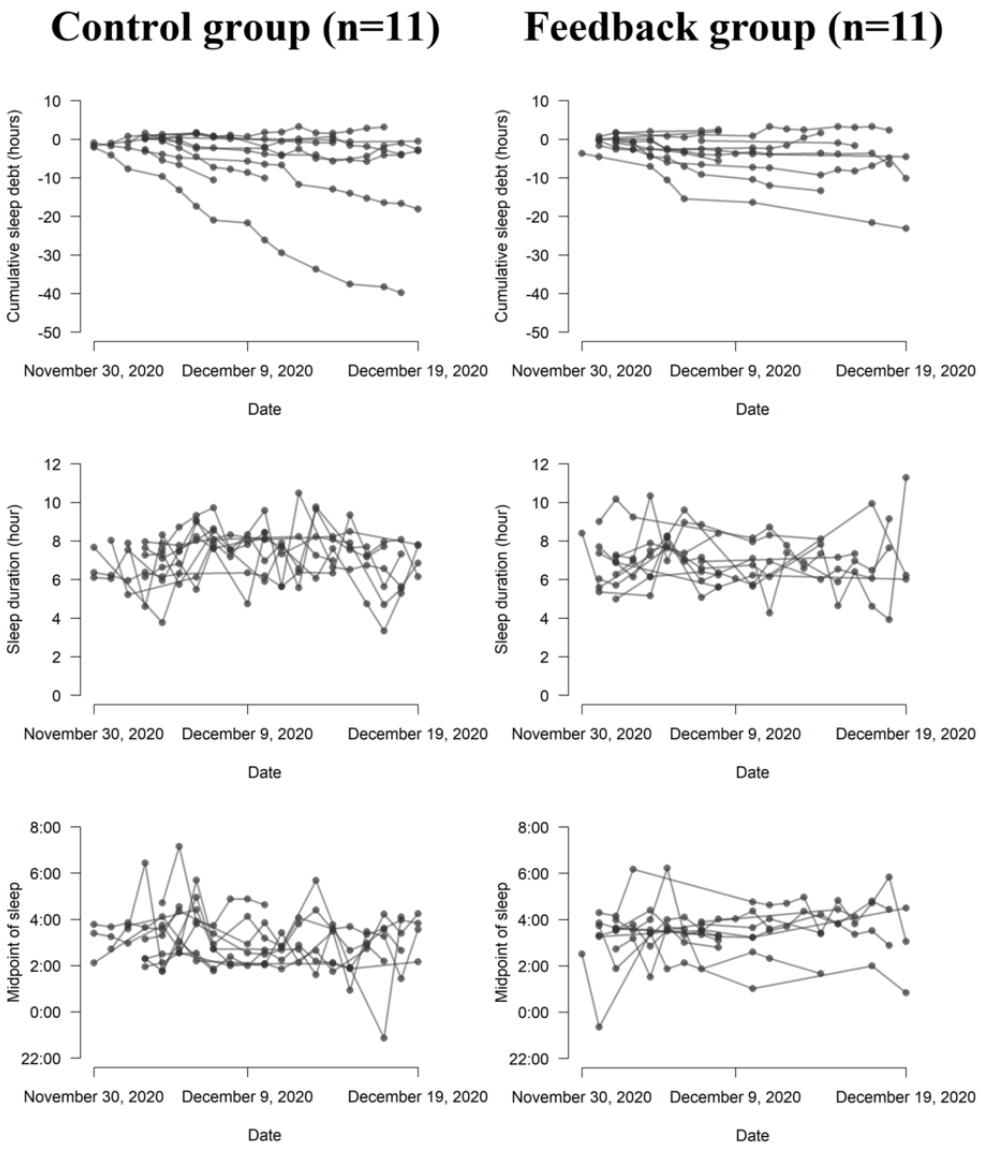
We examined group differences in the sleep variables such as cumulative sleep debt, sleep duration, and midpoint of sleep during the trial period. The data of 5 of the 27 (18%) participants were excluded from the analyses because of physical activity measurement failures (Figure 2). Thus, the data from 11 participants in the control group (mean age 43.3, SD 9.1 years; 6 males and 5 females) and 11 participants in the feedback group (mean age 41.3, SD 7.41 years; 9 males and 2 females) were analyzed.

Figure 3 presents the spaghetti plots of the computed sleep variables for all participants during the trial period. The Welch *t* test and Levene test for equality of variances indicated that there was no significant group difference in cumulative sleep debt per participant, $t_{17.32}$=0.64, *P*=.53; $F_{1,20}$=0.15, *P*=.70.

We subsequently examined the group differences in the daily trend, baseline level, and within-individual residuals for sleep duration and midpoint of sleep using the hierarchical Bayesian model. The within-individual residuals for the midpoint of sleep in the feedback group were significantly smaller than those in the control group (expected a posteriori for the difference: −15, 95% CI −26 to −4 min; Table 2). This was also the case for the within-individual residuals for bedtime (−18, 95% CI−31 to −4 min; Multimedia Appendix 4). Both groups showed no significant slope of the day in terms of sleep duration (control group: 0 95% CI −3 to 3 min; feedback group: 1, 95% CI −3 to 1 min) or midpoint of sleep (control group: −1, 95% CI −4 to 1 min; feedback group: 1, 95% CI −1 to 3 min).

XSL•FO
RenderX

Figure 3. Spaghetti plots of the estimated cumulative sleep debt (top panels), sleep duration (middle panels), and midpoint of sleep (bottom panels) per participant across the trial period. The left and right panels indicate the time series of the sleep variables for the control group and the feedback group, respectively.



XSL•FO
RenderX

Table 2. Results of the hierarchical Bayesian model for sleep duration and midpoint of sleep.

	Control group		Feedback group		Difference	
	EAP[a] (SD)[b]	95% CI[c]	EAP (SD)	95% CI	EAP (SD)	95% CI
Sleep duration[d]						
Intercept	7 h 33 min (26 min)	6 h 42 min to 8 h 23 min	7 h 31 min (33 min)	6 h 27 min to 8 h 36 min	−1 min (31 min)	−1 h 2 min to 58 min
Day	0 min (1 min)	−3 min to 3 min	1 min (1 min)	−3 min to 1 min	−1 min (2 min)	−4 min to 3 min
σ^{0e}	48 min (11 min)	29 min to 1 h 14 min	48 min (11 min)	29 min to 1 h 14 min	N/A[f]	N/A
σ^{y}	1 h 13 min (5 min)	1 h 4 min to 1 h 23 min	1 h 13 min (6 min)	1 h 3 min to 1 h 26 min	0 min (8 min)	−14 min to 16 min
Midpoint of sleep[d]						
Intercept	3:30 (21 min)	2:48 to 4:12	3:34 (26 min)	2:44 to 4:26	4 min (24 min)	−42 min to 52 min
Day	−1 min (1 min)	−4 min to 1 min	1 min (1 min)	−1 min to 3 min	2 min (1 min)	0 min to 5 min
σ^{0}	41 min (9 min)	27 min to 1 h 1 min	41 min (9 min)	27 min to 1 h 1 min	N/A	N/A
σ^{y}	1 h 1 min (4 min)	53 min to 1 h 9 min	46 min (4 min)	39 min to 54 min	*−15 min (5 min)*[g]	*−26 min to −4 min*

[a]EAP: expected a posteriori (expected value of the posterior distribution).

[b]SD of the posterior distribution.

[c]CI: credible interval.

[d]The models were run after controlling for age and sex.

[e]The difference in the interindividual variability for the intercept (σ^{0}) was not computed because σ^{0} was assumed to be equal between groups.

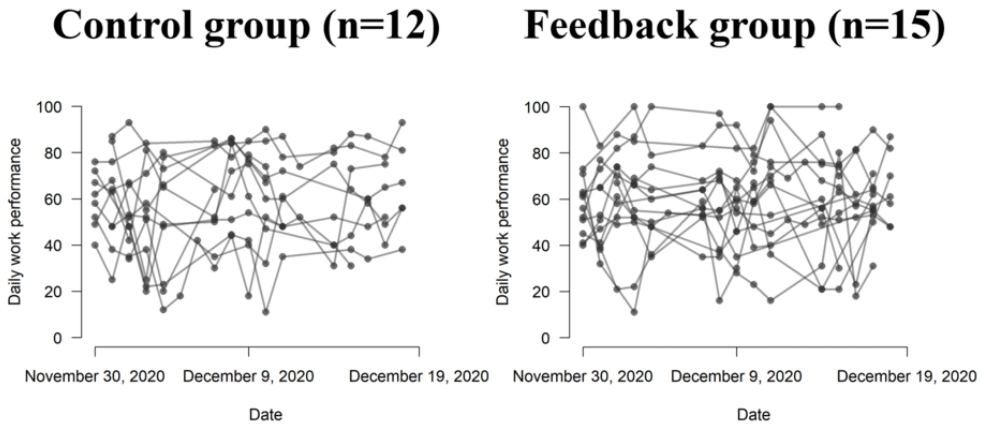
[f]N/A: not applicable.

[g]Italicized values denote statistically significant group effects, at a 95% CI.

Self-reported Work Performance Assessment

We examined the group differences in daily work performance during the trial period using 307 EMA records obtained from 27 participants (Figure 4). We examined the statistical properties of daily work performance by using the statistical model used to analyze the sleep variables. However, there were no significant differences between the groups (Table 3). When comparing the HPQ score before and after the intervention using a 2-way repeated ANOVA, the main effects of group, $F_{1,25}$=0.39, P=.54, and time, $F_{1,25}$=0.01, P=.94, and their interaction, $F_{1,25}$=0.05, P=.83, were not significant.

Figure 4. Spaghetti plots of the daily work performance recorded using ecological momentary assessment per participant across the trial period. The left and right panels indicate the time series of the work performance for the control group and the feedback group, respectively.



XSL•FO
RenderX

Table 3. Results of the hierarchical Bayesian model for daily work performance.

	Control		Feedback		Difference	
	EAP[a] (SD)[b]	95% CI[c]	EAP (SD)	95% CI	EAP (SD)	95% CI
Daily work performance[d]						
Intercept	57.59 (6.67)	44.60 to 70.54	61.45 (8.77)	44.24 to 78.82	3.85 (8.01)	−11.92 to 19.84
Day	0.32 (0.22)	−0.11 to 0.75	0.08 (0.17)	−0.26 to 0.41	−0.24 (0.27)	−0.77 to 0.29
σ^{0e}	17.06 (2.89)	12.38 to 23.54	17.06 (2.89)	12.38 to 23.54	N/A[f]	N/A
σ^{y}	13.28 (0.88)	11.68 to 15.14	12.62 (0.72)	11.31 to 14.12	−0.66 (1.14)	−2.94 to 1.50

[a]EAP: expected a posteriori.

[b]SD of the posterior distribution.

[c]CI: credible interval.

[d]The models were run after controlling for age and sex.

[e]The difference in the interindividual variability for the intercept (σ^0) was not computed because σ^0 was assumed to be equal between groups.

[f]N/A: not applicable.

Diurnal Slopes for EMA Scores

The recorded EMA scores of 27 participants (760 records) were used to examine the diurnal slopes in momentary mood and physical symptoms (Figure 2). The linear mixed effect model results showed significant interaction effects between group and time in physical symptoms (fatigue, P=.002; sleepiness, P<.001; and neck and shoulder stiffness, P<.001; Table 4). In addition, a multiple comparison test with Tukey correction showed that the EMA scores of physical symptoms at wake-up time were significantly lower than those at bedtime in the feedback group (fatigue, sleepiness, and neck and shoulder stiffness, P<.001). In the control group, a significant difference between wake-up time and bedtime was observed only for fatigue (P<.001; Figure 5).

XSL•FO
RenderX

Table 4. Results of the linear mixed effect model for mood and physical symptoms.

Mood and physical symptoms[a]	Coefficient (SE)	Df[b]	*P* value
Depressive mood			
Intercept[c]	41.27 (5.56)	23.28	<.001
Group	−9.87 (6.61)	23.42	.15
Time	−0.35 (1.10)	731.32	.75
Group×Time	0.74 (1.51)	731.6	.63
Anxiety			
Intercept[c]	35.75 (7.42)	23.31	<.001
Group	−16.89 (8.82)	23.46	.07
Time	−1.29 (1.56)	731.36	.41
Group×Time	3.03 (2.14)	731.66	.16
Stress			
Intercept	47.07 (8.02)	23.42	<.001
Group	−18.66 (9.54)	23.6	.06
Time	2.49 (1.80)	731.46	.17
Group×Time	0.46 (2.47)	731.82	.85
Fatigue			
Intercept	35.22 (7.24)	23.72	*<.001* [d]
Group	−16.47 (8.62)	24.03	*.07*
Time	16.07 (2.14)	731.78	*<.001*
Group×Time	9.31 (2.96)	732.39	*.002*
Sleepiness			
Intercept	55.75 (7.20)	24.08	*<.001*
Group	−9.81 (8.58)	24.51	*.26*
Time	1.27 (2.49)	732.14	*.61*
Group×Time	17.70 (3.42)	732.96	*<.001*
Neck and shoulder stiffness			
Intercept	43.33 (9.59)	23.27	*<.001*
Group	−12.99 (11.40)	23.37	*.27*
Time	−2.51 (1.64)	731.3	*.13*
Group×Time	8.46 (2.26)	731.51	*<.001*

[a]The df values correspond to the denominator df in ANOVA model.

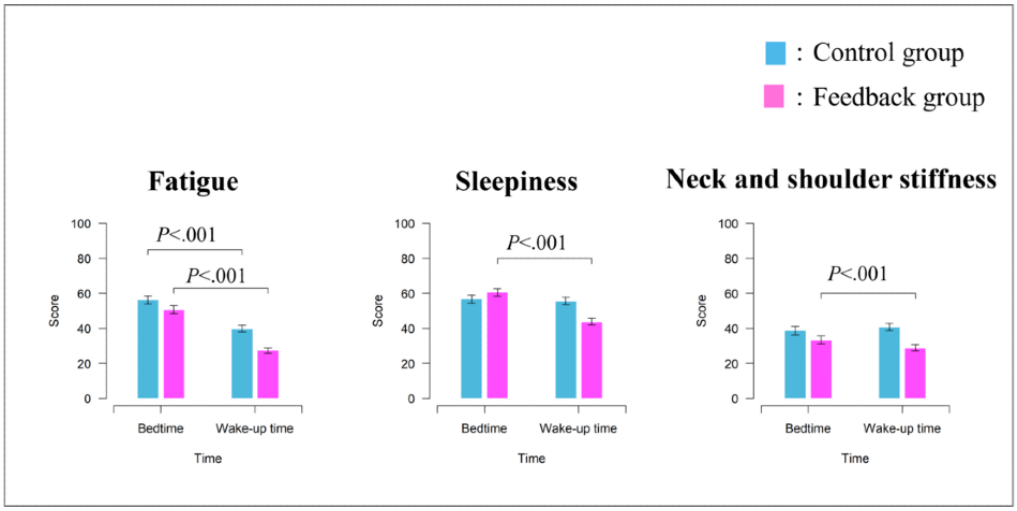
[b]The effects of group (control vs feedback group) and time (wake-up time vs bedtime) were assumed to be fixed effects, and those of individuals were assumed to be random effects.

[c]The control group and wake-up time were used as the reference categories for each variable; thus, the intercept indicates the expected EMA score of moods or physical symptoms for the control group during the wake-up time.

[d]Italicized values denote statistically significant interaction effects. All models were run by controlling for age and sex.

XSL•FO
RenderX

Figure 5. Average ecological momentary assessment scores for fatigue, sleepiness, and neck and shoulder stiffness at bedtime and wake-up time in the control group and the feedback group. The error bars indicate the SE per group.



Discussion

Principal Findings

We explored the effects of personalized feedback messages regarding the current sleep status on habitual sleep behavior and momentary mood and physical symptoms in Japanese office workers, using a unique cloud-based HIT system that included a web-based wearable activity monitor and a smartphone app. Specifically, we focused on group differences in the statistical properties of sleep variables and within-day momentary symptoms during the trial period. We found that the within-individual residuals for sleep timing were significantly smaller in the feedback group than in the control group. In addition, the diurnal slopes for physical symptoms (sleepiness, fatigue, and neck and shoulder stiffness) differed significantly between the feedback and control groups, largely because of better physical symptom scores in the feedback group at wake-up time. This is the first mHealth study to implement push-type sleep feedback based on objective measurements and to demonstrate improved sleep status and momentary symptoms associated with receiving the feedback message. The findings in this study suggest that objective push-type feedback messages may promote sleep self-management and solve habitual sleep behavior problems, despite the minor inconvenience.

Several mHealth apps have been developed for treating sleep disturbances such as insomnia [30,31], but the objective sleep feedback function has not been implemented in these apps. While users of consumer-grade wearable devices can take advantage of objective sleep measurements, these are regarded as pull-type interventions, as these devices require the user to actively access an app to receive feedback, which can sometimes be burdensome. Connecting mHealth apps with wearable sensing devices and sending objective push-type sleep feedback may be a feature to consider when developing or updating mHealth apps that target sleep disturbance.

Comparison With Previous Work

The within-individual residuals in sleep variables examined in this study can represent how sleep behavior varies across days, which is commonly referred to as intraindividual variability [22]. Recently, the stability of habitual sleep behavior has been considered a critical factor for physical and mental health, as previous studies have indicated that greater intraindividual variability in sleep behavior is associated with worse medical health conditions [25] and poorer psychological well-being [26]. In our study, participants may have attempted to improve their sleep habits by adopting strategies to stabilize their sleep timing (specifically their bedtime); for example, by not staying up late excessively. This inference is supported by evidence of a significant group difference in the within-individual residuals at bedtime but not in wake-up time. Similar results were reported by Murawski et al [40] who found improvement in the variability of sleep timing after the 3-month intervention. Given the fact that our survey period was relatively short, it is speculated that the improvement in sleep variability is an initial change caused by improved awareness of habitual sleep behaviors. Especially in modern industrial societies, the sleep-wake cycle adhering to social schedules, rather than endogenous circadian rhythms, leads to exposure to bright light at significantly different times from the natural environment. This causes disturbances in sleep and circadian rhythms, such as circadian misalignment [59] and social jet lag [60,61], which are linked to future health problems. Thus, the findings of this study suggest that sending personalized sleep feedback messages may potentially contribute to the primary prevention of physical and mental health problems as well as the improvement of the sleep-wake cycle.

Improvements in physical symptoms and stabilization of sleep timing were simultaneously observed in the feedback group, suggesting the covariant relationship between them. Indeed, previous studies have indicated that individuals with greater intraindividual variability in sleep timing and sleep duration show more dysregulated biomarkers related to endogenous circadian rhythm [23] and inflammatory functions [24], which can influence diurnal symptoms including sleepiness and fatigue. Therefore, improved physical symptoms at wake-up time in the feedback group may have been caused by stabilized sleep timing, mediated by regulated physiological systems. Thus, it is possible that improved physical symptoms were observed as a short-term effect or proximal outcome of the interventions

XSL•FO
RenderX

that stabilized sleep timing and, by extension, functioned as an incentive to improve sleep habits and adjust daytime activities. However, inconsistent results have been reported in previous studies investigating the relationship between sleep and physical symptoms. In a cross-sectional study, there was no relationship between sleep variability and daytime fatigue and sleepiness [62], whereas daytime sleepiness was reduced among university students who were instructed to stabilize their sleep-wake schedule experimentally [63]. These mixed results may be due to the use of survey designs that evaluate symptoms at a single time point. In contrast, we measured physical symptoms several times per day using the EMA technique, with finer temporal resolution than in previous studies, and with ecological validity, resulting in the discovery of improved physical symptoms at wake-up time in the feedback group. When acquiring physical symptom data, the timing of the measurements can also be an important factor.

A previous study demonstrated that the use of wearable devices improved sleep duration in a healthy population [29]. However, the hierarchical Bayesian model demonstrated that the slope of day for sleep duration and midpoint of sleep were not statistically significant; thus, we did not find an improvement in sleep duration or midpoint of sleep by using only sleep feedback. This discrepancy may be due to the prompts of feedback messages. During the survey period, participants' sleep debt did not accumulate much, and they seemed to receive messages indicating that their sleep status was better. Under such circumstances, they might have attempted to maintain their sleep status by stabilizing their sleep-wake cycle rather than to improve their sleep duration or sleep timing. In future studies, it will be necessary to provide further support for prolonging sleep duration and advancing sleep timing, in addition to an objective sleep feedback message. Improvements in sleep duration and timing can also be beneficial to daytime functions, such as momentary moods and work performance, which were not improved in this study.

Future Directions

The HIT app is primarily designed for collecting multidimensional data in daily life, unlike mHealth apps developed for treating sleep disturbance. Therefore, in the HIT app, the functions useful for improving sleep disturbances are limited to objective feedback messages, while other apps offer several support functions, such as psychoeducation, sleep hygiene, and data visualization [30]. It is possible to improve sleep duration and timing by incorporating additional intervention options into the HIT app. Behavioral instructions to build sufficient sleep pressure at night are an example of additional support. In the treatment of patients with insomnia, sleep restriction therapy, which induces mild sleep deprivation to build homeostatic sleep pressure, is used, and existing mHealth studies incorporating sleep restriction therapy demonstrate a significant improvement in insomnia severity and sleep efficacy [35,64,65]. Interventions that build sufficient homeostatic sleep pressure as part of daily living, for example, exercising in the evening and avoiding long naps, may improve sleep duration and daytime functions, including work performance. Other interventions include behavioral coordination that works on the endogenous circadian rhythm; for example, adjusting the timing of food intake and avoiding exposure to bright light before bedtime. Such expansions can facilitate the control of habitual sleep timing and enhance the applicability of an IoT system with mobile devices for the treatment of various sleep disorders. Measuring, integrating, and using multidimensional information, including environmental and behavioral data, requires additional research. Simultaneously, integrative health care information systems, such as an HIT system, may provide a solution and expand intervention options, facilitating the verification of their effectiveness.

Limitations

This study had several limitations. First, the small number of participants were recruited from a life insurance company in Japan, and the social schedules of participants appeared similar, limiting the generalizability of the results. Indeed, work-related factors, including occupation, job stress, work hours, shift work, and physically demanding work, are associated with habitual sleep duration and sleep quality [66-68], suggesting that sleep habits may differ by occupation. Thus, a representative study including various occupations and lifestyles is required to ensure generalizability of the findings of this study.

Second, in this study, the exclusive computational method estimating sleep debt was introduced by summing the differences between the estimated and expected sleep duration. Although similar methods have been used to estimate sleep debt in other studies (the difference between self-reported sleep need and sleep duration on weekdays) [69,70], sleep debt computed using these methods may not reflect the neurobehavioral impairment caused by chronic sleep loss commonly observed in the experimental condition [71,72]. Initially, we adopted this definition with the aim of having participants plan their daily activities to improve the status of sleep insufficiency, but it could be possible that actual sleep debt was not accurately estimated. Therefore, a reliable method to estimate current sleep status should be developed and used in future studies. Especially, combining EMA techniques and machine learning methods is expected to provide reliable sleep measurements in daily life. For instance, a recent study reported that machine learning techniques could estimate daily sleep quality by using complex life data obtained from EMA questionnaires [73]. Collecting and integrating multidimensional information would be meaningful not only in understanding the covariant associations of sleep behavior with daytime symptoms but also for developing novel sleep measurements.

Finally, participants could identify the group they were assigned to because only the participants in the feedback group received feedback messages at 9 AM. Considering that physical symptoms were measured using self-report evaluations, EMA recordings were affected by cognitive biases such as the Hawthorne effect (the inclination of people who participate in an experimental study to change or improve their behavior only because it is being studied and not because of changes in the experimental stimulus). Optimizing the study design would help clarify this point. For example, by sending intervention messages that are not related to habitual sleep behavior to participants in the control group, we can determine whether the finding is based

XSL•FO
RenderX

on a specific response to sleep feedback messages. In addition, implementing a microrandomized trial [74], which is equivalent to a within-individual randomized controlled trial, could also be helpful in investigating the effectiveness of the interventions. By randomizing whether feedback messages are sent, we could examine the causal relationships between sleep behaviors and daytime functions.

Conclusions

We conducted an mHealth trial with office workers and demonstrated that objective push-type sleep feedback stabilizes sleep timing and improves physical symptoms at wake-up time. However, we did not find evidence of prolonged sleep duration, advanced sleep timing, or improved work performance. Future research should incorporate specific behavioral instructions intended to improve sleep duration and sleep timing with the current protocol and investigate behavioral instruction effectiveness by integrating and using multidimensional information collected as part of daily life.

Acknowledgments

The authors would like to thank the director who afforded them the opportunity to conduct this study at the organization in which the participants work. This study was supported in part by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology, Japan, under grant 20H00569 and grant 17H00878, and Japan Science of Technology Agency Precursory Research for Embryonic Science and Technology grant JPMJPR19J3, Japan.

Data Availability

The participants of this study did not agree to their data being shared publicly; therefore, supporting data are not available.

Authors' Contributions

TN, KY, AK, and YY designed the study. HT, KS, and AK collected data. HT performed data analysis and wrote the first draft of the manuscript. All authors critically reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Screenshots of the health care Internet of Things app.
[DOC File , 476 KB - mhealth_v10i10e39150_app1.doc]

Multimedia Appendix 2
Validity of the Sciencenet activity monitor.
[DOC File , 617 KB - mhealth_v10i10e39150_app2.doc]

Multimedia Appendix 3
Sleep hygiene guide.
[DOC File , 32 KB - mhealth_v10i10e39150_app3.doc]

Multimedia Appendix 4
Additional analyses of the statistical properties of sleep variables.
[DOC File , 572 KB - mhealth_v10i10e39150_app4.doc]

References

1. Becker S, Miron-Shatz T, Schumacher N, Krocza J, Diamantidis C, Albrecht U. mHealth 2.0: experiences, possibilities, and perspectives. JMIR Mhealth Uhealth 2014 May 16;2(2):e24 [FREE Full text] [doi: 10.2196/mhealth.3328] [Medline: 25099752]
2. Kumar S, Nilsen WJ, Abernethy A, Atienza A, Patrick K, Pavel M, et al. Mobile health technology evaluation: the mHealth evidence workshop. Am J Prev Med 2013 Aug;45(2):228-236 [FREE Full text] [doi: 10.1016/j.amepre.2013.03.017] [Medline: 23867031]
3. Nikolova S. Top 3 therapy fields with the best market potential for digital health apps. Research2Guidance. 2017. URL: https://research2guidance.com/top-3-therapy-fields-with-the-best-market-potential-for-digital-health-apps [accessed 2021-08-01]

XSL•FO
RenderX

4. Ganesan AN, Louise J, Horsfall M, Bilsborough SA, Hendriks J, McGavigan AD, et al. International mobile-health intervention on physical activity, sitting, and weight: the Stepathlon cardiovascular health study. J Am Coll Cardiol 2016 May 31;67(21):2453-2463 [FREE Full text] [doi: 10.1016/j.jacc.2016.03.472] [Medline: 27050185]
5. Patel MS, Volpp KG, Rosin R, Bellamy SL, Small DS, Heuer J, et al. A randomized, controlled trial of lottery-based financial incentives to increase physical activity among overweight and obese adults. Am J Health Promot 2018 Sep;32(7):1568-1575. [doi: 10.1177/0890117118758932] [Medline: 29534597]
6. Klasnja P, Smith S, Seewald NJ, Lee A, Hall K, Luers B, et al. Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of HeartSteps. Ann Behav Med 2019 May 03;53(6):573-582 [FREE Full text] [doi: 10.1093/abm/kay067] [Medline: 30192907]
7. Carrasco-Hernandez L, Jódar-Sánchez F, Núñez-Benjumea F, Moreno Conde J, Mesa González M, Civit-Balcells A, et al. A mobile health solution complementing psychopharmacology-supported smoking cessation: randomized controlled trial. JMIR Mhealth Uhealth 2020 Apr 27;8(4):e17530 [FREE Full text] [doi: 10.2196/17530] [Medline: 32338624]
8. Ubhi HK, Michie S, Kotz D, Wong WC, West R. A mobile app to aid smoking cessation: preliminary evaluation of SmokeFree28. J Med Internet Res 2015 Jan 16;17(1):e17 [FREE Full text] [doi: 10.2196/jmir.3479] [Medline: 25596170]
9. Tighe J, Shand F, Ridani R, Mackinnon A, De La Mata N, Christensen H. Ibobbly mobile health intervention for suicide prevention in Australian Indigenous youth: a pilot randomised controlled trial. BMJ Open 2017 Jan 27;7(1):e013518 [FREE Full text] [doi: 10.1136/bmjopen-2016-013518] [Medline: 28132007]
10. Shan Z, Ma H, Xie M, Yan P, Guo Y, Bao W, et al. Sleep duration and risk of type 2 diabetes: a meta-analysis of prospective studies. Diabetes Care 2015 Mar;38(3):529-537. [doi: 10.2337/dc14-2073] [Medline: 25715415]
11. Cappuccio FP, Cooper D, D'Elia L, Strazzullo P, Miller MA. Sleep duration predicts cardiovascular outcomes: a systematic review and meta-analysis of prospective studies. Eur Heart J 2011 Jun;32(12):1484-1492. [doi: 10.1093/eurheartj/ehr007] [Medline: 21300732]
12. Zhai L, Zhang H, Zhang D. Sleep duration and depression among adults: a meta-analysis of prospective studies. Depress Anxiety 2015 Sep;32(9):664-670. [doi: 10.1002/da.22386] [Medline: 26047492]
13. Chami HA, Ghandour B, Isma'eel H, Nasreddine L, Nasrallah M, Tamim H. Sleepless in Beirut: sleep duration and associated subjective sleep insufficiency, daytime fatigue, and sleep debt in an urban environment. Sleep Breath 2020 Mar;24(1):357-367. [doi: 10.1007/s11325-019-01833-3] [Medline: 31028521]
14. Alves FR, de Souza EA, de França Ferreira LG, de Oliveira Vilar Neto J, de Bruin VM, de Bruin PF. Sleep duration and daytime sleepiness in a large sample of Brazilian high school adolescents. Sleep Med 2020 Feb;66:207-215. [doi: 10.1016/j.sleep.2019.08.019] [Medline: 31978864]
15. Ishibashi Y, Shimura A. Association between work productivity and sleep health: a cross-sectional study in Japan. Sleep Health 2020 Jun;6(3):270-276. [doi: 10.1016/j.sleh.2020.02.016] [Medline: 32360021]
16. Magee CA, Caputi P, Iverson DC. Relationships between self-rated health, quality of life and sleep duration in middle aged and elderly Australians. Sleep Med 2011 Apr;12(4):346-350. [doi: 10.1016/j.sleep.2010.09.013] [Medline: 21388876]
17. Sasaki N, Fujiwara S, Yamashita H, Ozono R, Monzen Y, Teramen K, et al. Association between obesity and self-reported sleep duration variability, sleep timing, and age in the Japanese population. Obes Res Clin Pract 2018;12(2):187-194. [doi: 10.1016/j.orcp.2017.10.008] [Medline: 29128435]
18. Yan B, Li R, Li J, Jin X, Gao F, Gao Y, et al. Sleep timing may predict congestive heart failure: a community-based cohort study. J Am Heart Assoc 2021 Mar 16;10(6):e018385 [FREE Full text] [doi: 10.1161/JAHA.120.018385] [Medline: 33666090]
19. Reutrakul S, Siwasaranond N, Nimitphong H, Saetung S, Chirakalwasan N, Ongphiphadhanakul B, et al. Relationships among sleep timing, sleep duration and glycemic control in type 2 diabetes in Thailand. Chronobiol Int 2015;32(10):1469-1476. [doi: 10.3109/07420528.2015.1105812] [Medline: 26595483]
20. Levandovski R, Dantas G, Fernandes LC, Caumo W, Torres I, Roenneberg T, et al. Depression scores associate with chronotype and social jetlag in a rural population. Chronobiol Int 2011 Nov;28(9):771-778. [doi: 10.3109/07420528.2011.602445] [Medline: 21895489]
21. de Souza CM, Hidalgo MP. Midpoint of sleep on school days is associated with depression among adolescents. Chronobiol Int 2014 Mar;31(2):199-205. [doi: 10.3109/07420528.2013.838575] [Medline: 24156519]
22. Bei B, Wiley JF, Trinder J, Manber R. Beyond the mean: a systematic review on the correlates of daily intraindividual variability of sleep/wake patterns. Sleep Med Rev 2016 Aug;28:108-124 [FREE Full text] [doi: 10.1016/j.smrv.2015.06.003] [Medline: 26588182]
23. Bei B, Seeman TE, Carroll JE, Wiley JF. Sleep and physiological dysregulation: a closer look at sleep intraindividual variability. Sleep 2017 Sep 01;40(9):zsx109 [FREE Full text] [doi: 10.1093/sleep/zsx109] [Medline: 28651371]
24. Okun ML, Reynolds CF, Buysse DJ, Monk TH, Mazumdar S, Begley A, et al. Sleep variability, health-related practices, and inflammatory markers in a community dwelling sample of older adults. Psychosom Med 2011;73(2):142-150 [FREE Full text] [doi: 10.1097/PSY.0b013e3182020d08] [Medline: 21097658]
25. Slavish DC, Taylor DJ, Lichstein KL. Intraindividual variability in sleep and comorbid medical and mental health conditions. Sleep 2019 Jun 11;42(6):zsz052 [FREE Full text] [doi: 10.1093/sleep/zsz052] [Medline: 30843059]

XSL•FO
RenderX

26. Lemola S, Ledermann T, Friedman EM. Variability of sleep duration is related to subjective sleep quality and subjective well-being: an actigraphy study. PLoS One 2013 Aug 14;8(8):e71292 [FREE Full text] [doi: 10.1371/journal.pone.0071292] [Medline: 23967186]
27. Perez-Pozuelo I, Zhai B, Palotti J, Mall R, Aupetit M, Garcia-Gomez JM, et al. The future of sleep health: a data-driven revolution in sleep science and medicine. NPJ Digit Med 2020 Mar 23;3:42 [FREE Full text] [doi: 10.1038/s41746-020-0244-4] [Medline: 32219183]
28. Tobin SY, Williams PG, Baron KG, Halliday TM, Depner CM. Challenges and opportunities for applying wearable technology to sleep. Sleep Med Clin 2021 Dec;16(4):607-618. [doi: 10.1016/j.jsmc.2021.07.002] [Medline: 34711385]
29. Crowley O, Pugliese L, Kachnowski S. The impact of wearable device enabled health initiative on physical activity and sleep. Cureus 2016 Oct 11;8(10):e825 [FREE Full text] [doi: 10.7759/cureus.825] [Medline: 27882272]
30. Aji M, Gordon C, Stratton E, Calvo RA, Bartlett D, Grunstein R, et al. Framework for the design engineering and clinical implementation and evaluation of mHealth apps for sleep disturbance: systematic review. J Med Internet Res 2021 Feb 17;23(2):e24607 [FREE Full text] [doi: 10.2196/24607] [Medline: 33595441]
31. Arroyo AC, Zawadzki MJ. The implementation of behavior change techniques in mHealth apps for sleep: systematic review. JMIR Mhealth Uhealth 2022 Apr 04;10(4):e33527 [FREE Full text] [doi: 10.2196/33527] [Medline: 35377327]
32. Simon L, Reimann J, Steubl LS, Stach M, Spiegelhalder K, Sander LB, et al. Help for insomnia from the app store? A standardized rating of mobile health applications claiming to target insomnia. J Sleep Res (forthcoming) 2022 May 27:e13642. [doi: 10.1111/jsr.13642] [Medline: 35624078]
33. Pulantara IW, Parmanto B, Germain A. Development of a just-in-time adaptive mHealth intervention for insomnia: usability study. JMIR Hum Factors 2018 May 17;5(2):e21 [FREE Full text] [doi: 10.2196/humanfactors.8905] [Medline: 29773529]
34. Pulantara IW, Parmanto B, Germain A. Clinical feasibility of a just-in-time adaptive intervention app (iREST) as a behavioral sleep treatment in a military population: feasibility comparative effectiveness study. J Med Internet Res 2018 Dec 07;20(12):e10124 [FREE Full text] [doi: 10.2196/10124] [Medline: 30530452]
35. Horsch CH, Lancee J, Griffioen-Both F, Spruit S, Fitrianie S, Neerincx MA, et al. Mobile phone-delivered cognitive behavioral therapy for insomnia: a randomized waitlist controlled trial. J Med Internet Res 2017 Apr 11;19(4):e70 [FREE Full text] [doi: 10.2196/jmir.6524] [Medline: 28400355]
36. Kuhn E, Miller KE, Puran D, Wielgosz J, YorkWilliams SL, Owen JE, et al. A pilot randomized controlled trial of the insomnia coach mobile app to assess its feasibility, acceptability, and potential efficacy. Behav Ther 2022 May;53(3):440-457. [doi: 10.1016/j.beth.2021.11.003] [Medline: 35473648]
37. Glazer Baron K, Culnan E, Duffecy J, Berendson M, Cheung Mason I, Lattie E, et al. How are consumer sleep technology data being used to deliver behavioral sleep medicine interventions? A systematic review. Behav Sleep Med 2022;20(2):173-187. [doi: 10.1080/15402002.2021.1898397] [Medline: 33757392]
38. Kang SG, Kang JM, Cho SJ, Ko KP, Lee YJ, Lee HJ, et al. Cognitive behavioral therapy using a mobile application synchronizable with wearable devices for insomnia treatment: a pilot study. J Clin Sleep Med 2017 Apr 15;13(4):633-640 [FREE Full text] [doi: 10.5664/jcsm.6564] [Medline: 28162145]
39. Reilly ED, Robinson SA, Petrakis BA, Kuhn E, Pigeon WR, Wiener RS, et al. Mobile app use for insomnia self-management: pilot findings on sleep outcomes in veterans. Interact J Med Res 2019 Jul 24;8(3):e12408 [FREE Full text] [doi: 10.2196/12408] [Medline: 31342904]
40. Murawski B, Plotnikoff RC, Rayward AT, Oldmeadow C, Vandelanotte C, Brown WJ, et al. Efficacy of an m-health physical activity and sleep health intervention for adults: a randomized waitlist-controlled trial. Am J Prev Med 2019 Oct;57(4):503-514. [doi: 10.1016/j.amepre.2019.05.009] [Medline: 31542128]
41. Stone AA, Shiffman S. Ecological momentary assessment (EMA) in behavioral medicine. Ann Behav Med 1994;16(3):199-202. [doi: 10.1093/abm/16.3.199]
42. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
43. Qian K, Koike T, Nakamura T, Schuller B, Yamamoto Y. Learning multimodal representations for drowsiness detection. IEEE Trans Intell Transport Syst 2022 Aug;23(8):11539-11548. [doi: 10.1109/tits.2021.3105326]
44. Beck AT, Steer RA, Brown GK. BDI-II: Beck Depression Inventory. 2nd edition. San Antonio, TX, USA: Psychological Corporation; 1996.
45. Kojima M, Furukawa TA, Takahashi H, Kawai M, Nagaya T, Tokudome S. Cross-cultural validation of the Beck Depression Inventory-II in Japan. Psychiatry Res 2002 Jul 31;110(3):291-299. [doi: 10.1016/s0165-1781(02)00106-3] [Medline: 12127479]
46. Hidano T, Fukuhara M, Iwawaki S, Soga S, Spielberger CD. New STAI for the State-Trait Anxiety Inventory-Form JYZ. Tokyo, Japan: Jitsumu Kyouiku Press; 2000.
47. Buysse DJ, Reynolds 3rd CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989 May;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]

XSL•FO
RenderX

48. Doi Y, Minowa M, Uchiyama M, Okawa M, Kim K, Shibui K, et al. Psychometric assessment of subjective sleep quality using the Japanese version of the Pittsburgh Sleep Quality Index (PSQI-J) in psychiatric disordered and control subjects. Psychiatry Res 2000 Dec 27;97(2-3):165-172. [doi: 10.1016/s0165-1781(00)00232-8] [Medline: 11166088]
49. Mollayeva T, Thurairajah P, Burton K, Mollayeva S, Shapiro CM, Colantonio A. The Pittsburgh sleep quality index as a screening tool for sleep dysfunction in clinical and non-clinical samples: a systematic review and meta-analysis. Sleep Med Rev 2016 Feb;25:52-73. [doi: 10.1016/j.smrv.2015.01.009] [Medline: 26163057]
50. Kessler RC, Barber C, Beck A, Berglund P, Cleary PD, McKenas D, et al. The World Health Organization health and work performance questionnaire (HPQ). J Occup Environ Med 2003 Feb;45(2):156-174. [doi: 10.1097/01.jom.0000052967.43131.51] [Medline: 12625231]
51. Kawakami N, Inoue A, Tsuchiya M, Watanabe K, Imamura K, Iida M, et al. Construct validity and test-retest reliability of the World Mental Health Japan version of the World Health Organization Health and Work Performance Questionnaire Short Version: a preliminary study. Ind Health 2020 Aug 07;58(4):375-387 [FREE Full text] [doi: 10.2486/indhealth.2019-0090] [Medline: 32173661]
52. Fukui I. The Depression and Anxiety Mood Scale (DAMS): scale development and validation. Jpn J Behav Ther 1997 Sep;23(2):83-93.
53. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. Sleep 2003 May 01;26(3):342-392. [doi: 10.1093/sleep/26.3.342] [Medline: 12749557]
54. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic sleep/wake identification from wrist activity. Sleep 1992 Oct;15(5):461-469. [doi: 10.1093/sleep/15.5.461] [Medline: 1455130]
55. Webster JB, Kripke DF, Messin S, Mullaney DJ, Wyborney G. An activity-based sleep monitor system for ambulatory use. Sleep 1982;5(4):389-399. [doi: 10.1093/sleep/5.4.389] [Medline: 7163726]
56. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: a probabilistic programming language. J Stat Soft 2017 Jan 11;76(1):1-32. [doi: 10.18637/jss.v076.i01]
57. Kuznetsova A, Brockhoff PB, Christensen RH. lmerTest package: tests in linear mixed effects models. J Stat Soft 2017 Dec 6;82(13):1-26. [doi: 10.18637/jss.v082.i13]
58. Lenth RV. Least-squares means: the R package lsmeans. J Stat Soft 2016 Jan 29;69(1):1-33. [doi: 10.18637/jss.v069.i01]
59. Baron KG, Reid KJ. Circadian misalignment and health. Int Rev Psychiatry 2014 Apr;26(2):139-154 [FREE Full text] [doi: 10.3109/09540261.2014.911149] [Medline: 24892891]
60. Wittmann M, Dinich J, Merrow M, Roenneberg T. Social jetlag: misalignment of biological and social time. Chronobiol Int 2006;23(1-2):497-509. [doi: 10.1080/07420520500545979] [Medline: 16687322]
61. Roenneberg T, Pilz LK, Zerbini G, Winnebeck EC. Chronotype and social jetlag: a (self-) critical review. Biology (Basel) 2019 Jul 12;8(3):54 [FREE Full text] [doi: 10.3390/biology8030054] [Medline: 31336976]
62. Kang JH, Chen SC. Effects of an irregular bedtime schedule on sleep quality, daytime sleepiness, and fatigue among university students in Taiwan. BMC Public Health 2009 Jul 19;9:248 [FREE Full text] [doi: 10.1186/1471-2458-9-248] [Medline: 19615098]
63. Manber R, Bootzin RR, Acebo C, Carskadon MA. The effects of regularizing sleep-wake schedules on daytime sleepiness. Sleep 1996 Jun;19(5):432-441. [doi: 10.1093/sleep/19.5.432] [Medline: 8843535]
64. Miller CB, Espie CA, Epstein DR, Friedman L, Morin CM, Pigeon WR, et al. The evidence base of sleep restriction therapy for treating insomnia disorder. Sleep Med Rev 2014 Oct;18(5):415-424. [doi: 10.1016/j.smrv.2014.01.006] [Medline: 24629826]
65. Aji M, Glozier N, Bartlett D, Peters D, Calvo RA, Zheng Y, et al. A feasibility study of a mobile app to treat insomnia. Transl Behav Med 2021 Mar 16;11(2):604-612. [doi: 10.1093/tbm/ibaa019] [Medline: 32227087]
66. Akerstedt T, Knutsson A, Westerholm P, Theorell T, Alfredsson L, Kecklund G. Sleep disturbances, work stress and work hours: a cross-sectional study. J Psychosom Res 2002 Sep;53(3):741-748. [doi: 10.1016/s0022-3999(02)00333-1] [Medline: 12217447]
67. Luckhaupt SE, Tak S, Calvert GM. The prevalence of short sleep duration by industry and occupation in the National Health Interview Survey. Sleep 2010 Feb;33(2):149-159 [FREE Full text] [doi: 10.1093/sleep/33.2.149] [Medline: 20175398]
68. Shockey TM, Wheaton AG. Short sleep duration by occupation group - 29 states, 2013-2014. MMWR Morb Mortal Wkly Rep 2017 Mar 03;66(8):207-213 [FREE Full text] [doi: 10.15585/mmwr.mm6608a2] [Medline: 28253230]
69. Léger D, Roscoat ED, Bayon V, Guignard R, Pâquereau J, Beck F. Short sleep in young adults: insomnia or sleep debt? Prevalence and clinical description of short sleep in a representative sample of 1004 young adults from France. Sleep Med 2011 May;12(5):454-462. [doi: 10.1016/j.sleep.2010.12.012] [Medline: 21474376]
70. Regestein Q, Natarajan V, Pavlova M, Kawasaki S, Gleason R, Koff E. Sleep debt and depression in female college students. Psychiatry Res 2010 Mar 30;176(1):34-39. [doi: 10.1016/j.psychres.2008.11.006] [Medline: 20079935]
71. Belenky G, Wesensten NJ, Thorne DR, Thomas ML, Sing HC, Redmond DP, et al. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. J Sleep Res 2003 Mar;12(1):1-12 [FREE Full text] [doi: 10.1046/j.1365-2869.2003.00337.x] [Medline: 12603781]

XSL•FO
RenderX

72. Van Dongen HP, Maislin G, Mullington JM, Dinges DF. The cumulative cost of additional wakefulness: dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. Sleep 2003 Mar 15;26(2):117-126. [doi: 10.1093/sleep/26.2.117] [Medline: 12683469]
73. Lim J, Jeong CY, Lim JM, Chung S, Kim G, Noh KJ, et al. Assessing sleep quality using mobile EMAs: opportunities, practical consideration, and challenges. IEEE Access 2022 Apr 1;10:2063-2076. [doi: 10.1109/ACCESS.2021.3140074]
74. Klasnja P, Hekler EB, Shiffman S, Boruvka A, Almirall D, Tewari A, et al. Microrandomized trials: an experimental design for developing just-in-time adaptive interventions. Health Psychol 2015 Dec;34S:1220-1228 [FREE Full text] [doi: 10.1037/hea0000305] [Medline: 26651463]

Abbreviations

BDI-II: Beck Depression Inventory second edition
BLE: Bluetooth Low Energy
CI: credible interval
EMA: ecological momentary assessment
HIT: health care Internet of Things
HPQ: Health and Work Performance Questionnaire
IoT: Internet of Things
mHealth: mobile health
PSQI: Pittsburgh Sleep Quality Index
SPF: Sleep Probability Function
STAI: State-Trait Anxiety Inventory

Edited by L Buis; submitted 13.05.22; peer-reviewed by M Aji, Q Chen; comments to author 08.06.22; revised version received 12.08.22; accepted 24.08.22; published 06.10.22.

Please cite as:
Takeuchi H, Suwa K, Kishi A, Nakamura T, Yoshiuchi K, Yamamoto Y
The Effects of Objective Push-Type Sleep Feedback on Habitual Sleep Behavior and Momentary Symptoms in Daily Life: mHealth Intervention Trial Using a Health Care Internet of Things System
JMIR Mhealth Uhealth 2022;10(10):e39150
URL: https://mhealth.jmir.org/2022/10/e39150
doi:10.2196/39150
PMID:36201383

©Hiroki Takeuchi, Kaori Suwa, Akifumi Kishi, Toru Nakamura, Kazuhiro Yoshiuchi, Yoshiharu Yamamoto. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 06.10.2022. 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.

XSL•FO
RenderX

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org

XSL•FO
RenderX