Contents

Original Papers

An Luu, Truong Nguyen, Van Cao, Trinh Ha, Lien Chung, Trung Truong, Tung Nguyen Le Nhu, Khoa Dao, Hao Nguyen, Phan Khanh, Khanh Le, Luu Tran, Phung Nhat, Duc Tran, Yen Lam, Catherine Thwaites, Jacob McKnight, Nguyen Vinh Chau, Jennifer Van Nui, Vietnam ICU Translational Applications Laboratory (VITAL) .................................................. 4

Exploring a Gaming-Based Intervention for Unemployed Young Adults: Thematic Analysis (e44423)
Ingjerd Straand, Asbjørn Følstad, Jone Bjørnestad .................................................. 14

User Perceptions of Visual Clot in a High-Fidelity Simulation Study: Mixed Qualitative-Quantitative Study (e47991)
Greta Gasciauskaite, Clara Castellucci, Amos Malorgio, Alexandra Budowski, Giovanna Schweiger, Michaela Kolbe, Daniel Fries, Bastian Grande, Christoph Nöthiger, Donat Spahn, Tadzio Roche, David Tscholl, Samira Akbas .................................................. 34

Digital Care Pathway for Patients With Sleep Apnea in Specialized Care: Mixed Methods Study (e47809)
Jari Haverinen, Terttu Harju, Hanna Mikkonen, Pia Liljamo, Miia Turpeinen, Jarmo Reponen .................................................. 46

Leveraging Generative AI Tools to Support the Development of Digital Solutions in Health Care Research: Case Study (e52885)
Danissa Rodriguez, Katharine Lawrence, Javier Gonzalez, Beatrix Brandfield-Harvey, Lynn Xu, Sumaiya Tasneem, Defne Levine, Devin Mann .................................................. 6

Testing Two Online Symptom Checkers With Vulnerable Groups: Usability Study to Improve Cognitive Accessibility of eHealth Services (e45275)
Klara Albrink, Dominik Schröder, Carla Joos, Frank Müller, Eva Noack .................................................. 79

The Role of Coherent Robot Behavior and Embodiment in Emotion Perception and Recognition During Human-Robot Interaction: Experimental Study (e45494)
Laura Fiorini, Grazia D’Onofrio, Alessandra Sorrentino, Federica Cornacchia Loizzo, Sergio Russo, Filomena Ciccone, Francesco Giuliani, Daniele Sancarlo, Filippo Cavallo .................................................. 92

Physical Therapists’ Acceptance of a Wearable, Fabric-Based Sensor System (Motion Tape) for Use in Clinical Practice: Qualitative Focus Group Study (e55246)
Audrey Lee, Patricia D’Onofrio, Emilia Farcas, Job Godino, Kevin Patrick, Elijah Wyckoff, Kenneth Loh, Sara Gomboatto .................................................. 106

Usability of an App for Medical History Taking in General Practice From the Patients’ Perspective: Cross-Sectional Study (e47755)
Klara Albrink, Dominik Schröder, Carla Joos, Frank Müller, Eva Noack .................................................. 121
<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Novel Continuous Real-Time Vital Signs Viewer for Intensive Care Units: Design and Evaluation Study</td>
<td>(e46030) Shiming Yang, Samuel Galvagn, Neeraj Badjatia, Deborah Stein, William Teeter, Thomas Scalea, Stacy Shackelford, Raymond Fang, Catriona Miller, Peter Hu, VS viewer study group.</td>
<td>136</td>
</tr>
<tr>
<td>Characterizing and Comparing Adverse Drug Events Documented in 2 Spontaneous Reporting Systems in the Lower Mainland of British Columbia, Canada: Retrospective Observational Study</td>
<td>(e52495) Erica Lau, Amber Cragg, Serena Small, Katherine Butcher, Corinne Hohl.</td>
<td>148</td>
</tr>
<tr>
<td>Usability and Evaluation of a Health Information System in the Emergency Department: Mixed Methods Study</td>
<td>(e48445) Christina Østervang, Charlotte Jensen, Elisabeth Coyne, Karin Dieperink, Annmarie Lassen.</td>
<td>164</td>
</tr>
<tr>
<td>Intention to Use an Electronic Community Health Information System Among Health Extension Workers in Rural Northwest Ethiopia: Cross-Sectional Study Using the Unified Theory of Acceptance and Use of Technology 2 Model</td>
<td>(e47081) Tesfahun Hailemariam, Asmamaw Atnafu, Lemma Gezie, Jens Kaasbøll, Jorn Klein, Binyam Tilahun.</td>
<td>178</td>
</tr>
<tr>
<td>Interest in mHealth Among Patients With Low Back Pain: Cross-Sectional Study</td>
<td>(e48729) Jonas Ipsen, Louise Sandal, Natalie Chang, Berit Schiøttz-Christensen, Karen Søgaard, Anders Hansen.</td>
<td>192</td>
</tr>
<tr>
<td>User-Centered Design and Usability of a Culturally Adapted Virtual Survivorship Care App for Chinese Canadian Prostate Cancer Survivors: Qualitative Descriptive Study</td>
<td>(e49353) Karen Young, Ting Xiong, Rachel Lee, Ananya Banerjee, Myles Leslie, Wellam Ko, Quynh Pham.</td>
<td>202</td>
</tr>
<tr>
<td>Effectiveness and User Perception of an In-Vehicle Voice Warning for Hypoglycemia: Development and Feasibility Trial</td>
<td>(e42823) Caterina Bérubé, Vera Lehmann, Martin Maritsch, Mathias Kraus, Stefan Feuerriegel, Felix Wortmann, Thomas Züger, Christoph Stettler, Elgar Fleisch, A Kocaballi, Tobias Kowatsch.</td>
<td>213</td>
</tr>
<tr>
<td>A Closed-Loop Falls Monitoring and Prevention App for Multiple Sclerosis Clinical Practice: Human-Centered Design of the Multiple Sclerosis Falls InsightTrack</td>
<td>(e49331) Valerie Block, Kanishka Koshal, Jaelenee Wijangco, Nicolette Miller, Narendar Sara, Kyra Henderson, Jennifer Reihm, Arpita Gopal, Sonam Mohan, Jeffrey Geftland, Chu-Yueh Guo, Lauren Oommen, Alyssa Nylander, James Rowson, Ethan Brown, Stephen Sanders, Katherine Rankin, Courtney Lyles, Ida Sim, Riley Bove.</td>
<td>229</td>
</tr>
<tr>
<td>Digital Triage Tools for Sexually Transmitted Infection Testing Compared With General Practitioners’ Advice: Vignette-Based Qualitative Study With Interviews Among General Practitioners</td>
<td>(e49221) Kyma Schnoor, Anke Versluis, Niels Chavannes, Esther Talboom-Kamp.</td>
<td>247</td>
</tr>
<tr>
<td>Usability and User Experience of an mHealth App for Therapy Support of Patients With Breast Cancer: Mixed Methods Study Using Eye Tracking</td>
<td>(e50926) Carolin Anders, Preetha Moorthy, Laura Svensson, Julia Müller, Oliver Heinze, Petra Knaup, Markus Wallwiener, Thomas Deutsch, Thao-Vy Le, Lina Weinert.</td>
<td>273</td>
</tr>
<tr>
<td>Feasibility and Design Factors for Home-Based Pulmonary Rehabilitation of Patients With Chronic Obstructive Pulmonary Disease and Chronic Lung Diseases Based on a People-Object-Environment Framework: Qualitative Interview Study</td>
<td>(e51150) Shih-Ying Chien, Alice Wong, Winston Tseng, Han-Chung Hu, Hsiu-Ying Cho.</td>
<td>290</td>
</tr>
</tbody>
</table>
Perspectives on the COVID-19 Vaccination Rollout in 17 Countries: Reflexive Thematic and Frequency Analysis Based on the Strengths, Weaknesses, Opportunities, and Threats (SWOT) Framework (e44258)
Vanja Kopilaš, Khrystyna Nasadiuk, Lucia Martinelli, Lenka Lhotska, Zoran Todorovic, Matjaz Vidmar, Helena Machado, Anna Svalastog, Sreko Gajovi ............................................................... 304

The Effect of a Video-Assisted Health Education Program Followed by Peer Education on the Health Literacy of COVID-19 and Other Infectious Diseases Among School Children: Quasi-Randomized Controlled Trial (e43943)
Xiaojuan Zhang, Yingkun Wen, Ning Han, Yawen Jiang ............................................................... 319

Characteristics, Opportunities, and Challenges of Osteopathy Based on the Perceptions of Osteopaths in Austria: Qualitative Interview Study (e45302)
Jonas Manschel, Jan Porthun, Ulrike Winkler, Jean Beuckels, David Martin ............................................................... 330

Trust in and Acceptance of Artificial Intelligence Applications in Medicine: Mixed Methods Study (e47031)
Daria Shevtsova, Anam Ahmed, Iris Boot, Carmen Sanges, Michael Hudecek, John Jacobs, Simon Hort, Hubertus Vrijhoef ............................................................... 344

A Machine Learning Approach with Human-AI Collaboration for Automated Classification of Patient Safety Event Reports: Algorithm Development and Validation Study (e53378)
Hongbo Chen, Eldan Cohen, Dulane Wilson, Myrtede Alfred ............................................................... 361

Testing the Feasibility and Acceptability of Using an Artificial Intelligence Chatbot to Promote HIV Testing and Pre-Exposure Prophylaxis in Malaysia: Mixed Methods Study (e52055)
Min Cheah, Yan Gan, Frederick Altice, Jeffrey Wickersham, Roman Shrestha, Nur Salleh, Kee Ng, Iskandar Azwa, Vimala Balakrishnan, Adeeba Kamarulzaman, Zhao Ni ............................................................... 374

Viewpoint

The Temperature Feature of ChatGPT: Modifying Creativity for Clinical Research (e53559)
Joshua Davis, Liesbet Van Bulck, Brigitte Durieux, Charlotte Lindvall ............................................................... 390

An Phuoc Luu¹, MPH; Truong Thanh Nguyen², MD; Van Thi Cam Cao², RN; Trinh Hoang Diem Ha², RN; Lien Thi Thu Chung², RN; Trung Ngoc Truong², MD; Tung Nguyen Le Nhu³, MD; Khoa Bach Dao², MD; Hao Van Nguyen⁴, MD; Phan Nguyen Quoc Khanh¹, MD; Khanh Thuy Thuy Le¹, MD; Luu Hoai Bao Tran¹, MD; Phung Tran Huy Nhat¹, BSc; Duc Minh Tran¹, PMD; Yen Minh Lam¹, MD; Catherine Louise Thwaites¹,³, BSc, MBBS, MD; Jacob Mcknight³, BEng, MSc, DPhil; Nguyen Van Vinh Chau¹,², MD; Jennifer Ilo Van Nui³, BA, MA, PhD; Vietnam ICU Translational Applications Laboratory (VITAL)⁴

¹Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam
²Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam
³Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
⁴See Acknowledgments

Corresponding Author:
Jennifer Ilo Van Nui, BA, MA, PhD
Oxford University Clinical Research Unit
764 Vo Van Kiet
Ho Chi Minh City, 70000
Vietnam
Phone: 84 362620124
Email: jvannuil@oucru.org

Abstract

Background: Wearable devices have been used extensively both inside and outside of the hospital setting. During the COVID-19 pandemic, in some contexts, there was an increased need to remotely monitor pulse and saturated oxygen for patients due to the lack of staff and bedside monitors.

Objective: A prototype of a remote monitoring system using wearable pulse oximeter devices was implemented at the Hospital for Tropical Diseases in Ho Chi Minh City, Vietnam, from August to December 2021. The aim of this work was to support the ongoing implementation of the remote monitoring system.

Methods: We used an action learning approach with rapid pragmatic methods, including informal discussions and observations as well as a feedback survey form designed based on the technology acceptance model to assess the use and acceptability of the system. Based on these results, we facilitated a meeting using user-centered design principles to explore user needs and ideas about its development in more detail.

Results: In total, 21 users filled in the feedback form. The mean technology acceptance model scores ranged from 3.5 (for perceived ease of use) to 4.4 (for attitude) with behavioral intention (3.8) and perceived usefulness (4.2) scoring in between. Those working as nurses scored higher on perceived usefulness, attitude, and behavioral intention than did physicians. Based on informal discussions, we realized there was a mismatch between how we (ie, the research team) and the ward teams perceived the use and wider purpose of the technology.

Conclusions: Designing and implementing the devices to be more nurse-centric from their introduction could have helped to increase their efficiency and use during the complex pandemic period.

(JMIR Hum Factors 2024;11:e44619) doi:10.2196/44619

KEYWORDS
vital signs; wearable devices; action learning; technology acceptance model; TAM; COVID-19; user-centered design; wearables; remote monitoring; technology acceptance; oximeter
Introduction

The popularity of portable wearable technologies that monitor health has increased substantially over the past decade due to their perceived utility, relatively simple implementation, and immediate feedback [1]. Wearable technology is used in both personal and clinical settings, and more recently in the context of the COVID-19 pandemic for diagnosis, remote monitoring, and other applications in both inpatient and outpatient settings [1-3]. Using wearable devices for COVID-19 care can result in infection control by reducing the amount of time that health care workers (HCWs) are physically with patients and providing continuous monitoring of vital signs for the early identification and potential treatment of deteriorating patients [2]. Specifically, remote monitoring of oxygen saturation using wearable devices became increasingly common during COVID-19 in hospital settings [4,5].

Despite the potential benefits, there have been many challenges noted in implementing and using wearable devices during COVID-19, including technical, social, and political spheres [1]. Technical challenges often include battery life, Wi-Fi or Bluetooth connections, and device communication. A few examples of social challenges are patients lacking technological confidence (eg, in older patients) and repeated device alerts or continuous monitoring making patients nervous, while political challenges could include regulatory issues for expanding the approval of devices for COVID-19-related medical situations [1]. Many of these challenges may be enhanced in low- and middle-income countries (LMICs), while the need for such integration is crucial, especially during pandemic situations [6,7].

There are several studies exploring the technical challenges of integrating wearable devices in trial settings during COVID-19 (eg, see [8]), but there is a lack of research surrounding the acceptability of such devices within these contexts and how attitudes may impact actual use [9]. Portable wearable devices could be a potential solution to allow for continuous monitoring of vital signs remotely and affordably for COVID-19 wards in LMIC settings; however, while advantageous, these devices cannot meet their full potential if the users do not agree to use them or realize their potential value [10]. Understanding user perceptions and needs as well as the context in which the technology is implemented is crucial for successful implementation [1]. User-centered approaches stress the importance of integrating both human factors and technical factors [11] while also paying attention to avoid excluding certain populations in the design [7]. User-centered approaches have been cited as a “critical success factor” in a variety of health-related technology projects [12].

From August to December 2021, when COVID-19 cases were increasing more rapidly than at any time previously in Ho Chi Minh City, Vietnam, there was an opportunity to integrate a prototype wearable device and monitoring system into the COVID-19-designated wards at the Hospital for Tropical Diseases (HTD). At this time, the HTD was overwhelmed with patients with COVID-19 and we needed to deploy something urgently that could help. Using pragmatic methods during the rollout of the device, we describe stakeholders’ use of the wearable device, aspects of acceptability, and under which circumstances its use would be most beneficial for improving the care of patients with COVID-19. The primary objective of this work was to support the implementation process of the wearable device in the hospital to improve patient care during a catastrophic period of the COVID-19 pandemic in Ho Chi Minh City, Vietnam.

Methods

Study Setting

This work took place within a larger project called the Vietnam ICU Translation Application Laboratory (VITAL) at the Oxford University Clinical Research Unit (OUCRU) and HTD. The goal of VITAL is to design and implement innovative technologies to improve patient care within the intensive care unit (ICU) at the HTD, with a longer-term goal of expanding these technologies regionally. In addition to the clinical and technological studies, there is an ethnographic study to explore the sociotechnical contexts of the ICU at the HTD and within ICUs in Vietnam more broadly. The VITAL multidisciplinary team was in place at the start of the COVID-19 pandemic.

In the first 100 days of the COVID-19 pandemic, Vietnam rapidly implemented a variety of public health measures resulting in relatively few cases and zero deaths [13]. Since that time, there were a few concentrated outbreaks (for example, in Da Nang in July 2020 and December 2020 in northern Vietnam). In May 2021, the cases started to increase on a countrywide basis, and by August 2021, the hospitals began to fill with patients with COVID-19. It was within this pandemic context that the wearable device was implemented at the HTD, and the VITAL study teams worked together throughout to improve its implementation.

The wearable device was selected by the company and was already integrated into a locally developed platform based on an available application programming interface, licensing, and availability. The device was medical grade and measured heart rate and blood oxygen levels, similar to a pulse oximeter. The wearable device was battery powered and each one connected to a tablet that was kept at the patients’ bedside. The tablets had a 3G or 4G sim card and sent the data to a cloud where multiple patients’ data were viewable by HCWs outside the patients’ rooms and isolation area.

Study Design

The aim of this work was to support the ongoing implementation of the wearable device rather than to follow a predefined, replicable study protocol, as would be used in trial settings, for example. Therefore, the work here describes the pragmatic rollout of the device. We used an action learning approach, including integration of multiple methods to assess the use and acceptability of the wearable device [14]. Action learning approaches rely on an iterative process of assessing local contexts, learning from relevant stakeholders, and using the information to improve an implementation or further develop a technology specific to the context [15,16]. As the wearable device started to be implemented in the HTD wards, our team
of HCWs, social scientists, and technology developers took the opportunity to work together to inform the implementation. Therefore, we adapted the methods as the situation changed and more insights were gained [14].

**Participants**

Potential participants included the HCWs from the HTD who were using the device in the wards during the implementation and corresponding ward heads. We estimated that a total of 30 doctors and 60 nurses would have worked in the wards where the wearable device was implemented and potentially used it in some form; therefore, we planned to recruit participants from this larger sample.

**Data Collection Methods**

**Informal Discussions and Observations**

We used an iterative process of engaging in informal discussions coupled with sense-checking discussions and observations during the implementation period. The informal and sense-checking discussions and observations were conducted with the team who was working directly in the wards, as well as with head nurses from the wards where the wearable device was being implemented. The informal discussions and observations were conducted during the implementation of the device.

**Feedback Survey Form**

We created the feedback form based on the components of the technology acceptance model (TAM) to assess the use and acceptance of the device. The TAM is used in a variety of disciplines to determine how individuals accept (or not) and use (or not) a given technology. Davis [17] developed this model based on components from the theory of reasoned action [18] and it consists of the following variables: use motivation (with perceived ease of use and perceived usefulness) and behavioral intention [17,19]. The model suggests that an individual will accept the use of a technology (ie, their behavioral intention) based on their perception of the technology’s usefulness and ease of use. Perceived usefulness refers to the perception that using the technology will enhance one’s work; for example, the wearable device will provide physicians and nurses some advantages (eg, remote monitoring). Perceived ease of use refers to the perception that the use does not add more work or effort to the work that could be enhanced; for example, using the wearable device will not increase nurses’ workload, despite its utility and simplicity [17]. The TAM framework was expanded twice to include attitudes as well as several other external factors [20]. The use of the TAM in health research has shown how perceived usefulness and perceived ease of use relates positively to attitude and behavioral intention [21]. The TAM has been criticized for being insensitive to the context or social factors, being simplistic, and following an assumption that users are rational decision makers, when indeed other factors play into decision making [22-24]. We used the TAM framework for its simplicity and because the categories of perceived usefulness and perceived ease of use were of relevance, but we also integrated other data collection methods alongside it to counter these limitations to some extent.

Based on the components of the TAM, we included 23 questions related to usefulness (n=5), ease of use (n=5), attitude (n=5), and behavioral intent (n=8) [25]. We asked these questions using a 5-point Likert scale (with scores of 5 being more favorable). We also added 2 open-ended questions and collected a variety of relevant demographic information (Multimedia Appendix 1). We piloted the tool in both English and Vietnamese and adjusted the form as needed. We used Google forms for electronic self-completion of the form and offered paper forms for hand-written self-completion. We explained the feedback form to the ward staff during team meetings and provided the link. The feedback form was distributed and completed in Vietnamese. We kept the feedback form link open for 7 weeks in total and started data collection after the implementation had been integrated into the wards so that users would have had experience using the device.

**User-Centered Workshop**

We held a user-centered workshop with a selection of HTD ward staff to explore user needs and ideas for development in more detail. Because we already had the technology and knew the spaces where implementation would be held, we followed an adapted version of the process described by Cooper et al [26]. With this approach, the workshop participants and facilitators set the scene as a busy COVID-19 ward during the peak of the pandemic. Then, the facilitators described the shells of users (personas), including a nurse and a doctor persona shell, and we had the workshop attendees describe who they imagined the nurse and doctor to be, as well as their behaviors and needs and the values each user group would find most essential. We based the conversation on the wearable technology that the participants had already used. Then, the group discussed solutions to the issues identified [26].

**Data Analysis**

Using the principles of action learning, we integrated the responses from informal discussions and observations into subsequent data collection, as well as summarized the content and grouped it into themes. For the analysis of the feedback survey form, we calculated mean scores for each variable and compared scores by profession. For the open-ended survey questions, we used content coding to summarize the responses topically. We presented the demographic data descriptively. We documented the responses from the user-centered design workshop as notes and summarized the results into main themes.

**Ethical Considerations**

In this paper, we are describing the processes that occurred as part of the development and implementation of a monitoring system; therefore, the work did not require ethics approval. Prior to the initiation of any activities, we held a meeting with ward heads to describe the work in more detail and obtain their agreement.

**Results**

**Device Implementation Within the HTD Context**

The wearable device was implemented in 3 wards starting in August 2021, including the adult ICU, Ward A, and Ward E.
We describe the implementation over a 5-month period from August to December 2021. During this period, these wards changed from COVID-19–designated and then back again to routine patient care settings, depending on the number of patients. Although the HTD was one of the COVID-19–designated hospitals, throughout the pandemic they offered routine patient care for specific diseases (eg, tetanus).

In addition to the rapidly changing physical spaces, the hospital management quickly deployed remote monitoring capacity using existing closed-circuit television cameras as a temporary solution to monitor very sick patients from outside the patients’ rooms. The remote monitoring was useful as it allowed for multitasking and prevented nurses and doctors from checking on patients more routinely in person. The hospital wards were at capacity during the study period. Prior to the pandemic, however, it was not unusual for the wards at the HTD to often be at maximum patient capacity. For example, in the adult ICU or during the rainy season, the number of dengue patients increases dramatically and the wards tend to be full.

Also, the workflow was organized differently during the pandemic period. Instead of nurses taking care of a few specific patients for the whole shift, 2 nurses and 1 doctor would instead go into the ward (in full personal protective equipment) as a team for 3 hours at a time while the other 2 nurses on shift completed admin work in the office. This meant that more coordination was needed, and often the team with the patients “need[ed] someone else to be [their] memory” as it was not easy to remember everything about all patients. The health care team’s workload, especially that of the nurses, ended up being more extensive for many reasons. One important reason is that, because of COVID-19 restrictions, there were also no families allowed in the wards who would help to look after patients in non–COVID-19 times; therefore, the majority of the care was left to the nurses. The patients were also more severely ill than previously in these wards and required more care by fewer staff.

**Device Use and Acceptability**

When we first distributed the feedback form, out of 90 potential participants, only 22 completed the survey (19 electronic and 3 paper forms), and 1 person stated that they did not use the technology and therefore no responses were recorded for that participant. Of the 21 respondents who completed the feedback form, 48% (n=10) were doctors and 48% (n=10) were nurses, with 52% (n=11) of the participants coming from Ward E (Table 1).

Overall, when assessing the TAM variables, the mean (SD) scores ranged from 3.6 (0.8) for perceived ease of use to 4.4 (0.6) for attitude, with behavioral intention (mean 3.9, SD 0.6) and perceived usefulness (mean 4.2, SD 0.7) scoring in between. Those working as nurses scored higher on perceived usefulness, perceived ease of use, attitude, and behavioral intention than did physicians (Table 2).

When asked, as an open-ended question, why participants would or would not use the wearable device in the future, of the 19 responses inputted, 15 participants wrote that they would use the system because of its convenience and usefulness in monitoring patients. However, in 2 of those responses, they also added comments that the device had limited perceived accuracy and transmission problems. Of the remaining 4 participants, 1 participant simply stated that the monitor was still in use, 2 participants wrote that they did not use the system anymore due to job location changes, and 1 participant wrote a few sentences about why the wearable device is not the “best choice,” highlighting its limited battery life, how the system had become additional work for the already overworked staff, and how it is not yet completely implemented.

**Table 1.** Demographic characteristics of the survey respondents (n=21).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>13 (62)</td>
</tr>
<tr>
<td>Men</td>
<td>8 (38)</td>
</tr>
<tr>
<td><strong>Age (years), median (IQR)</strong></td>
<td>35 (30-38)</td>
</tr>
<tr>
<td><strong>Occupation, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Doctor</td>
<td>10 (48)</td>
</tr>
<tr>
<td>Nurse</td>
<td>10 (48)</td>
</tr>
<tr>
<td>Other: nurses’ aid</td>
<td>1 (5)</td>
</tr>
<tr>
<td><strong>Primary ward during the implementation phase, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Adult intensive care unit</td>
<td>6 (29)</td>
</tr>
<tr>
<td>Ward A</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Ward D</td>
<td>3 (14)</td>
</tr>
<tr>
<td>Ward E</td>
<td>11 (52)</td>
</tr>
</tbody>
</table>
Integrating User Perceptions for Improved Implementation

As part of the action learning process, we supplemented the feedback form results with data from the observations and informal discussions during the 5-month period. There were 3 main observations. First, there was a mismatch between how we (ie, the research team) and the ward teams perceived the use of the technology. We quickly realized, from our observations and from informal discussions with the implementation team, that many of the nurses either did not use the wearable device or did not think that they used it even if they used it in some aspect (eg, connecting the device for the patients or changing batteries). Even after we clarified what we meant by “use,” there were still not additional participants who filled in the feedback form because they felt like they did not use the technology.

Second, the ward teams had varying perceptions of the technologies that are routinely implemented by the OUCRU team in the HTD wards as part of research projects. We heard from informal discussions with colleagues that the nurses assumed the wearable devices were from a research project, as is often the case with OUCRU projects, and therefore the nurses, in particular, ignored the device even if they had some role in its use. They did not see its potential benefit.

Finally, in order to make the device more useful for the ward staff, we realized during the meetings and informal discussions with the team that we needed to make the implementation and use of the device more “pro-nurse,” meaning we would need to emphasize how the device and its data were also useful and relevant to them. When discussing with the head nurse, the data were only displayed on the main screen in the staff room for one department. One suggestion was to move the tablet to the wall so that the nurses and others in the room (including the patients) could potentially see their vital signs. Because the devices and corresponding data were not in sight, it was easy to think that it was not relevant for the nurses and made it easier for them to ignore the device while with the patients.

User-Centered Design Workshop

With the information we had learned from the informal discussions, observations, and feedback form, we held a follow-up workshop on January 17, 2022, to discuss how we could make better use of the technology in the wards in COVID-19 situations in the future. The attendees included 2 doctors (1 man and 1 woman) and 3 nurses (2 women and 1 man). The participants discussed the behaviors and needs of the nurse and doctor persona. For both roles, the needs centered on having equipment and improved coordination. The nurses also mentioned more training needs, while the doctors’ needs were about the accuracy of monitoring (Textboxes 1 and 2).

There were 3 main value prop themes, including medical, technical, and patient themes. For medical aspects, the attendees discussed how the device should be able to provide highly accurate data, with appropriate alarms and cut-offs. For the technical theme, the device and software should be simple to connect and use, with a long battery life and stable connections during charging or switching devices. The display should be large and clear, and the data should be stored for a long period of time (ie, 7-10 days). Finally, for the patient theme, the device should be comfortable for the patients to wear to avoid them removing it.

There were several solutions discussed in the group to improve the use and efficacy of the wearable device (Table 3). Solutions included improving the credibility of the data, ideas to improve the ease of use, ways to make the alarms more consistent, and ideas for more ideal placement. One very specific issue that the group mentioned was that the alarms went off too much on the large display and the alarms were always red or black and blinking, and it was difficult to know if the device was turned off (due to patient discharge) or actually disconnected, which would require an intervention. The solution was to refresh the devices; however, if the alarms were excessive and not always indicating a real issue, trust in the device would remain low, so this was an important priority. They also suggested that the alarms and display on the tablet should be the same as the big screen, as they preferred screen consistency.

Another in-depth discussion was about moving the tablets to the walls and having the device plugged in all the time, which would solve the battery issues. They felt that the tablet could be set up on the wall but that brought up other issues about how to keep the device and watch safe after use. For some of the topics, the group used features of another wearable device that they had used in the wards in the past to inform their solutions (eg, device graphs and a line on the device for finger placement).

### Table 2. Mean technology acceptance model (TAM) scores by variable. The maximum score was 5.

<table>
<thead>
<tr>
<th>TAM variable</th>
<th>All participants, mean (SD)</th>
<th>Nurses, mean (SD)</th>
<th>Doctors, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived usefulness</td>
<td>4.2 (0.7)</td>
<td>4.3 (0.8)</td>
<td>4.1 (0.6)</td>
</tr>
<tr>
<td>Perceived ease of use</td>
<td>3.6 (0.8)</td>
<td>3.8 (0.7)</td>
<td>3.4 (0.6)</td>
</tr>
<tr>
<td>Attitude</td>
<td>4.4 (0.6)</td>
<td>4.6 (0.6)</td>
<td>4.2 (0.6)</td>
</tr>
<tr>
<td>Behavioral intention</td>
<td>3.9 (0.6)</td>
<td>4.0 (0.6)</td>
<td>3.7 (0.7)</td>
</tr>
</tbody>
</table>
Textbox 1. Behaviors and needs of the nurses.

Participant: Nurse Van is a 36-year-old woman. She is an administrative nurse and has a management job. She likes to have fun and has a family and 2 children. She is also responsible for bringing the kids to school and back.

Behaviors:
- Visit and provide direct patient care and monitor vital signs
- Carry out medical orders (ie, medications, blood tests, and nutrition)
- Assess, monitor, and hand over patients
- Work night duty
- Night shifts inform doctors on vital signs as prescribed

Needs:
- Equipment (eg, to measure blood pressure, temperature, oxygen levels, and heart rate)
- Training on diseases
- Teamwork and coordination

Textbox 2. Behaviors and needs of the doctors.

Participant: Doctor Huong is a 30-year-old woman. She is flexible and very active. She is not married and has no children and currently lives in a hotel. She is on night shift every 4 nights, and at times she visits her home in another town in Ho Chi Minh City, which is far from the Hospital for Tropical Diseases.

Behaviors:
- Prescribe medications
- Update medical records
- Perform examinations and change treatments
- Data entry
- Check vital signs in patient rooms (with a portable monitor that they move around) for examination and to detect abnormalities

Needs:
- Equipment (eg, monitors)
- Coordination with nurses (progress: medical records)
- Re-evaluation and working with other doctors
- Accuracy of vital sign monitoring

Table 3. Solutions for improvement.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Specific solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data credibility</td>
<td>Adding a graph for signal strength</td>
</tr>
<tr>
<td>Ease of use</td>
<td>- Adding a finger placement mark on the device</td>
</tr>
<tr>
<td></td>
<td>- Increasing the font size on the watch and tablet</td>
</tr>
<tr>
<td></td>
<td>- Tablets should be fixed on the wall</td>
</tr>
<tr>
<td>Alarms</td>
<td>- Reduce the alarm colors and blinking on the screen</td>
</tr>
<tr>
<td></td>
<td>- Use the same display on the screen and the tablets for consistency</td>
</tr>
<tr>
<td></td>
<td>- Refresh the tablets for more accurate alarms</td>
</tr>
<tr>
<td>Battery issues</td>
<td>Keep the tablet plugged in</td>
</tr>
<tr>
<td>Device placement</td>
<td>Placement on the wall (but only with an increase in font size)</td>
</tr>
</tbody>
</table>
Discussion

The HTD and OUCRU teams, along with the technology company, rolled out the wearable device in an extremely complex pandemic situation with a prototype system. In the end, the team used the device on over 100 patients. We assessed the usability and acceptance of the device over the implementation period when COVID-19 cases were peaking in the hospital and into the period when the COVID-19 cases were reducing. Similar to the literature on the topics, we found that the importance of understanding the users and their experiences using the device was crucial to get the most use out of these technologies.

There was a mismatch between our perception of who was using the device and those who thought that they were using or benefiting from the device on the ground. From the start, the device was designed and set up with doctors in mind, but in practice, the nurses’ roles and use were overlooked, even though they could also routinely use and benefit from the device. In our study, we found that the nurses who filled in the feedback form, on average, had slightly higher scores on 3 of the 4 TAM domains (ie, perceived usefulness, attitude, and behavioral intention), while the doctors, on average, scored the perceived ease of use slightly higher than the nurses. We know from the challenges with acquiring feedback that many nurses did not feel that they used the device even though they had some role in the device set-up and monitoring. Designing the device to be more nurse-centric from the early phases could have helped to increase the efficiency and definition of who is meant to use it. In the future, it is important to consider that the way the device is used might be dependent on the form of its use (eg, for triage, use in a pandemic emergency, or routine hospital use). We recommend the involvement of staff who could benefit from the technology, especially nurses in the hospital context, in the full implementation process. This could help to avoid mismatches in the perceptions of who the users are and who could and should benefit from the new technology. Research on integrating wearable devices during COVID-19 in Singapore also highlighted that device simplicity would encourage its use and the importance of making the technology fit into the current environment while not increasing or disrupting workflows [27].

The trust in the device and its data was an issue brought up several times during the implementation and feedback sessions. There are a variety of potential explanations for inconsistent data (eg, incorrect device placement or averaging of data); however, it reduced the credibility of the device for both doctors and, importantly, nurses. Data concerns about technology in clinic settings has been noted in other studies. For example, Faria et al [28] found that study clinicians reported that 36% of the data from a remote monitoring project were “invalid” for a variety of reasons, including low literacy of the patients and complexity of the device. Involvement from users from the beginning of the design and implementation process is crucial for design purposes but also to build trust and confidence in the devices [11]. While this project took place during COVID-19, which is a very specific circumstance, the broader findings resonate with research conducted prior to COVID-19 that focused on the implementation and scaling up of digital health technologies in LMICs. The recommendations also included integration of end-user feedback and engagement with all stakeholders throughout the design and implementation process [12].

There are limitations to this work. First, we did not collect data on the clinical worth or the accuracy of the data transmitted from the devices. Second, we focused on feedback from only health care staff (ie, doctors and nurses), and from only a subset of those who perceived that they used the device, which may have excluded some users and limited the overall sample size. We did not include patients who could also inform device acceptance, especially if used in noncritical cases where patients are moving around and conscious. Finally, the implementation setting for this work is not typical of other hospital settings in Vietnam or possibly other LMICs, as the HTD is a large referral hospital with an international research institute attached to it.

In anticipation of future (novel) pandemic situations or integration of wearable technologies into a range of clinical settings more broadly, it is important to fully understand if and how the wearable devices could be used more effectively by doctors, and importantly, nurses in the wards, for monitoring of deteriorating patients, especially in LMICs where resources are already stretched. Using an action learning approach during the implementation process highlights the importance of integrating user perspectives, ideas, and solutions into development and design.

Acknowledgments

We would like to acknowledge Nguy n Đ i Ph ú and Nguy n Th í Tin from the Hospital for Tropical Diseases and Nguy n n Đ í m Trình and Đ ng Ph ng Th o from the Oxford University Clinical Research Unit for their support and commitment in operating the remote monitoring system.

The members of the Vietnam ICU Translational Applications Laboratory (VITAL) group are as follows: Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam: An Phuoc Luu, ChanhQuang Ho, Duc Hong Du, Duc Minh Tran, Dung ThiPhuongNguyen, Gianh ThiNguyen, HaiBichHo, Hien Van Ho, HungManhTrinh, HuyQuangNguyen, KhanhNguyenQuocPhan, KhoaDinhVanLe, KienTrungDang, LamKhanhPhung, LieuThiPham, NgocThanhNguyen, NhatTranHuyPhung, PhuongTinhLe, QuyenThanHaNguyen, ThanhThiLeNguyen, ThyBuiXuanDoan, TrieuTrungHuynh, TrinhHuaKhangDong, VanMinhTuHoang, VanThaiThanhNinh, VuongLamNguyen, YenMinhLam, SayemAhmed, JosephDonovan, RonaldGeskus, EvelyneKestelyn, AngelaMcbride, GuyThwaites, LouiseThwaites, HugoTurner, JenniferIloVanNuil, and SophieYacoub; Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam: TamThaiCao, ThuyBichDuong, DuongThaiHai.
Ha, Nghia Dang Trung Ha, Chau Buu Le, Thu Ngoc Minh Le, Thao Thi Mai Le, Tai Thi Hue Luong, Phu Hoan Nguyen, Viet Quoc Nguyen, Nguyen Thanh Nguyen, Phong Thanh Nguyen, Anh Thi Kim Nguyen, Hao Van Nguyen, Duc Van Thanh Nguyen, Chau Van Vinh Nguyen, Oanh Kieu Nguyet Pham, Van Thi Hong Phan, Quy Tu Phan, Tho Vinh Phan, and Thao Thi Phuong Truong; University of Oxford, Oxford, United Kingdom: David Clifton, Mike English, Shadi Ghiasi, Heloise G reef, Jannis Hagenah, Ping Lu, Jacob McKnight, Chris Paton, and Tingting Zhu; Imperial College London, London, United Kingdom: Pantellis Georgiou, Bernard Hernandez Perez, Kerri Hill-Cawthorne, Alison Holmes, Stefan Karolcik, Damien Ming, Nicolas Moser, and Jesus Rodriguez Manzano; King’s College London, London, United Kingdom: Alberto Gomez, Hamideh Kerdegari, Marc Modat, and Reza Razavi; ETH Zurich, Zurich, Switzerland: Abhilash Guru Dutt, Walter Karlen, Michaela Verling, and Elias Wicki; The University of Melbourne, Melbourne, Australia: Linda Denesy and Thomas Rollinson.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Study tools in English and Vietnamese.

References


Abbreviations

HCW: health care worker
HTD: Hospital for Tropical Diseases
ICU: intensive care unit
LMIC: low- and middle-income country
OUCRU: Oxford University Clinical Research Unit
TAM: technology acceptance model
VITAL: Vietnam ICU Translation Application Laboratory

©An Phuoc Luu, Truong Thanh Nguyen, Van Thi Cam Cao, Trinh Hoang Diem Ha, Lien Thi Thu Chung, Trung Ngoc Truong, Tung Nguyen Le Nhu, Khoa Bach Dao, Hao Van Nguyen, Phan Nguyen Quoc Khanh, Khanh Thuy Thu Le, Luu Hoai Bao Tran, Phung Tran Huy Nhat, Duc Minh Tran, Yen Minh Lam, Catherine Louise Thwaites, Jacob McKnight, Nguyen Van Vinh Chau, Jennifer Ilo Van Nuil, Vietnam ICU Translational Applications Laboratory (VITAL). Originally published in JMIR Human Factors (https://humanfactors.jmir.org), 05.01.2024. This is an open-access article distributed under the terms of the Creative
Original Paper

Exploring a Gaming-Based Intervention for Unemployed Young Adults: Thematic Analysis

Ingjerd Jevnaker Straand¹, MSc; Asbjørn Følstad², PhD; Jone Ravndal Bjørnestad¹, PhD

¹Department of Social Work, University of Stavanger, Stavanger, Norway
²Department of Sustainable Communication Technologies, SINTEF Digital, Oslo, Norway

Corresponding Author:
Ingjerd Jevnaker Straand, MSc
Department of Social Work
University of Stavanger
Kjell Arholms hus
Stavanger, 4021
Norway
Phone: 47 93 22 22 89
Email: ingjerd.j.straand@uis.no

Abstract

Background: Promoting positive psychologies that promote resilience such as a growth mindset could be beneficial for young, unemployed adults, as many lack the self-esteem and self-efficacy to cope with job search adversity. These young people may be reached at scale through the web-based delivery of self-administered positive psychology interventions. However, past studies report unsatisfying user experiences and a lack of user engagement. A gaming-based experience could be an approach to overcoming these challenges.

Objective: Our research objective was to explore how young, unemployed adults experience a positive psychology intervention designed as a game to extract learning and principles for future intervention research and development.

Methods: To respond to the research question, a team of researchers at the University of Stavanger worked with designers and developers to conceptualize and build a gaming-based intervention. Feedback from the users was collected through formative usability testing with 18 young adults in the target group. Retrospectively, recordings and notes were transcribed and subjected to thematic analysis to extract learnings for the purposes of this paper.

Results: A total of 3 themes were identified that pinpoint what we consider to be key priorities for future gaming interventions for unemployed young adults: adaptation to user preferences (eg, need for responding to user preferences), empathic player interaction (eg, need for responsiveness to user inputs and a diverse set of interaction modes), and sensemaking of experience and context (eg, need for explicit presentation of game objectives and need for management of user expectations related to genre).

Conclusions: Feedback from end users in usability-testing sessions was vital to understanding user preferences and needs, as well as to inform ongoing intervention design and development. Our study also shows that game design could make interventions more entertaining and engaging but may distort the intervention if the game narrative is not properly aligned with the intervention intent and objectives. By contrast, a lack of adaptation to user needs may cause a less motivating user experience. Thus, we propose a structured approach to promote alignment between user preferences and needs, intervention objectives, and gameplay.

(JMIR Hum Factors 2024;11:e44423) doi:10.2196/44423

KEYWORDS
positive psychology intervention; digital mental health; serious gaming; intervention design; research through design; gaming-based intervention

Introduction

Positive Psychology for Unemployed Young Adults

Young people who are not in education, employment, or training (NEET) comprise, on average, 12.8% aged between 15 and 29 years in the Organisation for Economic Cooperation and Development countries [1]. Studies show that negative self-perceptions and a lack of perseverance are barriers to successful labor market inclusion [2-4], as the new labor market requires highly skilled workers who are not afraid of change,
challenges, and acquiring new skills [5,6]. For young people with weak beliefs in their capacity to learn, this could be a major risk factor for labor market exclusion, and this may in turn impact their overall well-being. Several researchers have studied the relationship between unemployment and mental health. McGee and Thompson [7] found a relationship between unemployment and depression in young adults and suggested the use of psychological interventions for the young and unemployed. The Norwegian NEET group is more likely to be recipients of health-related benefits, have poorer mental health, and lower levels of education compared with the average of the Organisation for Economic Cooperation and Development [8,9].

A qualitative inquiry into young people’s own experience of unemployment in Norway points to poor self-efficacy and lack of self-esteem that are reinforced through challenges and setbacks, even when these initially occur beyond the individual’s control [10,11], such as when there are insufficient training placements on offer for the vocational school pupils, a problem leading to a relatively large number of unqualified school dropouts in Norway [9]. Thus, there is a substantial rationale for exploring further how the public can offer training, not only in job-seeking skills, such as curriculum vitae (CV) writing and gaining work skills, but also in building psychological well-being and resilience to cope with such setbacks and challenges [12,13].

In the context of a broader research project, the Career Learning App, our study investigates the design and development of a web-based intervention using positive psychology to achieve beneficial changes [14]. Our broader research idea is that young people in the NEET group, henceforth referred to as young, unemployed adults (for the sake of simplicity and to reduce stigma), could benefit from building confidence in the possibility of learning and improving. The research idea stems from a body of work that has demonstrated positive results from offering high school students self-administered positive psychology interventions (PPIs) centered on growth mindset and challenge-seeking behaviors [15,16]. A growth mindset is the belief that human capacities are not fixed but can be developed and increased in response to one’s own efforts, good strategies, and help from others [17]. If a simple web-based PPI can influence high schoolers’ mindsets in ways that lead to positive academic outcomes [15,16,18-20], then it could also likely be beneficial to the young unemployed, leading to changes in how they engage with their contexts. Despite this strong rationale for the applicability of PPIs to facilitate well-being and personal growth in vulnerable populations, they have only, to a limited extent, been tested and used in the context of unemployment [13,21]. However, we cannot simply apply the PPIs designed for educational contexts; they need substantial adaptation to be relevant or usable for this new target group of young, unemployed adults [17]. For instance, the school-related examples used within the PPI to make them relatable are not relevant to this new target population. Furthermore, there is a lack of shared context to piggyback on to deliver the intervention and ensure that users will adhere to it. Thus, there is a need to design and develop a web-based PPI designed specifically for young, unemployed adults and their context. If successful in user studies, a resulting intervention may be used in forthcoming large-scale randomized controlled trials in Norway.

**Problems With Self-Administered Interventions**

Self-administered web-based interventions have the potential to support well-being and positive health changes in a large number of people at a moderate cost [22]. However, this introduces new challenges, illustrated in Figure 1. First, there is the challenge of adapting current PPIs to self-administered digital formats that are fit for the purpose of the intended user population. Second, there is the challenge of user motivation to obtain the users to complete and adhere to the intervention [23,24]. Past research suggests that users are not interested in or do not enjoy using digital mental health interventions [25], suggesting a need to work on the actual interventions themselves to increase engagement and user motivation.

**Exploring PPI as Gameplay to Be Relevant for Young People**

Past research suggests a need to adapt to the media preferences of young people and make the apps more visual and interactive to increase engagement and motivation among young people [26]. One possible approach to increasing engagement is to explore games and game elements. Starting from “where the young people are at” makes pedagogical sense [27], thus the application of game design is founded on young people’s own interests as a way to foster engagement and learning of positive psychologies. Although play and games are not unique to young humans [28,29], the average age of video game players is now 33 [30]. However, playing video games continues to be popular among young people [30,31]. Interactive digital games are increasingly used for purposes beyond entertainment, as
exemplified by the rise of health gaming apps for video gaming consoles. Game design elements are also increasingly applied to nongame contexts, for instance, by adding points and badges to nongame experiences, such as social media networks [32,33] or learning contexts [34]. When game design and game concepts are being applied for purposes beyond fun, they may be termed “serious games,” “learning games,” or “gamification” [35-40].

Game design applied to learning may be seen as a form of experiential learning (e.g., learning-by-doing) [41,42]. Game design has been successfully applied to mental health interventions [43,44] and educational contexts [34,38] in the past. Game design offers an approach to creating engaging experiences. Engagement is a complex and ambiguous term [45]. Our use of the term is in the sense of “emotional involvement” as in offering a pleasurable experience [32] and to describe how motivational, usable, and acceptable [46] the game would be in the eyes of the target audience, because this could be an important predictor of adherence. In general, there is insufficient research on the application of gaming and gamification to mental health, particularly in the well-being domain [47], and we have not found empirical studies that pursue to gamify positive psychology targeted specifically toward unemployed young people. There were no available gaming-based PPIs that could be used for the purposes of this study.

**Research Objective**

Our limited knowledge of how to adapt the intervention from an educational setting to a game-based format suggested a highly explorative approach, where we identified the need to design a game to explore this topic and to overcome the gaps in knowledge summarized in Figure 2. Thus, the research objective and question of this study were how to design a self-administered and digital PPI in a gaming format targeting young, unemployed adults and to explore how they engage with the game and whether they like using it.

**Methods**

**Setting: Learning in “Action”**

To answer our research question, this study used a human-centered design process [48,49], an approach that allows input from target users during design and development. The human-centered design process is a form of research with design as the primary outcome [50]. The collection of feedback from users during design and development impacts not only the design of the game but also the production of knowledge. As such, it is a form of participatory action-research [51,52] where “doing research” and “doing action” happen simultaneously. In the design research literature, this may also be referred to as “Research through Design” [53,54], where knowledge is produced through the design of the artifact and through the experience of the artifact. At the end of the project, we analyzed user feedback data thematically [55] to extract the learnings and design principles for the potential application toward creating user-friendly and user-relevant gaming-based PPIs for vulnerable populations. Figure 3 summarizes the “Research through Design” approach of this study.
Data Collection Through User Testing

We used testing with users to capture the user experience. Testing with users is important as a tool for obtaining the “design right” in game design [56]. Our project-specific usability study [57] was formative, and we were qualitatively evaluating early prototypes to answer how and why questions as a means to improve the design [58,59]. We aimed to capture participants’ thinking, for example, opinions, reasoning, and attitudes toward the prototype experience. The various prototypes thus functioned as exploratory “hypothesis testing” [60,61] and as boundary objects [62-64] that framed conversations with end users. These sessions with the end users lasted from 45 to 90 minutes, where users were asked to briefly describe their background and interests, followed by open “think-aloud” questions [65] related to a prototype experience, such as “What do you think about what you see here?” “What do you expect will happen now?” and “What do you think this is?” Thus, the participants were encouraged to verbalize their thoughts and experiences. Afterward, the participants were asked follow-up questions that were equally open, such as “What are you thinking now that you have seen this?” and “How would you describe this to a friend?” The objectives of usability testing were to collect feedback related to broad aspects of the intervention experience, namely its (1) engagement, (2) relatability, (3) understandability, and (4) potential for improvement.

Sample and Recruitment

The NEET group includes anyone not in economic activity from the age of 15 to 29 years [1]. The European Foundation for the Improvement of Living and Working Conditions has defined 7 subgroups of NEETs [66], ranging from the “classically unemployed” to young people who are caretakers, unable to work, or simply listed as “inactive.” How long an individual remains a NEET also varies significantly. The diversity of the population is not necessarily problematic for our study; when designing with users, one would usually strive for variance [49,67] rather than representativeness. Our study is exploratory and does not require large samples [68]. Our estimate required 20 users; however, this sample size was highly approximate, in line with qualitative studies in general [69,70]. Pragmatic needs in the design process guided the number of participants to a large degree and not, for instance, theoretical saturation.

We recruited from the Norwegian Labour and Welfare Administration (NAV), the national welfare institution of Norway that pays out unemployment wages and social support, and from a regional Individual Placement and Support (IPS) program, which offers placement support to young people with first-episode psychosis. We also recruited through the user testing platform, Teston (UserTesting). Our inclusion criteria were an age range of 18 to 29 years, with a “NEET background,” and because of the language in the game prototype, living in Norway and speaking Norwegian. We say “NEET background” and not “NEET status” because our participants from the IPS program were no longer in the NEET group by definition. We made a deliberate choice not to exclude based on the length of NEET status and unemployment. Although those who are entering the NEET group in the short term, the “in-betweener” [66], often find new employment without assistance [71] relatively quickly, even a short time out-of-work may increase the risk of exclusion [72]. We did not recruit participants on permanent disability allowance. Using multiple channels enabled quicker recruitment during the COVID-19 pandemic and increased the variance in our sample (as desired), as young people without the rights to receive unemployment benefits have fewer incentives to register with NAV [73]. We did not compare the experience of the intervention based on the recruitment channel because the groups were overlapping and experienced varying prototypes depending on the stages in the design process.

Participants

In total, 18 participants (12/18, 67% females, 6/18, 33% males) took part in the study during the 21 testing sessions; thus, some participants were involved more than once. Recruitment was particularly challenging because of the COVID-19 pandemic, and we found that it was difficult to recruit young, unemployed men. Remote participation through web-based technologies, such as the Zoom (Zoom Video Communications) platform, enabled the study to continue during the lockdown. We also experienced that this user group continued to prefer remote participation, even when restrictions were lifted. Table 1 summarizes the participant statistics and format of the usability test.
Table 1. Age distribution of participants and format of user testing.

<table>
<thead>
<tr>
<th>Recruitment channel</th>
<th>Format user testing</th>
<th>Age (years)(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPS(^b)</td>
<td>In person+Remote</td>
<td>19</td>
</tr>
<tr>
<td>IPS</td>
<td>In person+Remote</td>
<td>28</td>
</tr>
<tr>
<td>IPS</td>
<td>In person+Remote</td>
<td>18</td>
</tr>
<tr>
<td>IPS</td>
<td>In person+Remote</td>
<td>27</td>
</tr>
<tr>
<td>IPS</td>
<td>In person</td>
<td>22</td>
</tr>
<tr>
<td>IPS</td>
<td>Remote</td>
<td>19</td>
</tr>
<tr>
<td>IPS</td>
<td>Remote</td>
<td>18</td>
</tr>
<tr>
<td>IPS</td>
<td>Remote</td>
<td>23</td>
</tr>
<tr>
<td>Teston</td>
<td>Remote</td>
<td>18-29</td>
</tr>
<tr>
<td>Teston</td>
<td>Remote</td>
<td>18-29</td>
</tr>
<tr>
<td>Teston</td>
<td>Remote</td>
<td>18-23</td>
</tr>
<tr>
<td>NAV(^c)</td>
<td>Remote</td>
<td>21</td>
</tr>
<tr>
<td>NAV</td>
<td>Remote</td>
<td>18</td>
</tr>
<tr>
<td>NAV</td>
<td>In person</td>
<td>21</td>
</tr>
<tr>
<td>NAV</td>
<td>Remote</td>
<td>21</td>
</tr>
<tr>
<td>NAV</td>
<td>Remote</td>
<td>18</td>
</tr>
<tr>
<td>NAV</td>
<td>Remote</td>
<td>18</td>
</tr>
<tr>
<td>NAV</td>
<td>In person</td>
<td>22</td>
</tr>
</tbody>
</table>

\(^a\)For 3 (17\%) of the 18 participants, we only had an age interval provided to us.
\(^b\)IPS: Individual Placement and Support.
\(^c\)NAV: Norwegian Labour and Welfare Administration.

The Design and Development Process

Researchers at the University of Stavanger worked with designers and developers from a consulting company and potential future end users in an agile [60] and human-centered design [48,49] process. The objective of the process was to create an enjoyable gaming-based PPI targeting unemployed young adults. This process took place in 4 steps over approximately 9 months, from 2019 to 2020. In the first step, (1) design exploration in the form of a design sprint, a 5-day design and prototyping process [61], produced a minimum viable product of a gaming concept that was tested with 5 participants. Following a brief period for planning and procurement, we moved on to (2) agile development, consisting of 3 sprints, each lasting about a month. During this step, usability tests were conducted on 6 participants. This was followed by (3) refinement of content and prototypes, with involvement from behavioral intervention researchers to further develop and “add-in” the necessary intervention content. Finally, the prototypes were evaluated using (4) testing. In this step, 10 participants participated in usability testing and provided feedback on the final set of prototypes. Figure 4 summarizes the stepwise design and research process.

The intervention content needed to be adapted to be meaningful and relatable to the user group [74]. The basis for our gaming-based PPIs were growth mindset interventions from the “National Study Learning Mindset” [16] and its translated Norwegian version, “U-SAY” [5,15]. These are interventions that target high school students [15,16]. We added selected parts of cognitive behavioral therapy [75], specifically management of negative emotions, panic, and anxiety, to offer a more productive interpretation of stressors [76] that may occur during job search adversity [77,78]. During the first few days of the initial design sprint, a gaming concept, VitaNova, was developed where players can build a “new life” in a fictional narrative.
The Gaming-Based Intervention

The VitaNova gaming concept is a one-player fantasy game where the user plays a “no name,” an android character that can take on any skin to morph into another character with their skills and abilities. In particular, the player can choose between being Noomi and Twizzlesprock. However, as we learn through the game, your character has a backstory as the male character Abel, a former engineer and an outcast who sees himself as a failure. The game is designed in 3D and split into episodes (missions). The game starts with very little information and instructions, landing the user right into action. This was a design tactic to spark curiosity and make the users intrigued by the game so that they would want to explore it further. As a player, your first task is to escape from captivity, and then, gradually, more and more information is being revealed to you as the game progresses. Refer to Textbox 1 for an overview of the game narrative.

The visual design of games can influence how motivational and acceptable they are to the target population [46]. Therefore, to make the game look polished, cool, and professional and to keep users immersed and engaged [79], emphasis was placed on 3D design and detailing. The game was divided into episodes, or missions. The first 3 episodes were developed into a nearly fully functional game in Unity WebGL, a platform for building 3D games that can be used in a web browser. During user testing, we also showed the prototypes and the wireframes that were made in Figma. Figure 5 shows the prototype iterations of the game design.

Psychological content and tasks are entered into the gameplay to foster psychological well-being, teach a growth mindset, and offer psychoeducation and mental health tips. Some of this is interwoven into action in the form of interactive quizzes, dialogues, or other forms of interaction, such as a CV builder applied within the game. This was intended to be transferable to the end user situation to increase relevance, although the acquisition of such practical skills was not a target of the intervention. Furthermore, there was also psychological content that was external to the gameplay, such as embedded videos. When using externally sourced content, this was implemented in the game as “ruins from the past,” which the player could “find” in the game. The player would need to watch this content and use the information provided to complete the challenges and the in-game quizzes. Upon completing an episode of the game, the player was requested to write an answer to a reflective question where the user should answer as himself or herself, to encourage internalization of the messages that had been taught in the intervention through self-persuasion [80], and to transfer learning to the user’s own situation. We have included a further description of the prototypes in Multimedia Appendix 1.
**Textbox 1.** Description of the game narrative. (The 2 final missions were not included in the user testing.)

1. **Introduction**
   Wake up in the trunk of a moving vehicle. Use hacking skills to hack the lock. Find a tavern and interact with a bodyguard who refuses your entry as "no name." Find the Noomi skin and power unit and turn them on to enter the tavern.

2. **Tavern**
   Interact with Griff and Mia at the tavern to learn about this world. Try to receive help. Tell them that you must go to an old public office to pick up energy bars if you are going to receive any help.

3. **Learn**
   Find an abandoned public office where, among other things, you will discover many pieces of ancient psychological knowledge as ruins from the past.

4. **The sidekick**
   Use newly acquired knowledge to help out the depressed and anxiety-ridden droid Griff, who now becomes your sidekick, and help Mia the bartender.

5. **The bully**
   A bounty hunter is out looking for you to receive a reward from the boss at BetterJu Janus. Threatens your new friends at the tavern. Why are they after you?

6. **The chance**
   You discover that the only way to obtain all the answers is to try to join into BetterJu. Your friends tell you of a job posting that is open. Interactive curriculum vitae and job application process.

7. **The job interview**
   The job requires a different set of skills, which you acquire through entering the skin of Twizzlesprock. Job interview at the company BetterYou as Twizzlesprock.

8. **The escape**
   Discover who you really are from the overhearing conversation between your new boss, Janus, and a droid. Find your ex-girlfriend, who has been trapped but confirms your true identity. You both escape from the evil boss. The end.

**Figure 5.** Prototype iterations of VitaNova.

---

**Analysis**

The purpose of the usability testing was to synthesize findings that led to improvements and changes in the designed outcome. Retrospectively, we also conducted thematic analysis with the steps from Braun and Clark [55,81] as a practical guide: (1) data familiarization, (2) initial code generation, (3) search for themes, (4) review themes, (5) define and name themes, and (6) produce reports [81]. All 21 usability-testing sessions were recorded. The 7 most comprehensive usability tests were transcribed verbatim. The remaining data were analyzed based on recordings, researcher notes, and memos. Specifically, we used the transcribed data as our starting point and went back to recordings and memos to review codes and themes. The analytic process was iterative and creative, where we often moved back and forth between the data and codes [82]. All authors independently familiarized themselves with the data. Author 1 started with coding using the qualitative analysis software ATLAS.ti (ATLAS.ti Scientific Software). As a group, we discussed the findings and the initial codes in a workshop before moving over to paper-based coding and printing quotes from participants organized on large paper sheets. We used
diagramming, both digitally in Miro and in pen-and-paper sketches, to iterate themes and review codes in consensus meetings. Findings and concepts were discussed with other researchers, some of whom had acted as observers for the user testing or had watched recorded sessions. The quality of the analysis was ensured by researcher reflexivity, end user involvement, and method triangulation. Researcher reflexivity concerns activities that consider how researchers might have informed the research or biased outcomes [69]. Reflexivity was enabled through critical discussion of assumptions, themes, and codes in the team of researchers. The team also involved researchers not involved in the user testing or in the design process as an approach to validate the analysis based on these methods. End user involvement was supported through the iterative design process, where participant perspectives were sought at different levels of concept and design maturity. Specifically, we found it valuable to involve users from different recruitment sources (IPS, Teston, and NAV) to strengthen the credibility and transferability of the findings. Method triangulation was conducted by applying different approaches to design and user involvement at different phases of the process, allowing the assessment of themes or constructs from different perspectives. In particular, including data from the different phases of exploration, design, and evaluation was found to strengthen the credibility of the findings.

**Ethical Considerations**

All participants provided explicit and written consent to participate in the study and were rewarded with gift cards (approximately US $30/session) for their participation. The study was evaluated and approved by the Norwegian Centre for Research Data (approval number 131074) and the regional committees for medical and health research ethics in Norway (approval number 42128).

**Results**

**Overview**

As a tool to motivate and help young adults engage in work or education, the idea of an interactive and digital game was regarded by all participants as a “good idea”; it was described as “cool,” “unexpected,” and “motivating” upon first impression. Upon closer experience with the different gaming prototypes, we received different and more specific feedback. A total of 3 themes were constructed by the researchers through active analytic engagement with the data [55]. The themes pinpoint what we consider key priorities for future gaming interventions for unemployed young adults: (1) adapting to user preferences, (2) empathic player interaction, and (3) sensemaking of experience and context. Refer to Figure 6 for an overview of the themes and their key characteristics.

![Figure 6. Themes and theme characteristics.](https://humanfactors.jmir.org/2024/1/e44423)
Theme 1: Adapting to User Preferences

Sentiments about the game are based on enjoyment of the game experience, the genre, and whether the game meets user preferences, either by being targeted toward them or by allowing for experience customization. The participants expressed positive feelings and excitement related to the game. It was described as much more of an actual game than was expected:

I liked it; it was unexpected. [P3]

Seems like a cool concept. Never heard of it before. [P17]

The participants pointed to the use of humor and compared the game to pure entertainment media, such as commercially available games, or other parts of popular culture, such as films or television series:

It was actually pretty exciting [laughs]. Kind of funny, considering that you have included the welfare administration here. [P6]

The story was funny, it seemed a bit like a video game. [P10]

The game concept was further described as “something to do” or “something to allow the time to fly,” as pure entertainment or for relaxation purposes. The relaxing features were described as something that could make one more receptive:

If it’s like that you get new assignments once a week, then it can seem exciting. At least that is something to do. [P7]

So you are pretty relaxed when you get these questions, so then it is probably a little easier to answer... a little easier to reflect over this. [P4]

Others felt that the gameplay story could motivate them to do something by creating a sense of urgency, where the story would drive them to make more effort. A participant wanted proof within the game that it would be worth the effort to perform mundane tasks, such as updating their CV:

That’s how it is in life too: You have to make an effort yourself to get ahead; when you have to do something here [...], then you have to hurry, because someone is after you. Then you get the adrenaline to do it. [P16]

If there’s something in the game that can prove to me that it’s worth it, like writing a resume is worth it.... [P18]

However, several participants expressed uncertainty about whether the game would be “something for them,” suggesting either other preferences or not quite the right conceptual fit. In particular, the concept likely needed “aging up,” as it was perceived as something for a younger population. Furthermore, we learned from the participants, who spent more time gaming, that they had started out playing “adventure games” and “role-playing games” when they were younger, but that they had since moved on to playing “first-person shooter” games or other kinds of games:

It’s not a game I would have bought in the store. [P15]

I think if I had been younger then, yes. Because I have a lot of different types of games that I like to play, and now I like to play games where you shoot people, but earlier I liked playing games like that, where you follow a story, for example, it’s very different in... it’s very different for people what kind of games they like to play. [P2]

Designing for mobile phone use was considered important and was brought up as an improvement suggestion by nearly all users. Participants expressed that mobile phone use would make it easier to meet their own user preferences or the user preferences that they expected other young people to have:

I would have chosen an app or a mobile game. Or a course, if it was on [a] mobile [phone]. [P7]

Not everyone has a PC with them everywhere, so I wondered if this was on mobile. [P17]

If the game could be turned into a mobile game, that would be better. [P1]

Adapting to user preferences could also mean designing a customizable or more personalized experience. In particular, the choice of characters in the game is usually an arena for customization and personalization. One user commented that she would prefer to customize a character by selecting hair length, and body shape, etc rather than choosing between predefined female and male characters:

if you can choose male or female or... you don’t have to have these two, but you can choose what that person looks like. Since now a lot of young people... there are some young people who don’t want to be a man or a woman, so I think you... it’s very smart to make something like ‘do you want long hair’ ‘do you want short hair’ ‘do you want...' [laughs] [P2]

Others mentioned different strategies for choosing characters by either choosing a character that resembled themselves or identifying with them. Others would deliberately choose the extreme opposite of themselves. As one participant mentioned, “if your choice was the big male character, then you were likely ‘more vulnerable on the inside’ and ‘in need of protection.’”

Theme 2: Empathic Interaction With the Player

This theme describes the need for responsiveness to user inputs and the desire for a range of interaction modes. Although initially intrigued and enthused by the game design, users quickly became disappointed by the lack of functionality. Thus, this theme is based on the need for empathic interaction with the player in the game, where the game needs to take the user seriously by being responsive to user inputs, thereby allowing for actual contribution to the experience:

It’s very... you can see very easily that your answers don’t make much of a difference. It doesn’t matter what you choose. And I think if you’re going to have a game like this, you have to have a little change in what you say, how will it affect the game. [P2]

No, it’s just that I want to see that the people you’re talking to have something else or something more to

https://humanfactors.jmir.org/2024/1/e44423
Participants find the graphics visually appealing, although they do not feel that this is the most important aspect, saying that how the game works and how exciting and entertaining the game is are the most important parts:

*I think it looks pretty nice. There are many different types of games that can be in like... many different types of ‘art’ [styles] and so, yes, there are many games that can look one way, but can still be really good, but some games that look really nice, can be boring. So, don’t worry about how the game looks, but how the game works, that’s very important.* [P2]

Visually it was very nice... The story I am more unsure of...* [P12]

Some of the feedback indicates that our game was perhaps not fully developed as a game with the necessary combination of rules, goals, feedback, fantasy, and fun [32]. Participants wish for a wider range of interaction modes, such as moving more freely around in the game, having more challenges and tasks in the game, and having different ways of interacting with the characters. Users expect game-like interactions, not just choosing answers, reading, and writing:

*It is unusual for me that you cannot move around [in the game].* [P4]

*But I think that it would be a bit boring if it was just like that you had to read, and then click to choose answer options. [...] if I was sitting at home, and this was something I had to go through every week[...] then I’d just click quickly through it. And then I hadn’t properly read what it said.* [P7]

To support the learning objectives of the game, players had to answer reflective questions at the end of each gaming session, where they answered as themselves and not as the gaming character. This felt a bit “off” to the participants. Furthermore, several participants expressed a problem with articulating answers to those kinds of questions, expressing that they would not know what to write when asked:

*I didn’t expect that an assignment came up where you have to write about an experience from reality, sort of, which seems a little unusual to me since you are sitting inside this alternate world. But I had probably only written something about skateboarding. But it was very unusual for it to be like that.* [P4]

*I don’t know... at least I struggle a lot with tasks like that[...] I probably wouldn’t have written anything here.* [P5]

Although not as exciting as hoped for, the challenges and tasks in the game can still provide the user with a sense of achievement and act as an awakening for new thoughts; if not for them, then perhaps for someone else:

*You kind of get a little more confidence in yourself then. That you have actually managed something.* [P6]

...after all, it raises thoughts and yes... new ways of looking at things, I think, that it can start something in someone.* [P7]

**Theme 3: Sensemaking of Experience and Context**

This theme comprises a desire for an explicit presentation of the game objectives and a need to manage user expectations related to the gaming genre. The game was described as “cool but confusing:”

*Uh, well, it seems kind of cool, but it was a little hard to understand, I felt.* [P5]

Many participants mentioned sensemaking or lack of understanding in some form or another; they struggled to understand the point, the objective, or the mission to be completed in the game. Some participants pointed to a lack of logic or strangeness in the storyline and over-the-top reactions to what they perceived as minor happenings:

*So people want to buy parts of dead people so they can look how they want? Hum. That’s a very strange concept! [laughs].* [P2]

*She is stabbed now! ... And the taverna is burning. That was over-the-top. She just came for some food. This is over-the-top.* [P12]

The intervention messages in the game were not perceived by any of the participants. They were uncertain about what they were learning from the game experience. They were focusing on the details of the game narrative, trying to make sense of that, and, thus, the intervention part seemed “part of the fiction” and not clear what this was meant for:

*Didn’t learn anything. Well... I learned that there can be different ways to solve things, but I didn’t really learn anything[...] It was a bit difficult to understand the whole story, that is the whole thing.* [P1]

*Lots of talk about the brain, that the brain is a muscle, but don’t know what it can help with, it doesn’t make sense.* [P2]

The gameplay added complexity and was confusing to the participants. For most participants, there was a desire for an explicit presentation of game objectives, both in terms of what it should ultimately achieve for the end user (as an intervention) and what the objectives in the game narrative are. However, other participants felt that this uncertainty was part of the excitement:

...I should have known a little more what the goal was and what the meaning behind the game was. Because it seemed a bit like that, yes... a bit out of the blue. And you didn’t quite know what an anonymous person was and whether this was the future or whether this was a completely different world.* [P7]

*It was very interesting. It was very unusual for me with that kind of game. But I liked how it was. And you didn’t have very much information about what...* [P6]
One user expressed explicit concern about how relatable the contents could be if you use a context that is far removed from everyday life:

If it becomes too sci-fi, I think it might be difficult to transfer to reality. [P18]

Furthermore, we also identified a need to manage gaming genre expectations. The participants expressed a preference for certain genres over others; it may be difficult to cater to different preferences in terms of what games they like best to play. There are also certain expectations connected with different gaming genres that we were not so aware of in the research team, where users were trying to make sense of the game prototypes in relation to established genres, with expectations of gaming interaction to be similar to games in that genre. The participants asked us about the game in relation to genre concepts such as “open world,” “adventure,” and “role-playing” games:

So, I have a question, is this an open-world type of game where you go out to different places and pick up things or is this a text where you just follow what happens in the story? [P2]

I have a question: Is this open world—or just to follow a track, like? [P17]

It seems that the game genre was not clear to the participants, who pointed to different features of the gameplay prototypes that would take the game in different genre directions.

Discussion

Principal Findings

This study has used an iterative design process with active participation from potential users to develop an interactive game that aims to be user-friendly and engaging to be able to provide a vulnerable population with positive psychologies. As pointed out by past research, there is a strong rationale for promoting psychological well-being, for instance, to improve resilience [83] in the face of setbacks and challenges that occur as part of job search and being “out-of-work” [12,13,78], and may thus alleviate suffering [84]. A total of 3 themes were constructed from the user-based research that occurred through the design process of the gaming-based intervention: (1) adapting to user preferences, (2) empathic player interaction, and (3) sensemaking of experience and context. In the following section, we discuss the themes and how they could potentially be applied as designing principles for future self-administered gamified PPIs. Thus, the study sheds light on the application of game design for PPIs that aim to promote well-being and increase challenge-seeking in young, unemployed adults.

Comparisons With Previous Work

Adapting to User Preferences

This study expands the knowledge found in other studies on PPIs for young people, where the need to offer interactive, visual, and more engaging experiences has been identified [26,85,86]. Most participants expressed uncertainty about whether the game would be “something for them,” suggesting other preferences or not quite the right conceptual fit with their preferences. We interpret this to mean that there is a need to consider gaming genres and user preferences specifically, where a more refined user segmentation may be necessary [87,88]. For instance, in our group, some participants said that they only had an interest in certain kinds of games. Future studies could consider a more fine-grained targeting strategy based on preferences and interests and not simply age and employment status. A possibility would be to segment the population based on player types [89,90] or motivation [89,91-93], and to think more carefully about user preferences for different gaming genres before choosing a concept. In our study, we found that the selected game genre was perceived as engaging “not to them,” but “someone younger.” This was particularly true for the active gamers, who found the genre to be immature. Furthermore, most participants stated that they would have preferred a game designed for mobile use, indicating another kind of context of use than our initially planned use on PCs at home.

Empathic Player Interaction

Inside the game experience itself, the participants in our study were disappointed by the lack of features, functionalities, and opportunities to influence what was going to happen in the game, for example, player autonomy. A lack of autonomy may cause a more negative interpretation of an experience [94]. We interpret this as underdelivery, partly because of the overpromise of the first impression and the aesthetics of the graphics [95,96]. The participants expressed bleakly that in-game actions “do not matter” because, as players, they experienced an insufficient influence on the string of events in the game. Autonomy is an important motivator in self-determination theory (SDT) [97] and a lack thereof may contribute to reduced user motivation [13,35,97]. In part, the lack of autonomy and interactivity was caused by the requirements for a structured intervention set by the broader research project; each player needed to experience the same sequence of events. However, even within this frame, the game should be built to cater for somewhat more variation and focus on the interaction between the game and the player to meet user expectations. There is also a more specific need to be aware that within this target group of young, unemployed adults, many may feel in general that what they do does not matter [98]. Gameplay with insufficient df may unintentionally reinforce that message.

Sensemaking of Experience and Context

Participants had trouble making sense of the experience: our prototypes did not meet the participants’ expectations for genre, not fitting with role-playing games based on narratives and dialogues or open-world type games, where you move around freely in a 3D world to pick up items and battle with other characters. Furthermore, the game objective and rules were either not clearly presented to the participants or did not cater to sufficient player-game dialogue and manipulation of the experience. We found that there were tensions between the gameplay and the messages of the intervention, which could undermine the intervention and potentially threaten its effectiveness. This finding is depicted in Figure 7. The PPI
gameplay had a complex storyline, which confused the participants and made them miss out on their learning objectives. Past studies have pointed to psychological affordances and the importance of a “fertile soil” to make positive psychological interventions more likely to work [99-101]. In instructional or serious games used in education, Young et al [38] concluded the need to ensure that game objectives and learning objectives correspond and, further, that an overly complex gameplay can lead to misunderstandings and interfere with understanding. This seems transferable to gaming-based interventions. Other authors have referred to this as “relevant narrative,” which states that the narrative of the game should be relevant to the subject matter [102]. The choice of gameplay as a strategy for creating engagement for an intervention introduces a new context, which becomes the background for interpreting the messages of the intervention. The game design concept should be selected carefully and tested early with inexpensive methods, such as roleplay or paper sketches [32,56], to explore whether the gameplay is supportive of the intervention. In VitaNova, the gameplay goals implicitly reflected the learning goals, as the development of abilities was presented through the completion of in-game missions. However, because these learning goals were not explicitly communicated, the effectiveness of the intervention depended on the users themselves seeing the connection and transferring this knowledge to their own situation. Combined with the lack of clarity of game objectives and rules as well as an overly complex storyline, this led to confusion.

![Figure 7. Mismatch of goals between intervention and gameplay, in combination with insufficient clarity overall.](image)

**Practical Implications for Future Gaming-Based Interventions**

When revisiting the 3 themes and comparing them to related work, there appears to be a similarity between the identified findings of this study and the 3 basic psychological needs in SDT [103], which are: needs for autonomy, relatedness, and competence [92]; refer to Figure 8. As such, this study provides a form of bottom-up support for the usefulness of these constructs in designing and evaluating future gaming-based PPIs to understand how they might be more motivating to the user [92]. Further research should also investigate how and if a gaming-based PPI experience that does satisfy the relevant needs of autonomy, relatedness, and competence may contribute in itself to positive psychological outcomes for this population, as has been suggested [13].

Striking the right balance between learning and fun is a significant challenge, along with producing a relevant narrative [102] that supports intended learning. Ferrara [32] suggests a strategy for identifying the “gameness” that already exists in a context or situation rather than trying to tack it on. This may make it easier to transfer learning from the gaming space to everyday life [56,104]. However, moving from an idea to a game that works conceptually is challenging [105], and good intentions may be undermined by a seemingly fun yet unfit idea or concept, for example in the case of Disney and their first version of the game and exhibit “Habit Heroes,” intended to support healthy eating but rather reinforced stereotypes and made children feel bad about themselves [106]. Choosing an approach that “gamifies life” should thus be done with empathy, care, and frequent testing with users to avoid banalizing the situation and experiences of a vulnerable population, such as the young and unemployed. As such, a human-centered design approach is ideal because it starts with empathy [48]. However,
frequent playtesting [56] and usability evaluation [57] are also needed to reduce the risk of developing a concept that is not engaging with the intended audience [88] or that undermines or does not foster learning. Established game genres and concepts could be used as inspiration in early explorative ideation. The characteristics of existing games may be viewed as opposing values on a spectrum [32], and by imagining what the game-based PPI would look like in the form of existing game genres, a large volume of different ideas can be formed that may be tested early for fit with the PPI objectives and user preferences.

In Figure 9, we propose a broad but structured approach for how game-based PPI exploration may be executed, based on the lessons learned from our study and the discussion points in the preceding section. In this approach, insight into the user, context, gaming preferences, and gaming interests frames the design problem. It is also necessary to establish a clear and precise definition of PPI, including its underpinning mechanics, theories, and strategies that can help make the intervention effective. An alignment between the 2, forms the necessary “fertile soil” for the intervention game, where we ask how gameplay may support both user preferences and goals and PPI goals. Next, we propose working with existing genres and games to quickly generate many different ideas of what our gaming-based PPI may look and feel like. Promising concepts should be evaluated against relevant criteria, such as gameplay and user experience objectives and PPI objectives, and then made into prototypes for validation with user research.

The approach outlined here should be further detailed, refined, and validated in future research.

**Figure 8.** How the 3 themes correspond to basic psychological needs for relatedness, autonomy, and competence, in line with self-determination theory (SDT).
Figure 9. A proposed approach for designing future gaming-based positive psychology interventions (PPIs).

Pointers for Future Research

There is a range of possible strategies to choose from to work to improve an intervention and increase user motivation. In this specific study and for the purposes of this paper, we explored one possible strategy: to attempt to make the PPI more engaging, user-friendly, and relevant for young, unemployed adults by creating a gaming-based intervention. There are other alternative strategies that could increase relevance, appeal, and adaptation to the needs of the target audience. Kelders et al [24] suggest the use of design and persuasive design techniques, including reward, praise and reminders [107] as a tool to increase motivation and retention. Others [13] suggest the use of SDT [97], as we also found some support for this study. These strategies should be explored further in future studies.

Furthermore, the alignment between gameplay and intervention does not rely solely on the crafting of the game. Although our study grounded ideas on learning from past empirical research where PPIs had been applied to different contexts, there was a lack of clarity and theoretical grounding for the user experience in itself, including a clear definition of the learning [35] that should happen within the game design space. Incorporating learning theory, such as experiential learning [41], along with motivation theory, such as SDT [97] and persuasive design [108], as a more complete theoretical framework for the game designing process may provide a stronger direction to the conceptual work for the practitioners involved. Designing for...
behavioral and mindset change is increasingly relevant for design research and professional design practice [109], and there seem to be several gaps in understanding for design researchers and design teams who find themselves grappling with psychological and behavioral theories to produce interventions to support problem-solving of societal problems, such as youth unemployment.

Limitations
In this study, relevant participants were involved in a design process to capture their experience with designs in-the-making and take feedback into consideration in the design of revisions. We consider such early involvement a strength of the study. However, it also holds limitations; as the results are drawn from user experiences with prototype PPIs, the study does not provide user experiences resulting from a completed and verified PPI. Although the knowledge gained through the different stages of the design process is of substantial value to this area of research, future work is needed on experiences with fully functional gameplay PPIs to validate the findings of this study and to measure engagement, effectiveness, and adherence to the intervention. In addition, considering the fact that the positive psychologies implemented in the game mechanics were, to some extent, unclear to the participants after exposure, we cannot draw any conclusions on the experience of these in themselves at this stage. However, this was also not the purpose of this study.

Another limitation is the choice and availability of participants in the study. The population of unemployed young adults is highly heterogeneous. With our recruitment strategy, we are aware that we do not cover the entire range of end users, especially because we in part relied on voluntary registration and on contact with specific public welfare systems. Nevertheless, we find the involved participants to be within the scope of the studied PPI, and their feedback, hence, is of substantial benefit in understanding how the PPI may be experienced by representatives of this target group.

The third limitation concerns the context of the usability testing. Being observed by another person influences behavior (eg, Hawthorne effect), and participants likely spent much more time considering the prototypes than they would normally have. However, this approach was chosen because our interventions were prototypes and had unfinished functionality, which required a moderator to “fill the gaps” [110,111]. A fully self-administered and unmoderated use of a gameplay PPI would be a natural next step in future research.

Finally, it is important to note that, although our exploratory approach to insight into user perceptions of a game-based intervention for this target group is an important starting point for this area of investigation, future research is needed to establish the knowledge base needed to reliably provide such interventions. As part of this, we envision future studies with larger sample sizes and established scales as part of randomized controlled trials to gain further knowledge of the effectiveness of game-based interventions for this group and a basis for improvements in intervention design.

Conclusions
The study contributes insights into key user perceptions of game-based interventions for unemployed young adults. The contribution has implications for future game-like intervention design for this purpose. Our principal contribution is to explore engagement through a PPI, designed as an interactive game. We have described the iterative process of the development of a 3D-game concept, VitaNova, and have explored participants’ thoughts and feedback on their experiences. Although the participants were positive about the general idea of a game targeted toward unemployed young people, we found tensions between a PPI and an exciting game play and 3 themes that pinpoint priorities for future gaming implementations. Our study shows that interactive game design could make interventions more entertaining and engaging but can easily come into conflict with or undermine the intervention. We recommend aligning the gameplay narrative, objectives, and mechanics with intervention content and objectives to create engaging, relevant, and effective gaming-based PPIs that promote a more productive view of the challenges experienced by the young and unemployed.

Acknowledgments
The authors would like to thank the researchers at the University of Stavanger who contributed to this study: Venke F Haaland, Espen Sagen, Mari Rege, Hilde Ness Sandvold, and Wenche Ten Velden Hegelstad. The authors would also like to thank the Norwegian Welfare Administration in Sandnes and Skole- og Jobbresett at the University Hospital of Stavanger, as well as all the participants who took part in the study and gave their feedback on the different prototypes. The project is funded by the Norwegian Research Council (grant number 296390).

Conflicts of Interest
None declared.

Multimedia Appendix 1
A presentation of the different game prototypes.
[PDF File (Adobe PDF File), 2796 KB - humanfactors_v11i1e44423_app1.pdf ]

References


12. Isherwood K. Examining the effect of positive psychology interventions to improve psychological wellbeing, workplace motivation and re-employment prospects in North Wales. Bangor University. 2019. URL: http://tinyurl.com/3zvm6vxs [accessed 2023-08-02]


73. Straand et al. JMIR HUMAN FACTORS 2024 | vol. 11 | e44423 | p.31 https://humanfactors.jmir.org/2024/1/e44423

https://humanfactors.jmir.org/2024/1/e44423 JMIR Hum Factors 2024 | vol. 11 | e44423 | p.31 (page number not for citation purposes)


105. Straand et al. JMIR HUMAN FACTORS 2024 | vol. 11 | e44423 | p.32https://humanfactors.jmir.org/2024/1/e44423


**Abbreviations**

CV: curriculum vitae  
IPS: Individual Placement and Support  
NAV: Norwegian Labour and Welfare Administration  
NEET: not in education, employment, or training  
PPI: positive psychology intervention  
SDT: self-determination theory

©Ingjerd Jevnaker Straand, Asbjørn Følstad, Jone Ravndal Bjørnestad. Originally published in JMIR Human Factors (https://humanfactors.jmir.org), 18.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Human Factors, is properly cited. The complete bibliographic information, a link to the original publication on https://humanfactors.jmir.org, as well as this copyright and license information must be included.
User Perceptions of Visual Clot in a High-Fidelity Simulation Study: Mixed Qualitative-Quantitative Study

Greta Gasciauskaite¹, MD; Clara Castellucci¹, MD; Amos Malorgio¹, MD; Alexandra D Budowski¹, MD; Giovanna Schweiger¹, MD; Michaela Kolbe², Dr rer nat; Daniel Fries¹, MD; Bastian Grande¹, MD; Christoph B Nöthiger¹, MD; Donat R Spahn¹, MD; Tadzio R Roche¹, MD; David W Tscholl¹, MD; Samira Akbas¹, MD

¹Institute of Anesthesiology, University Hospital Zurich, Zurich, Switzerland
²Simulation Centre, University Hospital Zurich, Zurich, Switzerland

Corresponding Author:
Samira Akbas, MD
Institute of Anesthesiology
University Hospital Zurich
Raemistrasse 100
Zurich, 8091
Switzerland
Phone: 41 432532242
Email: Samira.Akbas@usz.ch

Abstract

Background: Viscoelastic hemostatic assays, such as rotational thromboelastometry (ROTEM) or thromboelastography, enable prompt diagnosis and accelerate targeted treatment. However, the complex interpretation of the results remains challenging. Visual Clot—a situation awareness-based visualization technology—was developed to assist clinicians in interpreting viscoelastic tests.

Objective: Following a previous high-fidelity simulation study, we analyzed users’ perceptions of the technology, to identify its strengths and limitations from clinicians’ perspectives.

Methods: This is a mixed qualitative-quantitative study consisting of interviews and a survey. After solving coagulation scenarios using Visual Clot in high-fidelity simulations, we interviewed anesthesia personnel about the perceived advantages and disadvantages of the new tool. We used a template approach to identify dominant themes in interview responses. From these themes, we defined 5 statements, which were then rated on Likert scales in a questionnaire.

Results: We interviewed 77 participants and 23 completed the survey. We identified 9 frequently mentioned topics by analyzing the interview responses. The most common themes were “positive design features,” “intuitive and easy to learn,” and “lack of a quantitative component.” In the survey, 21 respondents agreed that Visual Clot is easy to learn and 16 respondents stated that a combination of Visual Clot and ROTEM would help them manage complex hemostatic situations.

Conclusions: A group of anesthesia care providers found Visual Clot well-designed, intuitive, and easy to learn. Participants highlighted its usefulness in emergencies, especially for clinicians inexperienced in coagulation management. However, the lack of quantitative information is an area for improvement.

(JMIR Hum Factors 2024;11:e47991) doi:10.2196/47991

KEYWORDS
Visual Clot; blood coagulation; blood coagulation test; hemostasis; rotational thromboelastometry; situation awareness; user-centered design; visualization; user; perception; interpretation; thromboelastography; viscoelastic hemostatic; technology; coagulation; quantitative information

Introduction

Rapid hemostatic assessment is essential to targeted coagulation management in acute bleeding [1]. Increasingly used viscoelastic hemostatic assays, such as rotational thromboelastometry (ROTEM) or thromboelastography, enable faster insights into coagulation dysfunction than conventional laboratory tests. Standard coagulation assays are not optimal for managing acute hemorrhages that require rapid therapeutic action, as it often takes more than an hour to obtain the results [2,3].
and North American transfusion recommendations underline the advantages of viscoelastic hemostatic assays for managing trauma and severe perioperative bleeding, including the reduced need for transfusions, fewer perioperative complications, shorter hospitalization, and lower overall treatment costs [1,4,5]. Its usefulness has been demonstrated in many operative areas such as obstetrics [6,7], pediatric surgery [8], transplantation [9], cardiac surgery [10,11], neurosurgery [12], and burn surgery [13,14]. Viscoelastic hemostatic tests are also paramount in the diagnostic and treatment adjustment of hematological disorders, such as inherited afibrinogenemia, hemophilia, or multiple myeloma [15-18]. However, despite these technologies' widespread use and considerable advantages, their results' interpretation remains challenging and requires well-trained clinical personnel [19-21]. Visual Clot—a situation awareness-based visualization technology—was developed to support health care professionals in interpreting viscoelastic test results by reducing the complexity of their presentation. Based on raw ROTEM data, the results are displayed in real time as a 3D animated model of a blood clot to represent various elements of hemostasis, including platelets, plasmatic factors, and fibrin. It can also effectively illustrate the influence of heparin and hyperfibrinolysis [22] (Multimedia Appendix 1).

In a high-fidelity simulation study, anesthesiologists using Visual Clot were 2.2 times more likely to articulate the correct therapeutic approach. In addition, these anesthesiologists had a lower median time to administer the first appropriate targeted coagulation product. Overall, physicians presented with the results of viscoelastic testing using Visual Clot were approximately 56% more likely to provide accurate therapeutic interventions. In the same study, physicians were 3.5 times more likely to feel confident in their decisions when working with Visual Clot compared to traditional ROTEM results [23]. In the first computer-based study analyzing user perceptions of Visual Clot, participants described the technology as well-designed, easy to learn, and intuitive [24]. The guiding principles of the Visual Clot technology that result in enhanced situation awareness include Endsley's user-centered design principles [25]. Wittgenstein's philosophy as articulated in Tractatus Logico-Philosophicus [26], and insights from the National Aeronautics and Space Administration (NASA) publication “On Organization of Information: Approach and Early Work” by Degani et al [27]. Endsley's principles emphasize the use of direct visual representations of data to enhance situational awareness, a central principle in Visual Clot's data visualization. Wittgenstein's theory emphasizes the importance of logical representations that meaningfully correspond to the reality they are intended to represent. Visual Clot follows this principle by visually representing elements such as fibrin, platelets, plasmatic factors, hyperfibrinolysis, and bleeding. Following NASA's approach, Visual Clot strives to achieve the highest level of "order and wholeness" by consolidating all essential data into a single display. The primary goal of Visual Clot technology is to provide the care provider with situational information quickly and with minimal cognitive load. In this study, we aimed to capture and analyze perceptions of anesthesia personnel working with Visual Clot in a high-fidelity simulation to identify the strengths and recognize the potential for future improvements.

## Methods

### Ethical Considerations
The Cantonal Ethics Committee of the Canton of Zurich reviewed the study protocol and issued a declaration of no objection (Business Management System for Ethics Committees Number Req-2021-01112). Furthermore, each participant gave informed consent to use his or her data for research purposes. Participation was voluntary and without financial compensation.

### Study Design
We conducted a researcher-initiated single-center mixed qualitative-quantitative study at the University Hospital Zurich, Institute of Anesthesiology, Switzerland. Study participants were anesthesia personnel, including staff anesthesiologists, residents, and nurses. After participating in a high-fidelity simulation study of perioperative bleeding scenarios, where they worked with Visual Clot and ROTEM, we interviewed participants on their perceptions of Visual Clot technology.

As a second step, the same participants received an email invitation to participate in a survey a few weeks later. They rated statements we generated from identified and frequently mentioned themes in interview responses on a Likert scale.

### Previous High-Fidelity Simulation Study
In the high-fidelity simulation study [23], anesthesia teams, composed of a staff anesthesiologist, a resident, and an anesthesia nurse, participated in high-fidelity perioperative bleeding scenarios using either Visual Clot or ROTEM. The primary outcome of the study was correct targeted coagulation therapy. Secondary outcomes were time to targeted coagulation therapy, confidence, and workload.

ROTEM is the standard of care for managing acute hemorrhage in the study center, so all participants were familiar with the technology before participating [20]. Some participants had taken part in previous Visual Clot studies and, therefore, were already familiar with the technology [22,24].

Nevertheless, before the simulations began, we gave a 10-minute presentation that reviewed ROTEM and introduced Visual Clot. Multimedia Appendix 1 provides an instructional video of Visual Clot. Participants were invited to ask questions freely before starting work in the simulation environment. Each team solved 1 of 4 different perioperative bleeding scenarios, which were randomly allocated. We ended the scenarios when all necessary therapeutic measures were derived or, at the latest, after 15 minutes. Figure 1 illustrates an example of a Visual Clot printout used in the simulation study.
Figure 1. A Visual Clot result presentation showing a fibrin deficiency. The fibrin in the clot is shown as a dashed line, indicating its absence. The blood drops indicate the presence of a coagulation pathology.

Participant Interviews
After the simulations, we encouraged the participants to freely verbalize their thoughts in a distraction-free environment while the data collectors made field notes. The only suggestion to the participants before the interviews was to verbalize their positive and negative opinions of Visual Clot. The participants could define final adjustments in the collected answers at the end of the interviews.

Survey
In the second step, we formulated 5 statements to summarize the insights gathered during the interviews. The statements were submitted for evaluation on a 5-point Likert scale graded from “strongly agree” to “strongly disagree.” An email invitation was sent to all interviewed participants.

Outcomes and Statistical Analyses
Part I: Participant Interviews
Collected interview responses were translated from original German to English using a translation system DeepL (DeepL GmbH). Multimedia Appendix 2 provides the complete translated field notes.

The most commonly used terms in positive and negative responses were identified using the word count function. Word groups with the same root were united, excluding the frequently used filler words such as “to,” “and,” or “the” (Table 1). Using a template approach [28] we identified the major themes that dominated participants’ answers. As a result, we generated a coding tree (Figure 2). According to the coding template, we assigned statements to the themes. A total of 3 of the study authors, all anesthesiology residents GG, GS, and SA, rated the interview statements separately from each other using the coding tree (Figure 2). If the 3 investigators disagreed after multiple data coding, the final decision was taken in a joint discussion. Interrater reliability was calculated to investigate the consistency of the coding tree’s application.
Table 1. The most commonly used positive and negative terms to describe Visual Clot.

<table>
<thead>
<tr>
<th>Terms</th>
<th>Frequency, n</th>
<th>Terms</th>
<th>Frequency, n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy or easier</td>
<td>28</td>
<td>Missing</td>
<td>13</td>
</tr>
<tr>
<td>Good</td>
<td>16</td>
<td>Quantitative or quantification</td>
<td>11</td>
</tr>
<tr>
<td>Fast or faster</td>
<td>13</td>
<td>Information</td>
<td>9</td>
</tr>
<tr>
<td>See a problem</td>
<td>12</td>
<td>Values</td>
<td>8</td>
</tr>
<tr>
<td>Interpret quickly</td>
<td>11</td>
<td>Time</td>
<td>8</td>
</tr>
<tr>
<td>Simple</td>
<td>9</td>
<td>Fibrinogen</td>
<td>6</td>
</tr>
<tr>
<td>Understand</td>
<td>9</td>
<td>Numbers</td>
<td>6</td>
</tr>
<tr>
<td>Visual</td>
<td>8</td>
<td>Hyperfibrinolysis</td>
<td>6</td>
</tr>
<tr>
<td>Interpretation</td>
<td>7</td>
<td>Less</td>
<td>5</td>
</tr>
<tr>
<td>Intuitive</td>
<td>7</td>
<td>Confusing</td>
<td>4</td>
</tr>
<tr>
<td>At a glance</td>
<td>7</td>
<td>Flashing</td>
<td>4</td>
</tr>
<tr>
<td>Overview</td>
<td>7</td>
<td>Simplified</td>
<td>3</td>
</tr>
</tbody>
</table>

Figure 2. A coding tree representing the themes describing positive and negative user perceptions. ROTEM: rotational thromboelastometry.

Part II: Survey

The literature states that quantitative data can help generalize and confirm specific observations found in qualitative research [29-32]. For the subsequent survey, we defined 5 statements based on the previously identified themes. The same group of interviewed anesthesiologists was asked to rate them on 5-point Likert scales in a questionnaire created using Google Forms (Alphabet Inc). Participants were informed that the survey takes only a few minutes to complete, participation is voluntary, and no compensation is offered. The translated announcement of the survey invitation is displayed in Multimedia Appendix 3. The data collection was finished 3 weeks after the questionnaire was sent.

Statistical Analysis

The interview data analysis and figures were made using Microsoft Word and Excel (Microsoft Corp). We present the number of statements and their percentage distribution in the identified themes.

To define the interrater reliability of the coding template, we calculated Fleiss’ Kappa using R (version 4.0.5; R Foundation for Statistical Computing). We calculated every statement’s median and IQR for the survey analysis. We used the Wilcoxon signed rank test to determine the difference between the median and neutral answers. Statistical significance was indicated as $P<.05$. 
Results

Study and Participant Characteristics
Detailed information on the study and participants is provided in Table 2. Residents and nurses were the dominant participants in the interviews. The most experienced participant had 33 years of experience in anesthesia. The least experienced had less than 1 year. Residents and nurses also dominated the survey.

Table 2. Study and participant characteristics.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study characteristics</strong></td>
<td></td>
</tr>
<tr>
<td>Total number of interviewed</td>
<td>77</td>
</tr>
<tr>
<td>participants, n</td>
<td></td>
</tr>
<tr>
<td>Total number of participants</td>
<td>23</td>
</tr>
<tr>
<td>completed the survey, n</td>
<td></td>
</tr>
<tr>
<td><strong>Participant characteristics</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Interview participants</strong></td>
<td></td>
</tr>
<tr>
<td>Staff physicians, n (%)</td>
<td>8 (10.4)</td>
</tr>
<tr>
<td>Residents, n (%)</td>
<td>35 (45.5)</td>
</tr>
<tr>
<td>Anesthesiology nurses, n (%)</td>
<td>34 (44.2)</td>
</tr>
<tr>
<td>Anesthesia experience in years,</td>
<td>8 (3-10)</td>
</tr>
<tr>
<td>median (IQR)</td>
<td></td>
</tr>
<tr>
<td>Number of ROTEM(^a) interpretations per year, median (IQR)</td>
<td>26 (5-41)</td>
</tr>
<tr>
<td><strong>Survey participants</strong></td>
<td></td>
</tr>
<tr>
<td>Staff physicians, n (%)</td>
<td>7 (30.4)</td>
</tr>
<tr>
<td>Residents, n (%)</td>
<td>8 (34.8)</td>
</tr>
<tr>
<td>Anesthesiology nurses, n (%)</td>
<td>8 (34.8)</td>
</tr>
</tbody>
</table>

\(^a\)ROTEM: rotational thromboelastometry.

Part I: Qualitative Analysis of Interview Answers

Word Count Analysis
The most frequently used words and word combinations used to describe the advantages of Visual Clot were: easier or easy (26/77, 33.8%), interpret or interpretation (23.4%, 18/77 participants), quick or quickly (19.5%, 15/77 participants), visual, visualize, visualized, or visualization (19.5%, 15/77 participants), good (16.9%, 13/77 participants), faster or fast (15.6%, 12/77 participants). In the group of statements describing the limitations of Visual Clot, the words and word groups most frequently used were: ROTEM (23.4%, 18/77 participants), missing (information or values or numbers; 16.9%, 13/77 participants), quantitative or quantification (16.9%, 13/77 participants). Table 1 visually represents the most commonly used words in positive and negative perceptions.

Coding Tree
Figure 2 shows the generated coding tree, including 2 main domains and 9 themes. The interrater reliability of the tree raters was 0.856 (95% CI 0.831-0.880), indicating almost perfect agreement [33].

Statements Describing Visual Clot
Table 3 demonstrates examples of statements assigned to particular subtopics with participant counts and percentages.

A total of 4 comments were defined as positive but not assigned to any themes. There was 1 such statement in the negative group. A total of 19 comments were not assigned to any theme and were described as noncodable.
Table 3. Statements examples assigned to particular domain and subtopics with participant count and percentages.

<table>
<thead>
<tr>
<th>Major domain and subtopics</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive statements describing Visual Clot (179/319, 56.1%)</td>
<td>“Very simplified” (Participant 13).</td>
</tr>
<tr>
<td>Positive design features (63/295, 21.4%)</td>
<td>“Visual presentation” (Participant 21).</td>
</tr>
<tr>
<td>Positive usability features (30/295, 10.2%)</td>
<td>“Tells figuratively what to do” (Participant 26).</td>
</tr>
<tr>
<td>Intuitive and easy to learn (61/295, 20.7%)</td>
<td>“It can be perfectly integrated in the clinic” (Participant 24).</td>
</tr>
<tr>
<td>Time saving (21/295, 7.1%)</td>
<td>“A good tool to get an overview” (Participant 32).</td>
</tr>
<tr>
<td>Negative statements describing Visual Clot (113/319, 35.4%)</td>
<td>“You can see at-a-glance what is missing” (Participant 70).</td>
</tr>
<tr>
<td>Design problems (23/295, 7.8%)</td>
<td>“Very simplified” (Participant 13).</td>
</tr>
<tr>
<td>Usability problems (12/295, 4.1%)</td>
<td>“Visual presentation” (Participant 21).</td>
</tr>
<tr>
<td>Incompleteness: a lack of a quantitative component (62/295, 21.0%)</td>
<td>“Tells figuratively what to do” (Participant 26).</td>
</tr>
<tr>
<td>Need for training (15/295, 5.1%)</td>
<td>“It can be perfectly integrated in the clinic” (Participant 24).</td>
</tr>
<tr>
<td>Combination: Visual Clot and ROTEMa (8/295, 2.7%)</td>
<td>“A good tool to get an overview” (Participant 32).</td>
</tr>
</tbody>
</table>

aROTEM: rotational thromboelastometry.

**Positive Statements Describing Visual Clot**

**Positive Design Features**

Most comments were made on this topic, emphasizing that a “pictorially summarized” (participant 2) and “visually appealing” (participant 3) data presentation allows one to see “the relevant ROTEM information at-a-glance” (participant 9). Such a design supports health care professionals in making clinical decisions. It is essential in emergencies because the actual coagulation status is immediately visible (participant 14) and it is instantly apparent which hemostasis components are missing (participant 12).

**Positive Usability Features**

Visual Clot is “a good tool for broad application,” stated participant 72. It enables “pre-interpretation of the complex information” (participant 3) and focuses “on the essential” (participant 65). The benefits of the Visual Clot in urgent situations were also highlighted: the technology is “very good for emergencies,” stated participant 13.

**Intuitive and Easy to Learn**

As in the previous study [24], the Visual Clot was also described here as intuitive and easy to learn. “Very intuitive, short time needed to understand it,”—pointed out participant 6. It was underlined that visualizations provided by the Visual Clot are “quickly recognizable even by untrained persons or with little knowledge of coagulation” (participant 13).

**Time Saving**

The Visual Clot provides an “overview at-a-glance,” as participant 76 said. “I immediately saw what was missing,” stated participant 16. These features lead to quicker diagnosis—“focus is faster on the problem”—as participant 46 said, and thus to faster initiation of treatment.
Negative Statements Describing Visual Clot

Design Problems
Several ideas that could potentially enhance Visual Clot’s design were identified. Participant 47 pointed out that the presentation of platelets and fibrinogen are similar, and thus it is difficult to distinguish. Participant 48 also agreed: “I did not notice that platelets were missing because it was white and dashed like fibrinogen.” Some participants found that the Visual Clot is too dynamic—too much movement on the screen, which can lead to distraction and make the interpretation of the results difficult. “Even if coagulation status is fine, everything is moving, and you can poorly differentiate what is missing” (participant 53), “moves too much, even if everything is fine—distraction” (participant 55).

Usability Problems
Visual Clot is “confusing at the beginning”—stated participant 31 and added that it is “difficult to use without routine.” Visual Clot provides “too much information at once,”—participant 53 pointed out.

Incompleteness: Lack of a Quantitative Component
The central Visual Clot aspect criticized was the technology’s incompleteness in terms of lacking a quantitative component.

Several participants stated that the Visual Clot is “not precise” (participant 1), which can be explained in the words of participants 53 and 9, respectively, who said that in the Visual Clot “quantitative is missing” and that one “can get more information with the ROTEM.”

Need for Training
The main point identified in the participants’ opinions on this topic was the lack of experience working with this technology and that it is a very new tool not yet established in clinical practice.

Combination of Visual Clot and ROTEM
Several participants said they could benefit from combining the Visual Clot and ROTEM when interpreting coagulation assays. “A combination of Visual Clot and ROTEM would be perfect,” pointed out participant 19, while participant 74 said, “a combination of both is necessary.” There was no difference in positive and negative statements based on participants’ specialty or level of experience.

Part II: Analysis of Statements Assessed in the Survey

Figure 3 shows the detailed evaluation of the statements rated in the survey.

Discussion

Principal Findings
This mixed qualitative-quantitative study analyzed the perceptions of anesthesiology personnel regarding Visual Clot—a new situation-awareness and user-oriented visualization technology for viscoelastic hemostatic resuscitation—after the high-fidelity simulation study. User perceptions enable us to identify the positive aspects of the technology and reveal the potential for improvement in the future. After computer-based...
studies, this is the first time that Visual Clot has been evaluated in a high-fidelity simulation study, a validated process for testing a noncertified product in an environment that closely resembles clinical reality [34,35].

The principal findings demonstrate that the design features of Visual Clot have received the most positive comments. As in the previous computer-based Visual Clot study [24], the participants of this high-fidelity simulation study emphasized that the way this technology is designed provides a good overview of the clotting situation and is an additional help in the decision-making process during acute bleeding situations. Further, Visual Clot was described as intuitive and easy to learn. Participants repeatedly mentioned that the results of Visual Clot are quickly recognizable and understandable even by inexperienced clinicians. The main criticism concerned the lack of quantitative information.

Previous Visual Clot studies [22-24,36,37] underline the benefits of additional visualization technology, simplifying standard ROTEM data interpretation [20]. Anesthesia providers using Visual Clot in a high-fidelity simulation study were more likely to correctly administer targeted coagulation therapy and to give the first targeted coagulation product faster. In addition, participants demonstrated greater decision-making confidence with Visual Clot [23]. Moreover, the correctness of the clinical decisions was independent of previous rotational thromboelastometry knowledge and experience.

The superior participants’ performance when working with Visual Clot may be explained by its design supporting the strengths of human sensory perception. The Visual Clot was developed to assist care providers in managing highly complex coagulation situations, presenting the data in an awareness-oriented interface design. The main aim of this design is to convey the information as quickly as possible and with the lowest cognitive effort [25].

Principles of situation awareness-oriented and user-centered design enable effective data management and a comprehensive understanding of what is happening and thus help to stay situationally aware. This concept is essential in many domains, including medicine, where managing complex and dynamic situations is fundamental [38,39].

Its definition breaks down into three separate phases (1) the perception of environmental elements in the current situation within a volume of time and space, (2) understanding their meaning, and (3) their projection in the near future. Based on this, the Visual Clot data are visually represented, preprocessed, and simplified. The results of coagulation parameters are divided into 3 categories: too low, normal, or too high. Such information presentation increases diagnostic confidence, but numeric indicators are needed for precise data analysis and targeted treatment initiation. As previously indicated, the lack of quantitative information is reflected in user responses. It also explains the participants’ considerations that combining Visual Clot and ROTEM would be helpful in clinical decision-making.

Some other technologies based on situation awareness and user-centered design principles include Philips Visual Patient Avatar (Philips) [40], AlertWatch (AlertWatch Inc), Dynamic Lung Panel and PulmoSight (Hamilton Medical AG), HemoSight and Physiology Screen (Mindray Medical International Limited), and Alarm Status Visualizer (Masimo Corp) [41,42].

This study showed user perceptions regarding the new situation awareness-based, user-oriented technology for thromboelastometry data presentation—Visual Clot. It makes us aware of the user’s needs and could help us simplify information processing and decision-making in the future. An integral facet of advancing the technological framework informed by the results of this study lies in the prospect of merging quantitative data into the Visual Clot platform and presenting this merged information in a consolidated interface. This concerted integration promises to align both quantitative and qualitative data to provide a more complete and accurate representation of prevailing conditions. This integration can be achieved in a variety of ways, including the direct overlay of numerical values onto the Visual Clot visualization, or the parallel juxtaposition of a complementary graphical representation alongside the numerical data set.

Strengths and Limitations

This study has several strengths and limitations. The interview part of the study has the typical limitations of qualitative research. The findings of qualitative analysis cannot be extrapolated to larger populations with the same certainty as quantitative results because the findings are on the subjective basis and not tested for statistical significance [43]. However, the quantitative survey helped to provide greater insight into the importance of the main themes identified. Moreover, the interviewed participants were selected according to their availability in the clinical praxis and not randomly.

Furthermore, the number of participants in the survey was lower than in the interviews because not all participants in the simulation study completed the survey. Finally, it is a single-center study performed in a university hospital with high care standards in Europe. User perceptions may vary across diverse clinical settings in different parts of the world.

Conclusions

After previous studies investigating user perceptions of Visual Clot in computer-based simulation studies, this is the first study to analyze the user perceptions of Visual Clot in a high-fidelity simulation—the intermediate step between computer-based simulation studies in a laboratory and real-life use. In this study, Visual Clot appeared to be a well-accepted additional tool supporting health care professionals working with ROTEM. Based on participants’ perceptions, user-centered and situation awareness-oriented design, as shown in Visual Clot, can simplify the presentation of complex information and thus make critical decision-making quicker and more efficient. The benefits of this technology have been particularly highlighted in emergencies and even for care providers with little experience in coagulation management. Participants described Visual Clot as intuitive and easy to learn. The lack of a quantitative component has been identified as a significant limitation. These findings highlight the advantages of Visual Clot and its potential...
for improvement may help further develop this and other situation awareness-based technologies.

Acknowledgments
The authors are thankful to the study participants for their time and effort. The Institute of Anesthesiology of the University Hospital of Zurich, Zurich, Switzerland and the University of Zurich, Zurich, Switzerland funded this study.

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

Authors’ Contributions
GG, AM, CC, ADB, GS, MK, BG, CBN, TRR, DWT, and SA contributed to the conceptualization; GG, AM, CC, ADB, GS, MK, BG, CBN, TRR, DWT, and SA contributed to the methodology; GG, GS, and SA performed the formal analysis; AM, CC, ADB, GS, TRR, and DWT contributed to the investigation; GG, GS, and SA performed data curation; GG and DWT contributed to writing—original draft preparation; GG, DRS, CBN, DWT, SA, and DF contributed to writing—review and editing; GG and DWT contributed to visualization; DWT and SA performed supervision; and DWT performed project administration.

Conflicts of Interest
DRS’s academic department is receiving grant support from the Swiss National Science Foundation, Berne, Switzerland, the Swiss Society of Anesthesiology and Perioperative Medicine, Berne, Switzerland, the Swiss Foundation for Anesthesia Research, Zurich, Switzerland, Vifor SA, Villars-sur-Glâne, Switzerland and Vifor (International) AG, St. Gallen, Switzerland. DRS is cochair of the ABC-Trauma Faculty, sponsored by unrestricted educational grants from Novo Nordisk Health Care AG, Zurich, Switzerland, CSL Behring GmbH, Marburg, Germany, LFB Biomédicaments, Courtaboeuf Cedex, France and Octapharma AG, Lachen, Switzerland. DRS received honoraria or travel support for consulting or lecturing from Alliance Rouge, Bern, Switzerland, Danube University of Krems, Austria, European Society of Anesthesiology and Intensive Care, Brussels, BE, International Foundation for Patient Blood Management, Basel, Switzerland, Korean Society of Anesthesiologists, Seoul, Korea, Network for the Advancement of Patient Blood Management, Haemostasis and Thrombosis, Paris, France, Society for the Advancement of Blood Management, Mount Royal NJ, Alexion Pharmaceuticals Inc, Boston, MA, AstraZeneca AG, Baar, Switzerland, Bayer AG, Zürich, Switzerland, B. Braun Melsungen AG, Melsungen, Germany, Baxter AG, Glattpark, Switzerland, CSL Behring GmbH, Hattersheim am Main, Germany and Berne, Switzerland, CSL Vifor (Switzerland) Villars-sur-Glâne, Switzerland, CSL Vifor (International), St Gallen, Switzerland, Celgene International II Särl, Couvet, Switzerland, Daichi Sankyo AG, Thalwil, Switzerland, Haemonetics, Braintree, Massachusetts, United States, LFB Biomédicaments, Courtaboeuf Cedex, France, Merck Sharp & Dohme, Kenilworth, New Jersey, United States, Novo Nordisk Health Care AG, Zurich, Switzerland, Octapharma AG, Lachen, Switzerland, Pharmacosmos A/S, Holbaek, Denmark, Pierre Fabre Pharma, Aeschwil, Switzerland, Portola Schweiz GmbH, Aarau, Switzerland, Roche Diagnostics International Ltd, Reinhach, Switzerland, Sarstedt AG & Co, Sevelen, Switzerland and Nürnberg, Germany, Shire Switzerland GmbH, Zug, Switzerland, Takeda, Glattpark, Switzerland, Werfen, Bedford, MA, Zueilig Pharma Holdings, Singapore, Singapore. CBN is an inventor of Visual Patient and Visual Patient Predictive technologies, for which the University of Zurich and Koninklijke Philips N.V. hold patents, patent applications, design protections, and trademarks. Joint-development and licensing agreements exist with Philips Medizin Systeme Böblingen GmbH, Böblingen, Germany; Koninklijke Philips N.V., Amsterdam, The Netherlands; Philips Research/Philips Electronics Nederland BV, Eindhoven, The Netherlands; and Philips USA, Cambridge, Massachusetts, United States. Within the framework of these agreements, CBN receives travel support, lecturing and consulting honoraria, and may potentially receive royalties in the event of successful commercialization. CBN is an inventor of Visual Clot technology, with patent applications, design protections, and trademarks. Joint-development and licensing agreements exist with Philips Medizin Systeme Böblingen GmbH, Böblingen, Germany; Koninklijke Philips N.V., Amsterdam, The Netherlands; Philips Research/Philips Electronics Nederland BV, Eindhoven, The Netherlands; and Philips USA, Cambridge, Massachusetts, United States. Within the framework of these agreements, CBN receives travel support, lecturing and consulting honoraria, and may potentially receive royalties in the event of successful commercialization. CBN is an inventor of Visual Blood technology, for which the University of Zurich holds patent applications and design protections; potential royalties may follow successful commercialization. CBN is an inventor of Visual Blood technology, for which the University of Zurich holds patent applications and design protections; potential royalties may follow successful commercialization. CBN is an inventor of Visual Blood technology, for which the University of Zurich holds patent applications and design protections; potential royalties may follow successful commercialization.
Multimedia Appendix 1
Educational Visual Clot video.
[MOV File, 19983 KB - humanfactors_v11i1e47991_app1.mov]

Multimedia Appendix 2
Complete translated field notes of participant interviews.
[PDF File (Adobe PDF File), 669 KB - humanfactors_v11i1e47991_app2.pdf]

Multimedia Appendix 3
Translated announcement of the survey invitation.
[PDF File (Adobe PDF File), 204 KB - humanfactors_v11i1e47991_app3.pdf]

References

Download full-size versions of the cited references. These files are too large to be included in this email.


41. Products. mindray. 2023. URL: https://www.mindray.com/uk/products [accessed 2023-12-22]


Abbreviations

NASA: National Aeronautics and Space Administration
ROTEM: rotational thromboelastometry
Original Paper

Digital Care Pathway for Patients With Sleep Apnea in Specialized Care: Mixed Methods Study

Jari Haverinen\textsuperscript{1,2}, MSc, MHS; Terttu Harju\textsuperscript{3}, MD, PhD; Hanna Mikkonen\textsuperscript{4}, RNC; Pia Liljamo\textsuperscript{5}, RNC, PhD; Miia Turpeinen\textsuperscript{6}, MD, PhD; Jarmo Reponen\textsuperscript{2,6}, MD, PhD

\textsuperscript{1}Finnish Coordinating Center for Health Technology Assessment, Oulu University Hospital, Oulu, Finland
\textsuperscript{2}FinnTelemedicum, Research Unit of Health Sciences and Technology, Faculty of Medicine, University of Oulu, Oulu, Finland
\textsuperscript{3}Medical Research Center Oulu, Oulu Pulmonary Department, Oulu University Hospital and University of Oulu, Oulu, Finland
\textsuperscript{4}The Wellbeing Services County of North Ostrobothnia, Oulu, Finland
\textsuperscript{5}Finnish Institute for Health and Welfare, Department of Knowledge Brokers, Data and Analytics Unit, Helsinki, Finland
\textsuperscript{6}Medical Research Center Oulu, Oulu University Hospital and University of Oulu, Oulu, Finland

Corresponding Author:
Jari Haverinen, MSc, MHS
Finnish Coordinating Center for Health Technology Assessment
Oulu University Hospital
Kajaanintie 50
Oulu, PL10
Finland
Phone: 358 504095446
Email: jari.haverinen@pohde.fi

Abstract

\textbf{Background:} Sleep apnea is a significant public health disorder in Finland, with a prevalence of 3.7\%. Continuous positive airway pressure (CPAP) therapy is the first-line treatment for moderate or severe sleep apnea. From November 18, 2019, all patients who started their CPAP therapy at Oulu University Hospital were attached to a sleep apnea digital care pathway (SA-DCP) and were instructed on its use. Some patients still did not use the SA-DCP although they had started their CPAP therapy.

\textbf{Objective:} We aimed to study health care professionals’ (HCPs’) perspectives on the SA-DCP and its usefulness for their work; whether the main targets of SA-DCP can be reached: shortening the initial guiding sessions of CPAP therapy, reducing patient calls and contact with HCPs, and improving patients’ adherence to CPAP therapy; and patients’ perspectives on the SA-DCP and its usefulness to them.

\textbf{Methods:} Overall, 6 HCPs were interviewed in May and June 2021. The survey for SA-DCP users (58/91, 64\%) and SA-DCP nonusers (33/91, 36\%) was conducted in 2 phases: from May to August 2021 and January to June 2022. CPAP device remote monitoring data were collected from SA-DCP users (80/170, 47.1\%) and SA-DCP nonusers (90/170, 52.9\%) in May 2021. The registered phone call data were collected during 2019, 2020, and 2021. Feedback on the SA-DCP was collected from 446 patients between February and March 2022.

\textbf{Results:} According to HCPs, introducing the SA-DCP had not yet significantly improved their workload and work practices, but it had brought more flexibility in some communication situations. A larger proportion of SA-DCP users familiarized themselves with prior information about CPAP therapy before the initial guiding session than nonusers (43/58, 74\% vs 16/33, 49\%; \textit{P}=.02). Some patients still had not received prior information about CPAP therapy; therefore, most of the sessions were carried out according to their needs. According to the patient survey and remote monitoring data of CPAP devices, adherence to CPAP therapy was high for both SA-DCP users and nonusers. The number of patients’ phone calls to HCPs did not decrease during the study. SA-DCP users perceived their abilities to use information and communications technology to be better than nonusers (mean 4.2, SD 0.8 vs mean 3.2, SD 1.2; \textit{P}<.001).

\textbf{Conclusions:} According to this study, not all the goals set for the introduction of the SA-DCP have been achieved. Despite using the SA-DCP, some patients still wanted to communicate with HCPs by phone. The most significant factors explaining the nonuse of the SA-DCP were lower digital literacy and older age of the patients. In the future, more attention should be paid to these user groups when designing and introducing upcoming digital care pathways.
Introduction

Background

Sleep apnea is a significant public health disorder in Finland, with a prevalence of 3.7%. The prevalence of sleep apnea worldwide has been increasing in relation to the obesity pandemic [1]. Untreated sleep apnea increases cardiovascular diseases, accidents, likelihood of taking sick leave, and premature mortality [2]. The clinical severity of sleep apnea is defined based on 3 components: daytime sleepiness owing to sleep apnea, the apnea-hypopnea index (AHI), and arterial blood oxygen saturation [2]. Continuous positive airway pressure (CPAP) therapy is the first-line treatment for moderate or severe sleep apnea in addition to conservative therapy (ie, weight loss, avoidance of sleep-disturbing substances, and lifestyle issues) [2]. CPAP therapy is a safe and efficient treatment for sleep apnea, relieving both daytime and nighttime symptoms and improving traffic safety [3,4]. In Finland, the need for CPAP treatment and the number of outpatient visits in both specialized and primary care have increased considerably because of the increased number of patients with sleep apnea [5].

The digitalization of health care has been seen as a potential option for offering treatment to patients regardless of time and place and involving them in their own care [6,7]. In addition, digitalization has the potential to make health care systems more efficient [8]. Despite its potential to improve health care services, digitalization does not automatically guarantee better services [9]. It has been noted as a problem, for example, that digital services are not necessarily aligned with clinician and patient preferences [9]. The challenge is that, in some cases, they complement rather than substitute the current services, and care processes are not always redesigned to achieve the best benefits from digital services [9-12]. Citizens’ willingness and ability to use electronic services is also an obstacle to realizing the benefits of health care digitalization [13,14]. Because data breaches cause potentially catastrophic consequences, information security concerns have weakened patients’ adoption of digital health services [15,16]. The challenges of the technical implementation of digital services, such as missing functionalities and lack of interoperability with existing information systems, have weakened the willingness of health care professionals (HCPs) to use them [10].

Factors promoting the adoption of digital health services are their perceived benefits for patients and patients’ previous positive experiences with electronic services [13,15]. Previous studies showed that digital health interventions can improve patients’ adherence to their care [17,18]. For example, Aardoom et al [18] showed that adherence to CPAP therapy in patients with sleep apnea can be improved with digital interventions in the initial months of treatment. Adherence to the use of the digital health service has also been found in some studies to positively affect outcomes [19,20]. Good digital literacy promotes the use of digital health services; studies have found young people have better digital literacy than older age groups [13,21]. As the user base of digital health care services can be very broad, and users can have functional limitations owing to age or illness, the ease of use of these services is important in promoting their use [15,22].

Finland’s first phase of health care digitalization involved the digitizing of HCPs’ tools, such as electronic patient records; e-prescribing and digitalization have progressed well [23]. Currently, Finnish citizens are increasingly offered digital health care services and products [22,24]. Several countries, including Finland, have introduced new health technology assessment methods to ensure that digital health provides evidence-based benefits [16,22,25]. Digital care pathways (DCPs) are an example of digital health care services, and today, there are >300 DCPs in use in Finnish specialized care units [26]. One of the main goals of DCPs is to complement or replace traditional health care appointment visits [26]. In addition, DCPs aim to support and help in the self-treatment of long-term illnesses, monitoring, and adaptation to the illness, as well as enable patients to prepare for various health care procedures beforehand [26]. Several DCPs have been studied in Finland from the perspective of HCPs, organizations, and patients [7,10,13,27-31]. One of these DCPs is the sleep apnea DCP (SA-DCP), which was introduced at Oulu University Hospital (OUH) on November 18, 2019 [32]. All patients who start their CPAP therapy in OUH will be attached to the SA-DCP, that is, their patient data will be recorded in it, and they will be instructed on how to log in and use it [32]. When a patient starts on the SA-DCP, they register as an SA-DCP user through strong identification by accepting the terms of use and privacy statement and entering his or her contact information [33].

Objectives

In OUH, the CPAP therapy for patients with sleep apnea begins with an initial guiding session where patients are instructed on using their CPAP device. SA-DCP contains information and instructions about CPAP therapy; therefore, it would be desirable for patients to familiarize themselves with that information beforehand [32]. In this way, the initial guiding session of CPAP therapy could be shortened because the basic information about CPAP therapy would not need to be reviewed again during the sessions. The SA-DCP contains reliable information about sleep apnea, its treatment, and CPAP therapy [32]. With the introduction of SA-DCP, it would be desirable to reduce patients’ phone calls and other contacts with HCPs when information can be found in the SA-DCP. The SA-DCP also includes electronic messaging between patients and HCPs, which could reduce such calls [32]. The major aim of the SA-DCP is to increase patients’ adherence to CPAP therapy. However, there is still a challenge in that some patients with sleep apnea do not log in and use it.

The main aims of the study are as follows:

KEYWORDS

health services; telehealth; telemedicine; health personnel; sleep apnea syndromes; mobile phone

https://humanfactors.jmir.org/2024/1/e47809
1. To investigate HCPs’ perspectives on the SA-DCP and its usefulness for their work
2. To determine whether the main targets of SA-DCP can be reached: shortening the initial guiding sessions of CPAP therapy, reducing patient calls and contact with HCP, and increasing patients’ adherence to CPAP therapy.
3. To examine patients’ perspectives on the SA-DCP and its usefulness.

Methods

Study Participants and Data Collection
The study population included HCPs at the OUH and patients who had started their CPAP therapy at the OUH. The patient population consisted of 2 groups. SA-DCP users were patients who had registered with the SA-DCP. SA-DCP nonusers referred to patients who had not registered with the SA-DCP.

Interviews of HCPs
HCPs of the OUH were contacted via email. Overall, 6 HCPs participated in the interviews from May to June 2021. Of these, 4 (67%) HCPs worked with patients, 1 (17%) was a supervisor, and 1 (17%) connected patients with sleep apnea to the SA-DCP and booked their appointments. The interviews were conducted remotely using a structured questionnaire. The HCPs provided voluntary informed consent for the interview by submitting a signed document. The interviews were then recorded and transcribed.

Survey for Patients With Sleep Apnea
The first part of material collection was conducted between May and August 2021. With the help of OUH HCPs, the survey, along with an invitation to participate and information about it, was sent to SA-DCP nonusers by mail. Respondents could send their responses by prepaid mail or electronically using Webropol Ltd’s Webropol survey tool. SA-DCP users were informed about the study through the SA-DCP. They provided their consent and answered the survey using the SA-DCP questionnaire.

The second part of material collection was conducted between January and June 2022. Both SA-DCP users and SA-DCP nonusers were informed about the study with the annual device delivery in an assistive equipment center (AEC). They could send their responses by prepaid mail or answer electronically using Webropol Ltd’s Webropol survey tool.

The patients’ survey included multiple choice questions, 5-item Likert-type questions (with choices ranging from strongly disagree to strongly agree), and open-ended questions. In total, 33 SA-DCP nonusers and 58 SA-DCP users responded to the survey.

Remote Monitoring Data of CPAP Devices
Information about patients’ adherence to CPAP therapy was collected from the remote monitoring data of CPAP devices. The HCP of OUH carried out the material collection manually in May 2021 in connection with 1-year controls of CPAP therapy. The collected information was anonymized and provided to the researchers. In total, CPAP remote monitoring data were collected from 90 SA-DCP nonusers and 80 SA-DCP users.

Registered Data of Phone Calls
The information about the number of patients’ phone calls per year to an AEC was collected from Aurora Innovation Ltd’s TeleQ program. The registered phone call data were collected during 2019, 2020, and 2021.

SA-DCP Customer Feedback Survey
Patients using the SA-DCP had the opportunity to provide customer feedback using the SA-DCP survey tool. The patients provided informed consent through the SA-DCP that their customer feedback could also be used for research purposes. The customer feedback did not contain any personal information. Feedback on the SA-DCP from 446 patients between February 18 and March 24, 2022, was included in this study.

Statistical Methods
Patients’ survey data were analyzed using SPSS software (version 28.0; IBM Corp). Descriptive statistics were applied to calculate the mean and SD for continuous data and frequency and percentage for categorical data. Baseline differences between the groups were explored using a 2-tailed independent sample t test for continuous variables and chi-square test for categorical variables. A P value <.05 was considered statistically significant for all analyses.

Qualitative Analysis
Qualitative methods were used in this study to analyze the open-ended questions in patient surveys and interviews with HCPs. The collected material was first analyzed using an inductive content analysis method to obtain a comprehensive understanding [34]. Initially, the HCPs’ and patients’ responses to the open-ended questions were open coded. Subsequently, the analyzed data were grouped into subcategories, and then similar findings were combined into the main categories to enable the final analysis. Finally, the textual data were analyzed using the quantification method [35].

Ethical Considerations
The study followed the guidelines of the Finnish Advisory Board on Research Integrity [36]. According to Finnish Law (488/1999), this study was exempted from review by the institutional review board (ethics committee of Northern Ostrobothnia Hospital District). The respondents were informed of the study. All participants voluntarily participated in the study and provided their informed consent. The results were processed such that no participants were identifiable in the results or quotations of this study. Sensitive personal information was not collected. The data were processed and stored in a secure environment according to the procedures of the University of Oulu.

Results

The Number of New CPAP Therapies, SA-DCP Users, and Phone Calls in the Years Studied
The number of new CPAP therapies in OUH between 2019 and 2021 is presented in Table 1. The percentage of SA-DCP users...
has increased annually, but there are still patients who do not use the SA-DCP (Table 1). The number of phone calls per year to an AEC is presented in Table 1.

Table 1. New continuous positive airway pressure (CPAP) therapies, patients attached to the sleep apnea digital care pathway (SA-DCP), phone calls to an assistive equipment center (AEC) per year, and percentage of SA-DCP users.

<table>
<thead>
<tr>
<th>Year</th>
<th>2019</th>
<th>2020</th>
<th>2021</th>
</tr>
</thead>
<tbody>
<tr>
<td>The number of new CPAP therapies</td>
<td>1172</td>
<td>1645</td>
<td>1160</td>
</tr>
<tr>
<td>The number of patients attached to the SA-DCP</td>
<td>292&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1645</td>
<td>1160</td>
</tr>
<tr>
<td>The number of SA-DCP users</td>
<td>130</td>
<td>1006</td>
<td>935</td>
</tr>
<tr>
<td>Percentage of SA-DCP users</td>
<td>44.5</td>
<td>61.2</td>
<td>80.6</td>
</tr>
<tr>
<td>The number of phone calls per year to an AEC</td>
<td>2784</td>
<td>4068</td>
<td>4020</td>
</tr>
</tbody>
</table>

<sup>a</sup>The SA-DCP was introduced from November 18, 2019, onward.

**HCPs’ Perspectives on the SA-DCP and its Usefulness for Their Work**

On the basis of the interviews with HCPs, the main themes, facilitators, and barriers related to using the SA-DCP are presented in Textbox 1. According to the interviewed HCPs, they were unable to identify significant changes in their workload and working practices following the introduction of the SA-DCP. Only one responder perceived that his workload had slightly increased because the SA-DCP did not support integration with electronic patient record; therefore, patient data had to be transferred manually from one program to another (Textbox 1). However, HCPs reported that in some situations, the SA-DCP brought more flexibility to their work practices regarding patient communication (Textbox 1). For example, it enabled them to respond to patients’ DCP messages during nonurgent work times, not only prereserved times. HCPs also reported that the initial guiding session of CPAP therapy went more smoothly for SA-DCP users who had familiarized themselves with the information about CPAP therapy through the SA-DCP (Textbox 1). The interviewed HCPs hoped that patients would make more use of the SA-DCP and its possibilities so its benefits would be better used.
Textbox 1. Themes and perceived barriers and facilitators regarding implementation of the sleep apnea digital care pathway (SA-DCP) according to health care professionals (HCPs).

**Use rate of SA-DCP**

- **Barriers**
  - SA-DCP’s use rate had been lower than HCPs assumed it would be.
  - Some patients still thought that the only proper contact was personal contact with HCPs.
  - Patients’ previous experiences with the need to log in to several digital health care services reduced their motivation to use them.

- **Facilitators**
  - Reminder text messages about logging into the SA-DCP have been sent to patients since June 2020.

**Initial guiding session of continuous positive airway pressure (CPAP) therapy**

- **Barriers**
  - Some SA-DCP users and nonusers still had not familiarized themselves with the prior information about CPAP therapy in advance.

- **Facilitators**
  - The guidance went more smoothly for SA-DCP users who familiarized themselves with the prior information about CPAP therapy.
  - From the patients’ perspective, the instructional videos available in the SA-DCP were perceived as useful and clear.

**Patients’ communication practices with HCP**

- **Barriers**
  - There were still a lot of phone calls.
  - HCPs also had to be reminded that they should not always call patients in connection with treatment controls but send a message via the SA-DCP.

- **Facilitators**
  - The SA-DCP gives patients more flexibility to contact HCPs regardless of time and place. For example, the patient may be in a location where they cannot answer the HCP’s phone call.
  - HCPs may instruct the patient during a phone call to watch SA-DCP’s educational video to get a better understanding of the matter.

**Patients’ adherence to CPAP therapy**

- **Barriers**
  - There was no clear indication that patients’ adherence to CPAP therapy was higher with the introduction of the SA-DCP.

- **Facilitators**
  - Reports obtained from CPAP devices had increased some patients’ adherence to CPAP therapy.

**Integration of SA-DCP into existing information and communications technology systems**

- **Barriers**
  - The SA-DCP had to be used in a different web browser than electronic patient record (EPR).
  - Remote monitoring of CPAP devices requires a separate program, and remote monitoring data cannot be viewed via the SA-DCP.
  - Attaching patients to the SA-DCP is laborious and must be done manually by copying patient information from the EPR.
  - Data had to be copied manually from SA-DCP’s messages into patients’ care plans.

**Workload and work practices of HCPs**

- **Barriers**
  - HCPs’ workloads did not change with the introduction of the SA-DCP.

- **Facilitators**
  -
The SA-DCP brought more flexibility to HCPs’ work practices regarding communication with patients. The SA-DCP was a good way to deliver the necessary contact information to patients and thereby instruct them to reserve time for the necessary procedures by themselves.

The professionals also brought up ideas for the development of SA-DCP. They hoped that SA-DCP’s integration with other information and communications technology (ICT) systems would be improved. One factor that caused a large workload for professionals was arranging appointment times for patients. The time reserved for the patient may not always suit him or her, necessitating a discussion about a more suitable time. If the patient could book appointments through the SA-DCP, it would greatly reduce the professionals’ working hours. Two respondents mentioned that in the future, an initial guiding session of CPAP therapy could also be carried out remotely, but this would require that patients for whom this would be suitable should be identified in advance. The professionals hoped that all surveys and measurements made by the patients related to their treatment would be available in an electronic format. The hope was also that the SA-DCP’s calendar would automatically remind patients, for example, to renew equipment, giving them more responsibility for managing their own affairs. One respondent wished that instructional videos could be directly linked to SA-DCP’s messages so that patients would not have to search for them.

The HCP interviewees perceived digital services in health care as a positive thing. According to them, the services should be easy to use, and the real end users of the services should be included in their development. One respondent believed that patients will use digital health care services more frequently in the future, but such systems are always initially met with resistance. The respondent mentioned that at first, patients in Finland were against e-prescribing and the Patient Data Repository of Kanta Services, but today, such services are commonplace, and people use them smoothly.

Comparison of Characteristics Between SA-DCP Users and SA-DCP Nonusers

According to the patients’ survey, there were no statistically significant differences in age, sex, and smoking status between SA-DCP users and nonusers (Table 2). According to the remote monitoring data of CPAP devices, SA-DCP nonusers were older than SA-DCP users (mean 59.1, SD 13.8 vs mean 55.3, SD 10.8; \( P < .049 \); Table 3). Compared with nonusers, SA-DCP users perceived their own abilities to use ICT to be better (mean 4.2, SD 0.8 vs mean 3.2, SD 1.2; \( P < .001 \)); they used computers, tablets, or smartphones more often (58/58, 100% vs 27/33 81%; overall \( P = .002 \)); and they were more accustomed to using electronic services (mean 4.8, SD 0.5 vs mean 4.1, SD 1.2; \( P = .006 \); Table 2). There was no statistically significant difference in how regularly SA-DCP users and nonusers used the electronic services (Table 2). SA-DCP users thought that communication about SA-DCP and how to log in had been clear, although SA-DCP nonusers thought that it had not (yes 52/58, 91% vs yes 7/33, 24%; overall \( P < .001 \); Table 2). Compared with SA-DCP users, SA-DCP nonusers preferred phone calls or physical appointments with HCPs to manage their health-related issues (Table 2). Neither SA-DCP users nor SA-DCP nonusers had any major concerns about the data security and protection of digital health care services (Table 2).
Table 2. Patient responses to the survey.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Use of SA-DCP&lt;sup&gt;a&lt;/sup&gt;</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nonusers (n=33)</td>
<td>Users (n=58)</td>
</tr>
<tr>
<td></td>
<td>61.9 (11.6)</td>
<td>57.3 (12.0)</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td></td>
<td>.08</td>
</tr>
<tr>
<td>Sex, n (%)</td>
<td></td>
<td>.07</td>
</tr>
<tr>
<td>Male</td>
<td>26 (79)</td>
<td>34 (59)</td>
</tr>
<tr>
<td>Female</td>
<td>7 (21)</td>
<td>24 (41)</td>
</tr>
<tr>
<td>Physical training frequency, n (%)</td>
<td></td>
<td>.03</td>
</tr>
<tr>
<td>Daily</td>
<td>16 (49)</td>
<td>15 (26)</td>
</tr>
<tr>
<td>Weekly</td>
<td>13 (39)</td>
<td>37 (64)</td>
</tr>
<tr>
<td>Monthly</td>
<td>1 (3)</td>
<td>5 (9)</td>
</tr>
<tr>
<td>Less than monthly</td>
<td>1 (3)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>No physical training</td>
<td>2 (3)</td>
<td>1 (2)</td>
</tr>
<tr>
<td>Smoking, n (%)</td>
<td></td>
<td>&gt;.99</td>
</tr>
<tr>
<td>Yes</td>
<td>2 (6)</td>
<td>4 (7)</td>
</tr>
<tr>
<td>No</td>
<td>31 (94)</td>
<td>53 (93)</td>
</tr>
<tr>
<td>Adherence to CPAP&lt;sup&gt;b&lt;/sup&gt; therapy (own assessment; Likert scale 1-5), mean (SD)</td>
<td>4.8 (0.5)</td>
<td>4.7 (0.8)</td>
</tr>
<tr>
<td>Patients familiar with CPAP therapy before the initial guiding session</td>
<td>16 (49)</td>
<td>43 (74)</td>
</tr>
<tr>
<td>The patient became familiar through SA-DCP, n (%)</td>
<td>N/A&lt;sup&gt;c&lt;/sup&gt;</td>
<td>29 (67)</td>
</tr>
<tr>
<td>Average use of the CPAP device per night (hours), mean (SD)</td>
<td>6.3 (1.0)</td>
<td>6.3 (1.3)</td>
</tr>
<tr>
<td>Has CPAP therapy helped the patient’s sleep apnea?, n (%)</td>
<td></td>
<td>&gt;.99</td>
</tr>
<tr>
<td>Yes</td>
<td>28 (85)</td>
<td>49 (85)</td>
</tr>
<tr>
<td>No</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Cannot say</td>
<td>5 (15)</td>
<td>9 (15)</td>
</tr>
<tr>
<td>Information and communication technology skills (own assessment; Likert scale 1-5), mean (SD)</td>
<td>3.2 (1.2)</td>
<td>4.2 (0.8)</td>
</tr>
<tr>
<td>Patient’s computer, tablet, or smartphone use, n (%)</td>
<td></td>
<td>.002</td>
</tr>
<tr>
<td>Regularly (weekly)</td>
<td>27 (82)</td>
<td>58 (100)</td>
</tr>
<tr>
<td>Randomly (less often than weekly)</td>
<td>4 (12)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>None</td>
<td>2 (6)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>How accustomed is the patient to using electronic services (eg, banking services, appointment services, etc; own assessment)? (Likert scale 1-5), mean (SD)</td>
<td>4.1 (1.2)</td>
<td>4.8 (0.5)</td>
</tr>
<tr>
<td>If the patient uses electronic services, how regularly?, n (%)</td>
<td></td>
<td>.55</td>
</tr>
<tr>
<td>Daily</td>
<td>20 (69)</td>
<td>44 (76)</td>
</tr>
<tr>
<td>Weekly</td>
<td>7 (24)</td>
<td>11 (19)</td>
</tr>
<tr>
<td>Monthly</td>
<td>1 (3)</td>
<td>3 (5)</td>
</tr>
<tr>
<td>Less often than monthly</td>
<td>1 (3)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Would the patient choose an electronic service or a phone call as a contact method regarding her or his treatment?, n (%)</td>
<td></td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Electronic service</td>
<td>13 (39)</td>
<td>48 (83)</td>
</tr>
<tr>
<td>Phone call</td>
<td>20 (61)</td>
<td>10 (17)</td>
</tr>
<tr>
<td>If the patient could choose either an electronic service (eg, remote consultation) or a physical appointment regarding her or his treatment, which method would she or he prefer?, n (%)</td>
<td></td>
<td>.048</td>
</tr>
<tr>
<td>Electronic service</td>
<td>10 (30)</td>
<td>31 (53)</td>
</tr>
<tr>
<td>Physical appointment</td>
<td>23 (70)</td>
<td>27 (46)</td>
</tr>
<tr>
<td>Characteristics</td>
<td>Use of SA-DCP&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Nonusers (n=33)</td>
</tr>
<tr>
<td>---------------------------------------------------------------------------------</td>
<td>---------------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Patient concerns about the data security and protection of digital health care services (Likert scale 1-5), mean (SD)</td>
<td></td>
<td>2.6 (1.2)</td>
</tr>
<tr>
<td>Has communication about SA-DCP and how to log in to it been sufficiently clear?, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>7 (24)</td>
<td>52 (91)</td>
</tr>
<tr>
<td>No</td>
<td>22 (76)</td>
<td>5 (9)</td>
</tr>
<tr>
<td>Did SA-DCP increase the patient’s adherence to CPAP therapy?, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>N/A</td>
<td>42 (72)</td>
</tr>
<tr>
<td>No</td>
<td>N/A</td>
<td>16 (28)</td>
</tr>
<tr>
<td>Did the patient contact HCP&lt;sup&gt;e&lt;/sup&gt; during her or his treatment period?, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Through SA-DCP</td>
<td>N/A</td>
<td>17 (52)</td>
</tr>
<tr>
<td>Through another contact method</td>
<td>17 (100)</td>
<td>16 (48)</td>
</tr>
<tr>
<td>The contact was related to (total), n</td>
<td>26</td>
<td>37</td>
</tr>
<tr>
<td>Treatment of sleep apnea, n (%)</td>
<td>8 (31)</td>
<td>3 (8)</td>
</tr>
<tr>
<td>CPAP therapy, n (%)</td>
<td>15 (58)</td>
<td>23 (62)</td>
</tr>
<tr>
<td>Other issues, n (%)</td>
<td>3 (12)</td>
<td>11 (30)</td>
</tr>
<tr>
<td>Did patients who contacted HCP get the help they needed?, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>16 (94)</td>
<td>31 (94)</td>
</tr>
<tr>
<td>Through SA-DCP messaging</td>
<td>N/A</td>
<td>16 (52)</td>
</tr>
<tr>
<td>Through another contact method</td>
<td>16 (100)</td>
<td>15 (48)</td>
</tr>
<tr>
<td>Did the patient need to find additional information about his or her treatment without contacting HCPs during the treatment period?, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The patient got the information she or he needed</td>
<td>8 (89)</td>
<td>21 (91)</td>
</tr>
<tr>
<td>Through SA-DCP</td>
<td>N/A</td>
<td>8 (38)</td>
</tr>
<tr>
<td>Through another source (internet, patient organizations, etc)</td>
<td>8 (100)</td>
<td>13 (62)</td>
</tr>
</tbody>
</table>

<sup>a</sup>SA-DCP: sleep apnea digital care pathway.

<sup>b</sup>CPAP: continuous positive airway pressure.

<sup>c</sup>N/A: not applicable.

<sup>d</sup>Not available.

<sup>e</sup>HCP: health care professional.
Table 3. Remote monitoring data of continuous positive airway pressure (CPAP) devices.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Use of SA-DCP&lt;sup&gt;a&lt;/sup&gt;</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nonusers (n=90)</td>
<td>Users (n=80)</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>59.1 (13.8)</td>
<td>55.3 (10.8)</td>
</tr>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>48 (53.3)</td>
<td>49 (61.3)</td>
</tr>
<tr>
<td>Female</td>
<td>42 (46.7)</td>
<td>31 (38.8)</td>
</tr>
<tr>
<td>AHI&lt;sup&gt;b&lt;/sup&gt; at diagnosis, mean (SD)</td>
<td>32.5 (18.0)</td>
<td>30.6 (18.9)</td>
</tr>
<tr>
<td>AHI residual in treatment, mean (SD)</td>
<td>2.4 (2.7)</td>
<td>2.1 (3.9)</td>
</tr>
<tr>
<td>Percentage of nights CPAP was used, mean (SD)</td>
<td>92.6 (12.7)</td>
<td>91.5 (18.8)</td>
</tr>
<tr>
<td>Hours of CPAP use per night, mean (SD)</td>
<td>6.2 (1.5)</td>
<td>6.1 (1.8)</td>
</tr>
<tr>
<td>CPAP device mask leak, mean (SD)</td>
<td>3.1 (6.1)</td>
<td>2.8 (3.2)</td>
</tr>
<tr>
<td>CPAP device median pressure, mean (SD)</td>
<td>8.4 (2.2)</td>
<td>7.8 (2.0)</td>
</tr>
</tbody>
</table>

<sup>a</sup>SA-DCP: sleep apnea digital care pathway.

<sup>b</sup>AHI: apnea-hypopnea index.

**Patients’ Rationales for Using or Not Using SA-DCP**

Patients were asked about their rationale for using or not using the SA-DCP (Table 4). SA-DCP users mostly adopted the SA-DCP because they thought that signing up for the SA-DCP was part of their treatment process (42/58, 72%). SA-DCP nonusers did not adopt the SA-DCP mainly because they were unaware of it (15/33, 46%).

Table 4. Patients’ rationales for using or not using the sleep apnea digital care pathway (SA-DCP).

<table>
<thead>
<tr>
<th>Patients’ rationales for using the SA-DCP (n=58)</th>
<th>Values, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I thought signing up for SA-DCP was part of my treatment process</td>
<td>42 (72)</td>
</tr>
<tr>
<td>It was recommended to me</td>
<td>39 (67)</td>
</tr>
<tr>
<td>It allows me to take care of my affairs regardless of time and place</td>
<td>31 (53)</td>
</tr>
<tr>
<td>I am very accustomed to using electronic services</td>
<td>25 (42)</td>
</tr>
<tr>
<td>I can more easily get information about sleep apnea and its treatment</td>
<td>18 (31)</td>
</tr>
<tr>
<td>I prefer to use electronic services for my treatment</td>
<td>15 (26)</td>
</tr>
<tr>
<td>I can more easily get information about CPAP therapy</td>
<td>15 (26)</td>
</tr>
<tr>
<td>I can take care of things related to my care more safely during the current COVID-19 period</td>
<td>12 (21)</td>
</tr>
<tr>
<td>By using SA-DCP, I am more committed to my treatment</td>
<td>8 (14)</td>
</tr>
<tr>
<td>Other reasons</td>
<td>2 (3)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patients’ rationales for not using the SA-DCP (n=33)</th>
<th>Values, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am not aware of SA-DCP</td>
<td>15 (46)</td>
</tr>
<tr>
<td>I prefer physical appointments</td>
<td>13 (39)</td>
</tr>
<tr>
<td>I prefer phone calls</td>
<td>10 (30)</td>
</tr>
<tr>
<td>I am aware of SA-DCP, but I forgot to log in</td>
<td>9 (27)</td>
</tr>
<tr>
<td>I do not know how to use electronic services</td>
<td>7 (21)</td>
</tr>
<tr>
<td>I do not want to use electronic services</td>
<td>6 (18)</td>
</tr>
<tr>
<td>The use of electronic services is generally difficult</td>
<td>6 (18)</td>
</tr>
<tr>
<td>I do not receive personal help through electronic services</td>
<td>5 (15)</td>
</tr>
<tr>
<td>Other reasons</td>
<td>4 (12)</td>
</tr>
<tr>
<td>I am concerned about the data security and protection of electronic services</td>
<td>3 (9)</td>
</tr>
<tr>
<td>My sleep apnea treatment and CPAP therapy are balanced, so I do not need to contact health care professionals through any communication channel</td>
<td>3 (9)</td>
</tr>
<tr>
<td>My sleep apnea treatment and CPAP therapy are balanced, so I do not need additional information through any communication channel</td>
<td>1 (3)</td>
</tr>
<tr>
<td>I do not feel the need to log into SA-DCP as part of my CPAP therapy</td>
<td>1 (3)</td>
</tr>
</tbody>
</table>

aCPAP: continuous positive airway pressure.

Patients’ Prefamiliarization With CPAP Therapy Before the Initial Guiding Session

A larger proportion of SA-DCP users had familiarized themselves with prior information about CPAP therapy before the initial guiding session of CPAP therapy than SA-DCP nonusers (43/58, 74% vs 16/33, 49%; \( P = .02 \); Table 2). Among the 48 SA-DCP users who familiarized themselves with information about CPAP therapy beforehand, 29 (67%) performed it through the SA-DCP (Table 2). Most SA-DCP nonusers (6/16, 38%) said they had received the preliminary information from a spouse or a relative who had already used a CPAP device. The other sources of information for both groups were the internet (6/59, 10%), private health care providers (3/59, 5%), primary health care units (2/59, 3%), and the Duodecim medical information database (2/59, 3%). SA-DCP users also received information from occupational health care units (2/43, 5%) and the Facebook sleep apnea support group (2/43, 5%). Correspondingly, SA-DCP nonusers received information from specialized care units (2/16, 13%), research articles (1/16, 6%), and AEC (1/16, 6%).

The initial guidance sessions of CPAP therapy were carried out with small groups of patients (4-8 patients at a time). According to HCPs, the initial guiding sessions were smoother for patients who had already familiarized themselves with prior information about CPAP therapy through the SA-DCP (Textbox 1). The problem was that many patients still did not have prior information about CPAP therapy; therefore, most of the initial guiding sessions had to be implemented according to their needs. According to the HCPs, patients found the instructional videos available in the SA-DCP to be useful and clear (Textbox 1). Patients were also instructed to familiarize themselves with them and other information material found on the SA-DCP even after the sessions if they had further questions.
Patients’ Information Needs About Sleep Apnea and CPAP Therapy

SA-DCP includes electronic messaging functionality between patients and HCPs and information about sleep apnea, its treatment, and CPAP therapy. According to the survey responses of SA-DCP users, most patients looked for information about the SA-DCP, sleep apnea, self-treatment of sleep apnea, and cleaning and maintenance of the CPAP device (Table 5). The messaging functionality of the SA-DCP and its “frequently asked questions” function were not widely used; only 38% (22/58) of SA-DCP users used them (Table 5).

Table 5. Functionalities of the sleep apnea digital care pathway (SA-DCP) used by patients according to the SA-DCP users survey (N=58).

<table>
<thead>
<tr>
<th>The number of patients who familiarized themselves with the following information materials</th>
<th>Values, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Welcome to SA-DCP</td>
<td>55 (95)</td>
</tr>
<tr>
<td>Sleep apnea</td>
<td>45 (78)</td>
</tr>
<tr>
<td>Cleaning and maintenance of the CPAP(^d) device</td>
<td>43 (74)</td>
</tr>
<tr>
<td>Self-treatment of sleep apnea</td>
<td>42 (72)</td>
</tr>
<tr>
<td>CPAP therapy</td>
<td>39 (67)</td>
</tr>
<tr>
<td>Preparing for the CPAP therapy initial guiding session</td>
<td>38 (67)</td>
</tr>
<tr>
<td>Sleep apnea and driving ability</td>
<td>32 (55)</td>
</tr>
<tr>
<td>CPAP therapy in unusual everyday situations</td>
<td>28 (48)</td>
</tr>
<tr>
<td>Controls, rehabilitation, and social security</td>
<td>25 (43)</td>
</tr>
<tr>
<td>Frequently asked questions</td>
<td>22 (38)</td>
</tr>
<tr>
<td>The patient has communicated with the HCP(^b) in matters related to his or her treatment through the messaging functionality of SA-DCP</td>
<td>22 (38)</td>
</tr>
</tbody>
</table>

\(^{d}\)CPAP: continuous positive airway pressure.  
\(^{b}\)HCP: health care professional.

During the treatment period, both SA-DCP users and SA-DCP nonusers sought additional information regarding their treatment without contacting HCPs (23/58, 39% vs 9/33, 27%; \(P=.48\); Table 2). Among 23 SA-DCP users who sought more information, 8 (35%) performed it through the SA-DCP. The other reported information sources for SA-DCP users were the internet (9/23, 39%), Facebook sleep apnea support group (1/23, 4%), patient organizations (1/23, 4%), rehabilitation (1/23, 4%), and professional education (1/23, 4%). SA-DCP nonusers received additional information from the following sources: the internet (5/9, 56%), Facebook sleep apnea support group (1/9, 11%), scientific articles (1/9, 11%), and information material provided by private health care services providers (1/9, 11%). Most SA-DCP nonusers and SA-DCP users who sought more information about CPAP therapy and sleep apnea received the information they needed (8/9, 89% vs 21/23, 91%; \(P>.99\); Table 2).

Patients’ Communication Practices With HCPs

During the treatment period, 52% (17/33) of SA-DCP nonusers and 57% (33/58) of SA-DCP users contacted HCPs (Table 2). Among 33 SA-DCP users who contacted HCPs, 17 (52%) used the SA-DCP, and the rest used other contact methods (Table 2). SA-DCP users who preferred contact methods other than SA-DCP messages were older (mean 63.3, SD 9.4 vs mean 55.7, SD 10.8; \(P<.04\)). Phone calls were the most important form of contact for SA-DCP users (10/33, 30%). SA-DCP nonusers (9/17, 53%) mostly contacted HCP via phone calls. The next most common contact method for both groups was a physical visit to an AEC or a primary health care unit. The contact mostly concerned CPAP therapy; this was the case for 88% (15/17) of SA-DCP nonusers and 70% (23/33) of SA-DCP users (Table 2). Most SA-DCP nonusers and SA-DCP users who contacted HCPs received the help they needed (16/17, 94% vs 31/33, 94%; \(P=.41\); Table 2).

According to the phone call register data, the annual number of phone calls to an AEC was still high even after the introduction of SA-DCP (Table 1). An exact comparison of phone calls to AECs per patient between different years could not be made because the number of new CPAP therapies in the OUH varied between different years, and the number of annual phone calls also showed contacts with HCPs from patients whose CPAP therapies had started in previous years (Table 1). The results of 2019 mainly represent a situation in which the SA-DCP was not yet in use at OUH because it was introduced at the very end of 2019. The results of 2021 represent a situation in which the SA-DCP had been in use at OUH for approximately 2 years. HCPs also indicated the same; there was no significant decrease in the number of phone calls, and there were still many phone calls related to CPAP therapy (Textbox 1). HCPs emphasized that they try to guide patients during phone calls to use the SA-DCP more in matters related to their care. Although the patients’ affairs were handled mostly with phone calls, the HCPs thought the instructional videos and informational materials included in the SA-DCP were valuable. It was possible to better explain things to patients with them (Textbox 1). For example, HCPs may instruct the patient during a phone call to watch SA-DCP’s educational video to get a better understanding of the matter. The HCPs also emphasized that the SA-DCP is a
good way to deliver the necessary contact information to patients, allowing them to reserve time for the necessary procedures themselves (Textbox 1).

**Patient Adherence to CPAP Therapy**

According to the patients’ responses to the survey and remote monitoring data of CPAP devices, adherence to CPAP therapy was high in both groups (Tables 2 and 3). Both groups used the CPAP device on average for >6 hours per night and on >90% of nights (Tables 2 and 3). On the basis of the patients’ own assessments, adherence to CPAP therapy was high in both SA-DCP nonusers (mean 4.8, SD 0.5) and SA-DCP users (mean 4.7, SD 0.8; Table 2). In addition, according to the patients’ survey, 72% (42/58) of SA-DCP users reported that SA-DCP had made them more motivated to perform their own CPAP therapy (Table 2). Most patients in both groups believed that CPAP therapy helped them treat sleep apnea (Table 2). The remote monitoring data of CPAP devices showed that CPAP therapy had significantly reduced the number of AHIs for both groups (Table 3).

**Patient Feedback About SA-DCP**

A total of 446 patients responded to the customer feedback survey; their feedback is shown in Figure 1. Patient feedback on the SA-DCP was generally positive; most of them agreed or strongly agreed with the survey claims (Figure 1). When examining the results, it should be noted that the questions of the patient feedback survey are common to every DCP in the OUH. As the functionalities offered by DCPs vary according to the care chains of different diseases, not all the questions are necessarily valid for every DCP. For example, examinations are not offered through the SA-DCP.

![Figure 1. Patient feedback about the sleep apnea digital care pathway.](https://humanfactors.jmir.org/2024/1/e47809)
Discussion

Principal Findings

This study investigated whether the 3 main goals for introducing the SA-DCP at OUH were achieved. The first aim of introducing the SA-DCP was to shorten the initial guiding sessions of CPAP therapy on the assumption that the patients would have familiarized themselves with prior information about CPAP therapy in advance through the SA-DCP. The second main aim was to reduce the number of patients’ phone calls and contacts to HCPs, especially when the information can be found in the SA-DCP. The primary goal of implementing SA-DCP at OUH was to improve patients’ adherence to CPAP therapy. However, according to the results of this study, not all the objectives of introducing the SA-DCP were achieved.

On the basis of the HCP’s responses to this study, shortening the initial guiding sessions of CPAP therapy had not been fully achieved, although a significantly larger number of SA-DCP users had familiarized themselves with prior information about CPAP therapy compared with SA-DCP nonusers. In this regard, it can be said that SA-DCP has contributed to the better preparation of patients for sessions. The initial guiding sessions were smoother for patients who had already familiarized themselves with prior information regarding CPAP therapy through the SA-DCP. However, many patients still did not have prior information about CPAP therapy; therefore, most sessions had to be implemented according to their needs. Because digital services may require care process changes to get the most out of them, 2 HCPs mentioned that the initial guidance sessions could also be carried out remotely in the future; however, this would require that the patients for whom this procedure would be suitable should be identified in advance [11,12].

Despite previous studies showing that DCPs would make it possible to reduce the number of patient phone calls to HCPs, this did not happen in the case of SA-DCP [37,38]. The annual number of phone calls to an AEC was still high even after the introduction of SA-DCP, according to the phone call register data. As the number of patients’ phone calls related to CPAP therapy was still high, HCPs mentioned that it was difficult to assess the actual change in the number of phone calls. However, they perceived that the number of patient calls did not decrease significantly. Previous studies have shown that patients’ ability to use electronic services also promotes the use of digital health care services [39,40]. However, Jenssen et al [41] found that despite the regular use of new digital technologies and services such as electronic banking, few of their study participants supported using these tools for communicating with their HCPs. The same behavior pattern can also be observed in the case of SA-DCP. Although SA-DCP users in this study perceived their ability to use ICT to be good and used computers, tablets, or smartphones regularly and were accustomed to using electronic services, only approximately half of them contacted the HCP with SA-DCP messages when needed. Among SA-DCP users, phone calls were the most important other contact method. The notable finding was that SA-DCP users who preferred another contact method were older.

Patient concerns about data security and protection have weakened their willingness to use electronic communication methods in health care [15,42]. On the basis of this study, this would not be an explanatory factor for the low use of SA-DCP messages, as both SA-DCP users and SA-DCP nonusers were not significantly concerned about the data security and protection of digital health care services. Zanaboni and Fagerlund [43] discovered that communicating via electronic tools was less time-consuming from the patient’s perspective than communicating via phone calls. However, some participants indicated that the time elapsed to receive a response from the HCP was more important than the time spent using the service itself. Long response times have been seen as one of the most important reasons for patients’ dissatisfaction with electronic communication in health care [39,44]. In a Norwegian study, older patients hoped that their electronic messages would be answered the next day at the latest; otherwise, they experienced dissatisfaction with the service [39]. From the patients’ point of view, they may perceive that a phone call is a quick and convenient way to handle their health-related matters [45-47]. The fundamental difference is that a phone call involves real-time interactive communication, whereas SA-DCP messages can be defined as asynchronous communication [48]. The patient may ask follow-up questions during the phone call and the HCP can answer them immediately. When using electronic communication tools, there may be delays in answers to questions and possible follow-up questions because of asynchronous communication, as the patient and the HCP may not be dealing with the issue simultaneously [48].

One of the main goals of introducing the SA-DCP was to improve patients’ adherence to CPAP therapy. This study showed no statistical difference between SA-DCP users’ and nonusers’ adherence to CPAP therapy. Adherence to CPAP therapy was high in both groups according to the patients’ own estimates and remote monitoring data of CPAP devices. Both groups performed CPAP therapy regularly and reported that it helped them to treat their sleep apnea. In addition, 72% (42/58) of SA-DCP users reported that SA-DCP motivated them to perform their own CPAP therapy. Unfortunately, this study did not ask why the participants felt this. The role of the SA-DCP was to complement CPAP therapy by providing information and an electronic communication channel. It did not include clear mechanisms for influencing patients’ behavior related to their own health as digital health interventions typically do, for example, in relation to weight management [49-51]. The CPAP therapy clearly helped the participants in this study to reduce the number of AHIs. Presumably, the biggest motivation for performing CPAP therapy came from alleviating sleep apnea symptoms and not so much from using the SA-DCP; therefore, the SA-DCP was not a significant factor in explaining adherence to CPAP therapy.

This study investigated HCPs’ perspectives on the SA-DCP and its usefulness for their work. Although previous studies determined that DCPs could potentially free health care services capacity for other purposes and reduce the workload of HCPs, the results of this study do not support these results in the case of SA-DCP [27,28]. The HCPs who participated in the study were unable to define significant changes in their workload and
work practices after the introduction of SA-DCP. The primary aim of HCPs was for patients to use the SA-DCP more so that its benefits could be better used. Previous studies have highlighted that DCPs can promote work flexibility, for example, by enabling HCPs to respond to patients’ DCP messages at nonurgent, not only reserved times [27,31]. The responses of HCPs in this study pointed out the same. With the help of the DCP, patients can access the information it contains before and after contact with HCPs, thus reducing patient follow-up questions [12]. From this perspective, HCPs felt that educational videos and information materials on SA-DCP were beneficial because, through them, the patients could better understand things. From a technical point of view, the SA-DCP’s weak integration with existing ICT systems was seen as one of its key shortcomings and an area for future development. The lack of interoperability with existing ICT systems has been found to weaken the willingness of HCPs to use digital health care services and increase their workload [10,52]. According to the interviewed HCPs, lack of integration reduced the fluency of their work, increased the workload of one responder, and can cause risks from the perspective of information protection and patient safety when patient information is copied manually between different programs.

Digital health care services are intended to help patients become more active actors, more adherent to their own care, and change their behavior in a more favorable direction for their health [49-51,53,54]. Promising results have already been achieved, for example, in treating obesity with the help of digital services [30,51]. In the case of SA-DCP, it was hoped that patients would be active and familiarize themselves with the information contained in it about sleep apnea and CPAP therapy. According to the patient survey, most SA-DCP users have done so. Although most SA-DCP users familiarized themselves with the information in SA-DCP, there was no statistically significant difference in the proportion of SA-DCP users and nonusers who sought additional information about their illness or CPAP therapy. From this perspective, it cannot be said that SA-DCP users are more active actors. It has been established that digital health care services can lower the threshold for patients to contact HCPs [37,54,55]. According to this study, there was no statistically significant difference between the percentage of SA-DCP users and SA-DCP nonusers who contacted HCPs during their treatment period. However, this study did not ask how often the patients contacted the HCPs. On the basis of the results of this study, it seems that patients sought additional information about their illness or contacted HCPs when they had a real need, regardless of the information source or communication method.

Most SA-DCP users thought that the treatment they received through the SA-DCP was good; it was fine technically, a safe service, and the information it contained was clear and understandable. However, some patients still did not use SA-DCP, although the relative number of active SA-DCP users increased during the study period. Lack of digital literacy is one of the barriers to promoting the use of digital health care services. Older adults, in particular, tend to have lower digital literacy than the general population [39,56]. Mannheim et al [40] emphasized in their study that older adults are not a homogeneous group in terms of digital literacy and should also be better included when designing digital health care services [40]. On the basis of the patient survey, there was no statistically significant difference in the age of SA-DCP users and SA-DCP nonusers, but based on remote monitoring data from CPAP devices, SA-DCP nonusers were older. According to this study, SA-DCP nonusers perceived their abilities to use ICT to be worse; they used computers, tablets, or smartphones more rarely and were less accustomed to using electronic services than SA-DCP users. SA-DCP nonusers preferred phone calls or physical appointments to manage their health-related issues with HCPs. The results showed a statistically significant difference in how clearly the patients perceived the communication about SA-DCP. Only 24% (7/38) of SA-DCP nonusers considered communication to be clear, and ignorance of the SA-DCP was the most common reason for them not to use the SA-DCP. After the diagnosis of sleep apnea, the patients received an information letter containing information about the disease and its treatment. This letter also included information on the SA-DCP and how to use it. Did SA-DCP nonusers think the SA-DCP was not adequately explained because they did not want to use digital health care services in the first place and preferred to conduct their health-related issues through phone calls or physical visits? They may not have paid attention to the SA-DCP information letter if they do not typically use or are not willing to use digital health care services or if they perceive they have weak skills in using them.

One of the key findings of this study is that the nonuse of SA-DCP and its functionalities among patients with sleep apnea means that its full potential is not being used. This can be seen, for example, in the initial guiding sessions of CPAP therapy, when some patients still come without prior knowledge. Although the number of SA-DCP users increased during the years covered by this study, not all SA-DCP functionalities were significantly used. In particular, this was reflected in the fact that SA-DCP messages were not widely used; therefore, the number of calls to AECs was not reduced. This study found that lower digital literacy and older age were significant factors in explaining the nonuse of the SA-DCP. Older SA-DCP users more often favored other contact methods, such as phone calls, when contacting HCPs during their treatment period. In the future, special attention should be paid to how digital health care services are designed according to the needs of older adults with weak digital literacy. Care processes should be better adapted to the requirements of digital health care services. Clearly, only the traditional information letter about SA-DCP is not sufficient to encourage all patients to adopt it. If there are challenges in deployment, patients could be more actively encouraged to adopt the SA-DCP and offered support. Previous studies have highlighted that the desire of older adults to use digital health care services can be supported by offering guidance and peer support [39,56]. Studies have also emphasized that both professionals and patients should be closely involved in DCP development to obtain the best benefit and that development should be a continuous process [10,29]. With age, various functional limitations, such as diminished eyesight related to diabetes or deteriorated motor skills owing to rheumatism, can increase and thus make it more difficult to use digital services [57,58]. Therefore, special attention should be
paid to the usability and accessibility of digital services. The real end users should be involved in the design process, as the interviewed HCPs highlighted [22,25,58].

Limitations

Our study had some limitations. Patients with sleep apnea give up CPAP therapy for different reasons, which can bias this study’s data regarding patients’ adherence to CPAP therapy. Most patients who responded to the survey had continued CPAP therapy for ≥1 year, and remote monitoring data on CPAP devices were collected in connection with 1-year control. Unfortunately, when the study was carried out, no information was available on the proportion of SA-DCP users and SA-DCP nonusers who had discontinued CPAP therapy. This would have provided additional information about patients’ adherence to CPAP therapy. Previous results have highlighted that high attrition rates hinder achieving the full benefits of digital health care services. During the implementation of the study, the SA-DCP did not enable the automatic collection of log data on the activity of patients using the SA-DCP, but through the automatic log data, it was only possible to determine that the patient had used the SA-DCP. Therefore, this study did not examine patients’ adherence to SA-DCP use, but only whether they had used the service.

On the basis of the study’s results, approximately half of SA-DCP users still contacted HCPs in a way other than through SA-DCP messages, although they reported having good digital literacy. Most SA-DCP nonusers also preferred phone calls to contact HCPs. However, in this study, SA-DCP users and SA-DCP nonusers were not asked why some preferred phone calls to contact HCPs instead of electronic messaging. Future research is needed to better understand this behavior pattern. This study did not ask patients how many times they contacted HCPs; it only investigated whether the patients contacted HCPs during their treatment period. Information on the number of contacts would have provided valuable information on whether using the SA-DCP can lower the threshold for contacting HCPs.

One of the goals of the SA-DCP was to increase patients’ adherence to CPAP therapy, and most SA-DCP users felt this was the case. Although the results of the survey and the remote monitoring data of the CPAP devices showed that there was no statistically significant difference in adherence to CPAP therapy between the groups, it would have been beneficial to ask SA-DCP users why most of them felt that the SA-DCP had increased their adherence to CPAP therapy. However, this was not investigated in this study. The sample size of the interviewed HCPs was small in this study. However, the answers to the HCPs were mostly consistent. Most of them thought there were no significant changes to their workload and work practices; there were still many phone calls from patients. At the time of writing, the SA-DCP did not enable the automatic collection of log data about the number of electronic messages. If this information had been available, it would have enabled a better comparison between the volumes of phone calls and SA-DCP messages.

Conclusions

According to this study, not all the goals set for introducing the SA-DCP have been achieved. The HCPs who participated in the study could not define significant changes in their workload and work practices after the introduction of SA-DCP. The SA-DCP has brought more flexibility to HCPs’ work practices regarding patient communication. Despite using SA-DPC, some patients still wanted to communicate with HCPs by phone. Adherence to CPAP therapy was high in both SA-DCP users and nonusers. Patients’ lower digital literacy and older age were the most significant factors explaining the nonuse of the SA-DCP. In the future, more attention should be paid to how these user groups should be considered in the design and introduction of the DCPs.

Acknowledgments

This research was funded by the Strategic Research Council (336278 and 358417). The authors are grateful to the study participants who generously gave their time to share their views on the sleep apnea digital care pathway. The authors would also like to thank the health care professionals from Oulu University Hospital who helped in the practical implementation of the study.

Authors’ Contributions

All authors participated sufficiently in the work to take public responsibility for the appropriate portions of the content. JH, TH, and JR were responsible for study conception and design. JH, HM, PL, and TH performed data acquisition. JH analyzed and interpreted the data. JH, TH, MT, PL, and JR drafted the manuscript.

Conflicts of Interest

PL and HM have been involved in the national and regional development of the digital care pathways of Health Village. All other authors declare no other conflicts of interest.

References


Abbreviations

AEC: assistive equipment center
AHI: apnea-hypopnea index
CPAP: continuous positive airway pressure
DCP: digital care pathway
HCP: health care professional
ICT: information and communications technology
OUH: Oulu University Hospital
SA-DCP: sleep apnea digital care pathway

Edited by A Kushniruk; submitted 02.04.23; peer-reviewed by S Kapila, J Roland, Y Sapanel; comments to author 23.07.23; revised version received 31.08.23; accepted 20.01.24; published 22.02.24.

Please cite as:
Haverinen J, Harju T, Mikkonen H, Liljamo P, Turpeinen M, Reponen J
Digital Care Pathway for Patients With Sleep Apnea in Specialized Care: Mixed Methods Study
JMIR Hum Factors 2024;11:e47809
URL: https://humanfactors.jmir.org/2024/1/e47809
doi:10.2196/47809
PMID:38386368
Leveraging Generative AI Tools to Support the Development of Digital Solutions in Health Care Research: Case Study

Danissa V Rodriguez¹, PhD; Katharine Lawrence¹,², MPH, MD; Javier Gonzalez², BS; Beatrix Brandfield-Harvey¹, BA; Lynn Xu¹, MPH; Sumaiya Tasneem¹, MPH; Deane L Levine¹, MPH; Devin Mann¹,², MS, MD

¹Department of Population Health, New York University Grossman School of Medicine, New York, NY, United States
²Medical Center Information Technology, Department of Health Informatics, New York University Langone Health, New York, NY, United States

Abstract

Background: Generative artificial intelligence has the potential to revolutionize health technology product development by improving coding quality, efficiency, documentation, quality assessment and review, and troubleshooting.

Objective: This paper explores the application of a commercially available generative artificial intelligence tool (ChatGPT) to the development of a digital health behavior change intervention designed to support patient engagement in a commercial digital diabetes prevention program.

Methods: We examined the capacity, advantages, and limitations of ChatGPT to support digital product idea conceptualization, intervention content development, and the software engineering process, including software requirement generation, software design, and code production. In total, 11 evaluators, each with at least 10 years of experience in fields of study ranging from medicine and implementation science to computer science, participated in the output review process (ChatGPT vs human-generated output). All had familiarity or prior exposure to the original personalized automatic messaging system intervention. The evaluators rated the ChatGPT-produced outputs in terms of understandability, usability, novelty, relevance, completeness, and efficiency.

Results: Most metrics received positive scores. We identified that ChatGPT can (1) support developers to achieve high-quality products faster and (2) facilitate nontechnical communication and system understanding between technical and nontechnical team members around the development goal of rapid and easy-to-build computational solutions for medical technologies.

Conclusions: ChatGPT can serve as a usable facilitator for researchers engaging in the software development life cycle, from product conceptualization to feature identification and user story development to code generation.

Trial Registration: ClinicalTrials.gov NCT04049500; https://clinicaltrials.gov/ct2/show/NCT04049500

(JMIR Hum Factors 2024;11:e52885) doi:10.2196/52885

KEYWORDS
digital health; GenAI; generative; artificial intelligence; ChatGPT; software engineering; mHealth; mobile health; app; apps; application; applications; diabetes; diabetic; diabetes prevention; digital prescription; software; engagement; behaviour change; behavior change; developer; developers; LLM; LLMs; language model; language models; NLP; natural language processing

Introduction

Health care has undergone a digital transformation, resulting in a growing reliance on software engineering for medical use cases, including health care research. However, little guidance exists for health researchers on how to effectively develop digital health interventions [1]; in particular, software development challenges that include expertise gaps in coding, custom development needs, high costs, and time constraints result in
multilevel barriers to designing and deploying a usable, scalable, and sustainable digital health product [1].

Generative artificial intelligence (GenAI) technologies such as ChatGPT can potentially support researchers in health technology endeavors by providing foundational frameworks and processes for the software development life cycle [2]. These systems can help reduce time and enhance precision for technology-based research projects by supporting both nonprogrammers and experienced programmers in code development, troubleshooting, and cleaning [2]. Moreover, the ability to use GenAI to generate content from different perspectives (expert or nonexpert) can facilitate and improve communication between technical and nontechnical team members of multidisciplinary teams. For example, a nontechnical team member can write their ideas in natural text and then use GenAI to request assistance in creating discussion points to communicate to a technical team audience. GenAI tools may also help health technology researchers refine research questions, identify appropriate theoretical frameworks and models, and leverage popular implementation strategies such as design thinking to build effective, theory-grounded, and evidence-based digital health interventions. ChatGPT (OpenAI, Microsoft Corporation) has already demonstrated feasibility as a support tool for clinical decision support development in health care [3], and more broadly as a coding copilot in programming and engineering [4,5].

This study explores the use of ChatGPT to recreate a personalized automatic messaging system (PAMS), which was developed as part of a digital health research initiative to support patient engagement with a commercial digital diabetes prevention program (dDPP). We examine the capacity, advantages, and limitations of ChatGPT to support product ideation and conceptualization, intervention content development, and the software engineering process including software requirement generation, software design, and code production. This paper provides insights to support the GenAI-assisted development of computational tools that are usable, reliable, extensible, and in line with the standards of modern coding practices. The framework includes prompts for both the intervention conceptualization as well as the main phases of the software development process.

**Methods**

**Settings and Intervention Development Context**

In previous work [6], we described the development of PAMS, a novel integrated multicomponent communications platform, to promote patient-provider communication and patient engagement in a commercial dDPP (Noom; Noom, Inc). The PAMS intervention included early prototyping and user testing, a technical development phase, and a randomized controlled trial. The core content and user experience features of PAMS were identified, prototyped, and evaluated using the well-established design thinking “discover, define, design, and test” approach to iteratively gather information, define, design, and refine the engagement intervention [7]. Stakeholders included: patients with prediabetes and their support network (eg, caregivers and partners), primary care providers, health technologists, programmers and computer scientists, behavioral change theorists and subject matter experts, the research administrative team, and dDPP product developers and coaches. The main components of this PAMS intervention include (1) a theory-driven behavior change message library, (2) a personalized automated message system delivery platform (SMS text messaging–based), and (3) EHR-integrated data visualizations. The PAMS messaging library uses an integrated framework that combines established theoretical models for behavior change with human-centered design strategies to maximize the evidence-based conditions for behavior change and the user acceptance and use of a digital health product. The technical development of PAMS followed an agile software development approach based on incremental 2-week sprint cycles consisting of requirement planning, design, development, and testing of a specific set of functional features. In this paper, we will recreate this development process using GenAI (ChatGPT).

**ChatGPT-PAMS Experiment Design**

To evaluate the effectiveness of using GenAI to support the development of digital tools in medical settings, our experiment is based on recreating PAMS using GenAI (ChatGPT) and evaluating human-generated vs ChatGPT-generated documentation. To accurately capture the ideation and development process, our multidisciplinary team reviewed all documentation and processes used in the early stages of PAMS conceptualization, including supporting theoretical models, content and features, and technical development. We then recreated these processes via a series of prompts for ChatGPT-4 to assist with the generation of theory, content, user stories, requirement documents, design diagrams, and the code for a subset of the requirements. Outputs from ChatGPT were reviewed and compared to human-generated documentation by 11 evaluating team members. Evaluators consisted of clinicians, behavioral scientists, programmers, and research staff working in digital health and technology for behavior change research. Collectively, they represent more than 50 years of clinical, research, design, and computer science experience. The evaluators independently rated the quality of various aspects of information provided by ChatGPT on a Likert scale, where higher ratings indicated greater quality of information (1: very poor; 2: poor; 3: acceptable; 4: good; 5: very good; N/A: not applicable). Aspects of evaluation included: understandability (Does this output make sense given the context of the study and prompts?), novelty (Were new ideas generated?) [3], usability (Does this create a usable output?), relevance (Does this create a useful output?), efficiency (Would having these outputs have saved time?), and potential for bias (What unintended consequences might arise from these outputs?) [6]. Evaluators were also asked to give an overall score on the quality of the ChatGPT output (Overall, how good would you say this output is?). Post review, a group debrief was conducted, using a semistructured interview guide to facilitate discussion regarding perceptions of outputs and rationale for ratings.

**Ethical Considerations**

Ethical considerations helped guide the initial development of research methods and reduce potential risks for participants in...
the original study implementation with the PAMS intervention [7]. Recreating the technical development of a system previously built as part of the dDPP randomized controlled trial (NCT04049500) has not introduced any new risks to patients. Patients were not involved in this research examining the use of GenAI in the development of digital health care solutions. No patient data was used in the prompt generation phase.

Regarding ethical considerations for the methods used in this paper, as an attempt to mitigate evaluator biases, we worked with a diverse team of evaluators who were aware of the initial study but were not necessarily involved in the technical development. Additionally, we understand the limitations and concerns of the use of ChatGPT including possible hallucinations and incorrect answers. Thus, we emphasize the need for human expertise to identify correct and incorrect outputs and have flagged this as a study consideration. When developing the GenAI-based solution, we used the same considerations for data security, patient usability, accessibility, and data privacy used in the original human-developed solution.

**Prompt Generation Framework**

Prompt engineering focuses on the skill of designing and creating effective prompts that guide ChatGPT to produce the best possible output for your task. We followed existing literature [8-11] combined with our expertise and experimentation to provide a framework that yields the best result when developing a digital solution like PAMS (Figure 1).

**Results**

**PAMS Concept and User Experience Generation**

**Overview**

Core components of the PAMS intervention were conceptualized and designed via an underlying behavior change theory, design principles and personas, and a message content library.

**Underpinning Behavior Change Theory and Approach**

**Human-Generated Solution**

Leveraging behavior change literature review and interviews with behavior change theory content experts (n=4), the research team initially identified ten unique behavior change theories and six process models that were considered to be an appropriate fit for the aims of the overall intervention. A unique model was developed that captured (1) the relevant underlying behavior change theory, (2) implementation strategies, and (3) unique contexts of the technology environment (Figure 2A).

As a researcher looking to enhance engagement and communication in a digital diabetes prevention program application, there are several well-accepted behavior change theories and frameworks that can inform your intervention development. Here are some prominent areas:

1. **Transtheoretical Model (TTM) or Stages of Change Model** This model proposes that individuals go through a series of stages when modifying their behavior. These stages include pre-contemplation, contemplation, preparation, action, and maintenance. Tailoring interventions based on the individual’s stage of change can be effective in promoting engagement and behavior change.

2. **Social Cognitive Theory (SCT)** This theory emphasizes the reciprocal interaction between personal factors, environmental factors, and behavior. It suggests that self-efficacy (confidence in one’s ability to perform a behavior) and observational learning (learning through observing others) play crucial roles in behavior change. Designing interventions that enhance self-efficacy and provide opportunities for observational learning can enhance engagement and communication.

3. **Self-Determination Theory (SDT)** SDT focuses on intrinsic motivation and the fulfillment of psychological needs, including autonomy, competence, and relatedness. Building interventions that support autonomy by providing choices, fostering a sense of competence, and creating a supportive social environment can promote engagement and communication.

4. **Health Belief Model (HBM)** HBM suggests that behavior change is influenced by perceived susceptibility to a health problem, perceived severity of the problem, perceived benefits of taking action, perceived barriers to change, and cue to action. Interventions can be designed to address these factors and enhance engagement and communication.

5. **Motivational Interviewing (MI)** MI is a client-centered approach that aims to enhance intrinsic motivation and commitment to change by exploring and resolving ambivalence. Using MI techniques within the dDPP application can facilitate engagement and communication between patients and clinicians.

6. **Cognitive-Behavioral Therapy (CBT)** CBT focuses on identifying and modifying maladaptive thoughts, feelings, and behaviors. Incorporating CBT principles, such as cognitive restructuring and behavior activation, into the dDPP application can improve engagement and communication.

7. **The COM-B Model** This model suggests that behavior change is influenced by three components: capability, opportunity, and motivation. Designing interventions that address these three components can promote engagement and communication around the dDPP application.

These theories and frameworks can serve as valuable guides to inform the development of your novel engagement intervention for the digital diabetes prevention program. It is important to adapt and tailor them to the specific needs of your target population and context.

A. **ChatGPT generated output** Prompt: You’re a researcher studying how to improve patient-centric engagement and communication around digital behavior change technologies. The specific digital tool is a digital diabetes prevention program (dDPP) application, which you are looking to enhance with a novel engagement intervention. What are some well-accepted behavior change theories or frameworks that you can use to inform your intervention development?

GenAI Solution

When prompted, ChatGPT identified seven relevant well-accepted behavior change theories and frameworks to inform a “dDPP support intervention” (Figure 2B). It did not provide information on the transtheoretical domains framework, or the taxonomy of behavior change techniques, but when prompted on these, identified both as acceptable strategies for use.

**User Experience: Design Principles, Personas, and Messaging Content**

Human-Generated Solution

The research team used a human-centered design approach to identify key design principles, defined as the set of considerations that form the basis of the PAMS product (Figure 3B). These were developed from insights gathered via a review of relevant digital behavior change research, consultation with content and theoretical experts in digital health and implementation science (n=3), and two group interviews (n=9). From these insights, five relevant fictional personas were designed to capture the various phenotypes of user engagement with the commercial dDPP, along with unique user journeys developed to describe their projected engagement with the program over time (Figure 3D). Overall, over 193 unique messages were developed, each grounded by a relevant behavior change technique and tailored to an individual phenotype’s user journey. These elements were continuously revisited and refined during the testing phases of the dDPP research. This included a 6-month near-live user testing phase consisting of nine patients engaging with various iterations of the PAMS prototype, and a 12-month live single-arm pilot phase consisting of 25 patients using PAMS-beta with the commercial dDPP platform.
GenAI Solution
ChatGPT was prompted from multiple perspectives (researcher, clinician, and patient) to identify key design principles (Figure 3A) and sample solutions for the PAMS intervention. It also provided common engagement phenotypes for digital health tool users, based on patterns of use, frequency, duration, and “other elements.” Of note, nonadopters were not identified within the initial round of phenotypes. ChatGPT also developed personas for each of the identified engagement phenotypes, including persona names, backgrounds, and individual journeys. ChatGPT was able to produce five to ten unique messages targeted toward each phenotype and to adapt these messages based on various additional prompts. The user types or personas generated by ChatGPT are consistent with the human-generated users and cover all the phenotypes identified in our previous research (eg, mapping to a specific behavior change technique and reflecting a key design principle; Figure 3C).
PAMS Technical Development

Overview
The technical development includes a PAMS requirements document and architectural design and code.

Technical Requirements (User Stories)

Human-Generated Solution
Following the data collection and intervention design period, we created, as a team, a series of user stories (Figure 4B) which were followed by system requirements to describe the intended use cases, features, and challenges of the proposed PAMS software. Initial system requirements represent the “minimum viable product” that was developed, piloted, and further refined (Figure 4D). Our development team followed software engineering principles to generate the requirements document.
Figure 4. User stories and technical requirements outcome of ChatGPT vs human-generated outputs. API: application programming interface; CDC: Centers for Disease Control and Prevention; EHR: electronic health record; PAMS: personalized automatic messaging system; REDCap: Research Electronic Data Capture; UX: user interface.

A. ChatGPT-generated output. Prompt: You are a software engineer working in a medical school and designing a system to support a workstation and control for prehabilitation patients aiming to increase patient adherence to the digitalwellness program. In the trial, patients are obsessed and enrolled in the API for seamless data exchange. Additionally, they are provided with a Fitbit as an easy-to-use support data collection. You need to define a system that would collect patient data from a digital well-being app and analyze it to determine the adherence to the patient intervention in the prehabilitation phase. Every week, each patient will be assigned a new adherence as a function of their adherence score on the previous week. The adherence score will be calculated based on the following thresholds: 0 to 100 points for the â€œlow adherenceâ€ category, and 0 to 500 points for the â€œgood adherenceâ€ category. The system shall determine if the patient has met the adherence criteria for each behavior metric over all three weeks. The system shall determine if the patient has met the adherence criteria for each behavior metric over all three weeks. The adherence score shall be calculated by the formula: adherence score = (number of successful adherence events / total number of adherence events) * 100. The system shall be designed to provide feedback to the user when they are likely to fall below the adherence criteria.

GenAI Solution
We used the output of the “feature construction phase” to inform the GenAI output for requirements. During the initial stages of the prompting phase, we refrained from suggesting solutions, allowing ChatGPT to generate potential solutions autonomously. We reviewed and evaluated these outputs, eliminating impractical or incompatible solution paths that did not align with the intentions or capabilities of our team. Once we reached a satisfactory outcome but faced uncertainty regarding the next steps, we instructed ChatGPT to assume a different “personality” (eg, software architect) and used the previous outputs as a foundation for the new role’s initial prompts. Throughout this process, we encouraged each “personality” to seek clarifications by asking questions and provided feedback without biasing toward any predetermined solution. We repeated this process at least four times for each personality type, engaging in a comprehensive review of the outcomes to inform the next iteration. B. Human-generated output.

As a domain expert...

1. As a designer, I want to be informed of the features and performance metrics of the prototype.
2. As a developer, I want to be able to modify the prototype to meet new requirements.
3. As a user, I want to use the prototype to test new features.
4. As a manager, I want to track progress and provide feedback.

As a domain expert...

1. I want to know how the prototype will perform in the real-world setting.
2. I want to ensure that the prototype is scalable and robust.
3. I want to test the prototype with real users.
4. I want to ensure that the prototype meets all regulatory requirements.

As a developer...

1. I want to have easy access to the technical documentation.
2. I want to be able to integrate the prototype with existing systems.
3. I want to ensure that the prototype is efficient and scalable.
4. I want to test the prototype with real users.

As a manager...

1. I want to have clear communication with the development team.
2. I want to ensure that the prototype meets all regulatory requirements.
3. I want to test the prototype with real users.
4. I want to ensure that the prototype is scalable and robust.

As a user...

1. I want to have easy access to the prototype.
2. I want to be able to provide feedback on the prototype.
3. I want to ensure that the prototype meets all regulatory requirements.
4. I want to test the prototype with real users.
back-and-forth roleplay with multiple personalities (researcher, architect, and developer), transitioning to a different personality when it became evident that the current one could no longer progress without additional feedback (Figures 4A and 4C).

Architectural Design

Human-Generated Solution

After the requirement phase, our software development team developed the PAMS architectural diagram, which is a graphical representation of the system that includes (1) a set of components (eg, a database and computational modules) that will perform a function required by the system; (2) the set of connectors that will help in coordination, communication, and cooperation between the components; and (3) conditions for how components can be integrated to form the system (Figure 5B).
GenAI Solution

For the GenAI-generated architectural design, we leveraged the outputs of the requirement phase and the available ChatGPT plugins to designate the GenAI model as a software engineer and proceeded to develop an architectural diagram. During this process, we engaged in iterative prompting and provided explicit instructions to ChatGPT, specifying the use of Amazon Web Services (AWS) for development, integration of external systems such as Twilio (Twilio Inc) and REDCap (Research Electronic Data Capture; Vanderbilt University), and the adoption of a microservice approach to facilitate the efforts of our development team (Figure 5A).
**Code**

**Human-Generated Solution**

PAMS components include several lambda functions that execute its engagement or adherence algorithm, messaging, and data manipulation functionalities. Most of the functions are coded and developed using Python (Python Software Foundation) and Scala (École Polytechnique Fédérale Lausanne) as programming languages. AWS was used for the development of PAMS [12]. Our developers followed our microservice approach design using an event-driven model [13,14]. The main components of PAMS are AWS lambda functions which are triggered by different events such as updates to S3 buckets, modifications on DynamoDB (AWS) tables, or CloudWatch (AWS) events. External interactions of PAMS use application programming interface calls, which secure effective data transfer (Figure 6B).

**Figure 6.** Code for the function that calculates patient adherence and engagement trends. ChatGPT vs human-generated outputs.

A. ChatGPT generated output. Prompt: You are an experienced software developer tasked with writing code for the business logic. Write code in Scala for the function responsible for calculating engagement trends for patients every three weeks. Assume all other functionality is being worked on independently and mock calls to services and data sources are needed.

B. Human generated output. The code was generated by our developer following traditional software engineering principles and standardized code development styles.

**GenAI Solution**

To facilitate the generation of the coded solution using ChatGPT, we assigned the role of a software engineer to the model and specifically requested it to generate Scala code for a specific functionality, namely the “calculate engagement trends” function. Consistent with the iterative nature of the GenAI-based software development process, we engaged in a back-and-forth interaction with ChatGPT, iterating over the prompt and its output while providing expert guidance to ensure optimal results. While allowing ChatGPT to generate free text, we evaluated each output for accuracy and adherence to the desired specifications (Figure 6A).

**Internal Review of Human Vs GenAI Outputs**

The 11 evaluators participated in the output review process. All had familiarity or prior exposure to the original PAMS intervention. Overall, evaluators rated the ChatGPT-produced outputs as positive for the theoretical background and design phase in terms of understandability, usability, novelty, relevance, and efficiency. For these two components, the question about completeness showed the most variability with divided opinion among “agree” and “disagree” and the bias was mostly categorized as “neither agree nor disagree.” For the first part of the technical development (user stories and requirement documents), most of the raters found the ChatGPT output positive in terms of understandability, usability, and relevance. In terms of completeness and novelty, requirements were better rated than the user stories which represent an interesting output since requirements are derived from the user stories. We hypothesize that our raters were expecting better user stories, but once these were defined, they considered ChatGPT to be effective at turning these into the requirements. In terms of bias, similar to the theoretical background and design phase, the most popular answer was “neither agree nor disagree.” For the more technical pieces of the development that required software engineering knowledge, specifically the architectural diagram and code elements, results showed the highest N/A responses. These higher levels of N/As were associated with lower levels of expertise (eg, coding experience) since only 2 of the 11 evaluators had computer science backgrounds. However, the overall score excluding the N/As was positive for the technical component.

**Discussion**

**Results Summary**

This study leveraged ChatGPT-4 to recreate content features and software development of PAMS. ChatGPT served as a usable facilitator for researchers engaging in the software development life cycle, from product conceptualization to...
feature identification, and user story development to code generation. GenAI technologies facilitated effective communication and understanding within our multidisciplinary team by providing well-described features and supporting the role of a software engineer. Our findings indicate that the ChatGPT-generated output is comprehensive, albeit with occasional ambiguities that required clarification or adjustment by the research team. The ChatGPT-generated output exhibited a high level of accuracy in capturing the intended requirements. We found that ChatGPT supported a highly efficient development process, producing over 5 days what initially required more than 200 human hours from content and technical experts. The results suggested that by efficiently prompting ChatGPT and leveraging the expertise of our team, we could have significantly reduced the time we invested in initial system modeling and conceptualization phases as well as technical phases of software development (coding). Overall, GenAI technologies like ChatGPT offer a promising approach to efficient software development.

While promising, some significant limitations to ChatGPT’s outputs should be noted. In the design phase, while ChatGPT was able to provide general guidance in tool design (eg, app vs web-based vs EHR solution) it was unable to provide evidence to support its rationale for these choices. This lack of reference support has been well-documented and has a material impact on researchers looking to build upon an evidence base for their health technology interventions. Similarly, when asked to provide theoretical frameworks to support behavior change, it offered only a partial list, initially excluding the COM-B (capability, opportunity, motivation, behavior) model upon which the original PAMS intervention was based, and needed prompting from our behavior change expert to provide more specific guidance. In the context of code generation, we focused on testing a specific function, namely the Calculating Patient Engagement feature, which is the core functionality of our system. However, the initial attempts exhibited nonidiomatic constructs and contained bugs (no efficient loops and wrong logic). Finally, we observed that ChatGPT overlooked certain suggested features during the design phase, resulting in the generated code occasionally demonstrating unnecessary complexity and disregarding some of the best practices and features of the target programming language. We believe that further iterations would have improved the code quality, encompassing better adherence to coding standards and the inclusion of desired business features, such as handling edge cases and capturing more nuanced engagement trends. Nevertheless, we reached a point of diminishing returns with ChatGPT where we determined that engaging an experienced developer would have expedited the code generation process and ensured a more robust implementation.

These limitations highlight the ongoing importance of human expertise in the development process, especially in scenarios where theoretical expertise, intricate coding practices, and business-specific requirements are involved. The lack of rationale to support the generated results shows the value of having human experts on the team who can interpret the results. ChatGPT needs to be used as a support tool but not the source of truth; thus, we always trusted and relied on human experts to validate the ChatGPT-generated results before moving to the next phase. Overall, it is important to have human experts in the system development process to guide the outputs in terms of reprompting the system (support the decision-making on acceptable output) and ensuring their accuracy. Moreover, results are highly dependent on the quality of the prompts which emphasizes the role of prompt engineering. The results show that well-structured prompts (role + problem description + ask) that infuse human expertise into every iteration are key to obtaining good results (Figure 1). As part of our prompt framework described in the methodology section, results showed that detailed problem explanations, clear asks, and roleplaying are an excellent combination to guide accurate results. We suggest asking ChatGPT questions using different roles, asking for clarification if needed, and in cases of wrong outputs, redirecting the prompts.

Related Work
There is near-universal interest in understanding the impacts of GenAI and large language models (LLMs) on human social structures, including the experience of work and the production of work-related outputs in health care and more broadly [15,16]. In health care, LLMs are poised to impact everything from care delivery experience, diagnostic reasoning and cognitive skills, training and education, and the overall composition of the workforce [17]. These theoretical disruptions are tempered, however, by acknowledging that in its current state, GenAI tools remain suboptimal, with ongoing issues in accuracy, reliability, usability, cost, equity, and ethics.

In commercial spaces, ChatGPT-enabled products designed to assist with coding and software development are already being developed (eg, OpenAI Codex [OpenAI] and CodeGPT [CodeGPT]). These tools can help generate novel code, debug and analyze code issues, assist in code refactoring, and provide code documentation. As yet, however, their usefulness in terms of quality has not been extensively evaluated, and costs and other considerations may make them inaccessible to health care researchers. ChatGPT-enabled tools for front-end design (eg, integrating ChatGPT with Figma [Figma, Inc]), user testing (including synthetic user testing), and prototyping have also been created, all allowing health technology research teams with limited design resources to take advantage of tools from product and experience design to create their interventions. Overall, commercial LLMs have been demonstrated to improve worker efficiency and productivity, through “co-pilot” support services that automate low-skills tasks, organize and present information, and surface insights [18]. Brynjolfsson et al [18] found that a ChatGPT-supported tool providing conversational guidance for customer support agents increased worker productivity by almost 14%. The authors further found that these productivity benefits accrued disproportionately to less-experienced and lower-skill workers, allowing less-skilled or newer workers to experience more rapid gains; the authors posit that high-skill workers may have less to gain from artificial intelligence technologies like ChatGPT offer a promising approach to efficient software development.
intelligence assistance due to tacit knowledge reinforcement rather than new knowledge or skill development. Our work suggests that both less-experienced, lower-skill workers and high-skill workers can benefit, with novices benefiting more from new knowledge (if accurate) and skill development and experts benefiting from knowledge validation and offloading of high-effort low-value tasks.

In the academic computer science literature, ChatGPT has been evaluated as a tool for collaborative software design [4], including to improve code quality refactoring, requirements elicitation, and general design solutions [5], and fix programming bugs [19]. Similar findings are reflected in our work, including the caveats of requiring human oversight. Other authors have identified important ethical issues in using GenAI solutions for software engineering, which were not considered in this study [20].

Within health care, a growing body of research has explored the feasibility of GenAI tools (mostly ChatGPT) in a variety of use cases, including answering patient questions [3,21], creating suggestions to optimize clinical decision support [22], generating a history of present illness summaries [23], and overall examination performance [24]. In general, these papers find promising signals for the accurate and acceptable use of GenAI tools, but with many current-state caveats for their optimal, safe, and scaled use. Key areas of concern include reliability (particularly around hallucinations and citation fabrication), reproducibility, and recency of data inputs. While research in this area will continue to grow, as more test cases comparing GenAI performance to that of clinical staff will be undertaken, further work is needed to create validated and generalizable outcome measures. Future work must also ensure that the variety of GenAI tools (including general commercial LLMs, health care–specific LLMs, and internally developed tools) are equally evaluated.

Limitations
There are several limitations to this study. First, no research team members have expertise in prompt generation for GenAI tools; as a result, our prompting reflects the a priori perspectives, biases, and knowledge gaps of our team, and are therefore particularly subject to issues of framing, recall, and confirmation bias that may influence the interpretation of the results. Second, our research team members, who acted as prompt engineers in this study, were highly familiar with the project and participated in the human-based design process; thus, they were aware of what deviations from human-based design to address by reprompting the system. As a result, we have introduced bias in the prompting process and results reflect higher accuracy. Third, the absence of robust tools to objectively measure the “quality” of current ChatGPT outputs poses challenges to accurately and objectively assess its performance. Furthermore, in this case, the output reviewers were not blinded to the human vs ChatGPT outputs, given the complexity of this study and the difficulty in providing enough research context to support independent blind review. Finally, broader limitations of the technology, such as potential hallucinations and concerns about behavioral changes of responses over time, deserve acknowledgment, as they could have implications for the practical applications and long-term viability of GenAI in health care research contexts. Future research efforts should address these limitations to enhance and replicate our findings.

Implications and Future Directions for Exploration
We are considering several future directions for the use of ChatGPT in our digital health intervention development. We envision increasing our expertise in prompt engineering (add expert prompt engineers to the team) to actively use ChatGPT to further develop PAMS features, particularly for additional messaging content. We anticipate this will save our research team considerable time and effort. We may also use ChatGPT to facilitate more time-consuming aspects of our research documentation, including both coding documentation and larger research archival work (eg, meeting minutes and recording intervention decision-making). Overall, we feel ChatGPT and related tools can be effectively leveraged within health care technology research teams with a spectrum of technical expertise, serving to both augment existing skills and supplement skill gaps. For those with expertise in computer science or programming, we imagine ChatGPT can assist by automating high-effort, low-impact tasks or repetitive work that is considered important but often deprioritized as more urgent tasks arise (eg, code documentation). For those without preexisting programming skills, we imagine ChatGPT can offer technical support, including educational tools and skill-building opportunities. Overall, this process will both validate existing knowledge and create new knowledge for teams, as well as potentially improve interteam communication and collaboration.

Conclusions
In this study, we explored the use of the GenAI tool ChatGPT to recreate a novel digital behavior change intervention which our research team had previously developed to support patient engagement and adherence to a commercial dDPP. Specifically, we reviewed and evaluated the capacity and limitations of ChatGPT to support digital health research intervention ideation, design, and software development, finding it a feasible and potential time- and resource-saving tool to support research teams in developing novel digital health products and technologies. At the same time, we identified gaps in ChatGPT outputs that may limit its effective use for both novel and advanced technology developers, particularly around the completeness of outputs. Future directions will include the development of more targeted artificial intelligence–based tools to support health care researchers with all levels of software or engineering skills, as well as the development of improved tools to objectively evaluate GenAI outputs.
Acknowledgments

The National Institute of Diabetes and Digestive and Kidney Diseases (1R18DK118545-01A1, principal investigator: DM) funded this research.

Conflicts of Interest

None declared.

References


Abbreviations

- **AWS**: Amazon Web Services
- **COM-B**: capability, opportunity, motivation, behavior
- **dDPP**: digital diabetes prevention program
- **GenAI**: generative artificial intelligence
- **LLM**: large language model
- **N/A**: not applicable
- **PAMS**: personalized automatic messaging system
- **REDCap**: Research Electronic Data Capture

©Danissa V Rodriguez, Katharine Lawrence, Javier Gonzalez, Beatrix Brandfield-Harvey, Lynn Xu, Sumaiya Tasneem, Defne L Levine, Devin Mann. Originally published in JMIR Human Factors (https://humanfactors.jmir.org), 06.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Human Factors, is properly cited. The complete bibliographic information, a link to the original publication on https://humanfactors.jmir.org, as well as this copyright and license information must be included.
Testing Two Online Symptom Checkers With Vulnerable Groups: Usability Study to Improve Cognitive Accessibility of eHealth Services

Kaisa Savolainen1, DA (Art and Des), MSc (tech); Sari Kujala1, PhD
Department of Computer Science, Aalto University School of Science, Espoo, Finland

Corresponding Author:
Kaisa Savolainen, DA (Art and Des), MSc (tech)
Department of Computer Science
Aalto University School of Science
Konemiehentie 2
Espoo, 02150
Finland
Phone: 358 503101457
Email: kaisa.savolainen@aalto.fi

Abstract

Background: The popularity of eHealth services has surged significantly, underscoring the importance of ensuring their usability and accessibility for users with diverse needs, characteristics, and capabilities. These services can pose cognitive demands, especially for individuals who are unwell, fatigued, or experiencing distress. Additionally, numerous potentially vulnerable groups, including older adults, are susceptible to digital exclusion and may encounter cognitive limitations related to perception, attention, memory, and language comprehension. Regrettably, many studies overlook the preferences and needs of user groups likely to encounter challenges associated with these cognitive aspects.

Objective: This study primarily aims to gain a deeper understanding of cognitive accessibility in the practical context of eHealth services. Additionally, we aimed to identify the specific challenges that vulnerable groups encounter when using eHealth services and determine key considerations for testing these services with such groups.

Methods: As a case study of eHealth services, we conducted qualitative usability testing on 2 online symptom checkers used in Finnish public primary care. A total of 13 participants from 3 distinct groups participated in the study: older adults, individuals with mild intellectual disabilities, and nonnative Finnish speakers. The primary research methods used were the thinking-aloud method, questionnaires, and semistructured interviews.

Results: We found that potentially vulnerable groups encountered numerous issues with the tested services, with similar problems observed across all 3 groups. Specifically, clarity and the use of terminology posed significant challenges. The services overwhelmed users with excessive information and choices, while the terminology consisted of numerous complex medical terms that were difficult to understand. When conducting tests with vulnerable groups, it is crucial to carefully plan the sessions to avoid being overly lengthy, as these users often require more time to complete tasks. Additionally, testing with vulnerable groups proved to be quite efficient, with results likely to benefit a wider audience as well.

Conclusions: Based on the findings of this study, it is evident that older adults, individuals with mild intellectual disability, and nonnative speakers may encounter cognitive challenges when using eHealth services, which can impede or slow down their use and make the services more difficult to navigate. In the worst-case scenario, these challenges may lead to errors in using the services. We recommend expanding the scope of testing to include a broader range of eHealth services with vulnerable groups, incorporating users with diverse characteristics and capabilities who are likely to encounter difficulties in cognitive accessibility.

(Keywords: eHealth; online symptom checkers; usability; cognitive accessibility; web accessibility; qualitative research)
Introduction

Background

Given the widespread use and popularity of eHealth services, there is a growing need for more accessible services to all potential user groups [1]. In recent years, more emphasis has been placed on accessibility and inclusion; for example, the European Union Accessibility Act has been incorporated into and enforced as national law since June 2022 [2]. As health care services are often public services, it is important that they serve a broad range of users. Furthermore, usability has been recognized as a key component of eHealth applications, and users may face problems with using the applications due to their health conditions [3]. In addition, patients with chronic illness have been reported to encounter more cognitive challenges [4]. Thus, extra attention should be paid to the usability of eHealth applications.

Universal design and design for all address these requirements by aiming at designing services that are usable by and accessible to all user groups regardless of their age, abilities, or possible disabilities [5]. Usability is a high-level term that indicates how a system can be used by specified users in a certain context of use to achieve specific goals with regard to effectiveness, efficiency, and satisfaction [6]. Accessibility, which is a part of usability, describes how a system can be used by people with the widest range of needs, characteristics, and capabilities [6,7]. Thus, accessibility covers all sorts of users with different limitations. A concept that has been addressed by several research papers [8,9] is web accessibility (or e-accessibility), which refers to the accessibility of web services.

In this paper, we address cognitive accessibility, which refers to accessibility beyond physical and sensory capabilities, and thus takes into account varied human characteristics such as intellectual disabilities, attention difficulties, reading problems, autism spectrum disorders, and low language skills [10]. Cognitive accessibility is an important aspect of web accessibility as it involves a large number of users and has a high impact on usability [10]. A summary of the relationship between these concepts is presented in Figure 1.

Figure 1. The relation of cognitive accessibility to usability and accessibility. Note that the sizes and positions of the circles are indicative.

This research focuses on cognitive accessibility within the context of the 2 most frequently used online symptom checkers in Finnish public primary care across numerous municipalities in Finland. Online symptom checkers are used by people seeking health-related guidance, and these services typically provide an urgent assessment and suggest guidance based on the symptoms reported by the user [11]. Patients can use the 2 examined symptom checkers to book appointment times for doctors and laboratory tests or obtain medical help for the most common health issues. First, patients report their symptoms and submit them to the health care center through the symptom checker. Health care professionals receive patient inquiries with an urgency rating, decide on actions to be taken, and inform patients.

Patients are generally highly satisfied with symptom checkers, but younger and more highly educated people have been more likely to use them [11]. For example, symptom checkers enable patients to access health care anytime and anywhere. Therefore, it is essential to ensure that all user groups, including individuals in vulnerable situations, can use these services effectively. Symptom checkers can also empower users as a means of facilitating their health care [12]. However, the accuracy of the symptom checkers depends on how well patients are able to communicate their symptoms when using the tools [13]. As these services spread and are used by a wider range of individuals, it is crucial to also evaluate their usability and accessibility with a more diverse set of users.

Prior Work

Vulnerable Groups

Many public eHealth services and their poor usability and accessibility can cause challenges for certain user groups [14]. These user groups are, thus, in a potentially vulnerable situation in using the service and at risk of digital exclusion [1]. This is especially problematic because research has shown that digital exclusion can cause social exclusion [15]. Public health services must, thus, address the needs of potentially vulnerable groups, including people who are disadvantaged by health, economic, cultural, or social conditions [16], such as older adults, migrants, mental health service users, and the unemployed [16,17].

Older adults are the largest group to face challenges in using digital health services [18,19]. As people age, their cognitive abilities may weaken, with cognitive load being identified as the most significant accessibility barrier for older adults [20]. Memory changes can also affect learning, information processing, and language comprehension [21,22]. Additionally, older adults often struggle with focusing their attention, particularly when multitasking [21,22]. Moreover, older age
groups tend to use eHealth services less frequently than younger demographics. A Finnish study examining an online symptom checker (referred to as service A in this study) observed that individuals aged 20-39 years used the service more actively compared with older age groups, relative to their representation in the population [23]. This suggests that enhancing service usage entails prioritizing usability and accessibility from the perspective of older users as well.

Migrants represent a growing demographic that often faces challenges when accessing health services in their new country of residence [1]. Language barriers and a lack of digital skills are common issues encountered by this group [1]. Additionally, individuals with intellectual disability are another vulnerable population impacted by the digitalization of health services [24]. They have been noted to experience more difficulties in finding information on the internet and understanding online information compared with the general population [25].

Previous research suggests that vulnerable groups, such as older adults and individuals with mild intellectual disabilities, encounter cognitive challenges when using technology [26,27]. Therefore, the development of more accessible eHealth services would enable these groups to access health information more easily [25,28], thereby enhancing their sense of empowerment concerning their health issues.

The preferences or needs of older adults or individuals with mild intellectual disabilities are often overlooked in the majority of eHealth studies [29,30]. It is imperative to better consider these user groups during the design of eHealth services [17,28]. Many eHealth applications could greatly benefit from the application of universal design principles [29], which facilitate understanding the needs of potentially vulnerable groups and inform the design of more inclusive and usable services [31,32]. Consequently, this enables vulnerable groups to derive as much benefit from eHealth systems as the rest of the population [33,34]. Indeed, universal access approaches can offer benefits to anyone [35]. Therefore, to gain a better understanding of the challenges faced by vulnerable groups when using services, it is essential to conduct testing with a diverse group of users.

**Usability Testing of Symptom Checkers**

The usability of symptom checkers has been examined in prior research; however, there has been limited emphasis on potentially vulnerable user groups, such as older adults, migrants, and those with intellectual disability [36-38]. Moreover, research on usability in the eHealth domain frequently concentrates on quantitative aspects (eg, the number of errors, task completion times, and usability questionnaires) and typically involves a large number of users [12,36,38]. However, the qualitative aspect of usability studies is also crucial for gaining a deeper understanding of the thoughts and reasons behind errors, as well as capturing the patient’s perspective at a broader level [3,39]. Additionally, while a System Usability Scale (SUS) questionnaire provides a numeric score for experienced usability, it alone is not adequate for evaluating usability. Instead, it should be complemented with other measures, such as task completion rates or more qualitative approaches, to ascertain which aspects of a service require improvement and how best to address them [39].

Marco-Ruiz et al [13] conducted research on symptom checkers and emphasized the significance of testing with real users to comprehend the cognitive processes involved when using a new system to record health data. Furthermore, they noted that the user base accessing symptom checkers is highly diverse, with some individuals possessing higher health literacy and experience in recording online information, while others may have very limited or no experience [13].

**Goal of the Study**

The goal of our study is to gain a deeper understanding of cognitive accessibility in the context of eHealth services. Therefore, our paper focuses on addressing the following research questions:

- What kind of challenges do vulnerable groups face in using eHealth services?
- What needs to be considered when testing with vulnerable groups?

The structure of this paper is as follows: In the next section, we describe the methods used in this study, followed by the presentation of results. Subsequently, we discuss the findings and overarching contributions of this study, concluding with our final remarks.

**Methods**

**Approach and Researcher Background**

Our qualitative study adopts a case study approach, wherein the cognitive accessibility of eHealth services was assessed through usability testing of 2 online symptom checkers. The research team comprised 3 researchers: The first researcher, a human-computer interaction student, conducted the initial 8 tests as part of their master’s thesis work. Subsequently, a second researcher, a senior researcher with expertise in human-computer interaction (who served as the thesis advisor), conducted the remaining 5 tests. Additionally, a third senior researcher with backgrounds in human-computer interaction and eHealth oversaw the entire study.

**Context and Study Setting**

We conducted a usability test of 2 Finnish online symptom checkers in 2 phases in Finland during the Spring and Fall of 2021. The tested services were Omaolo (DigiFinland Oy) [40] and Klinik Access (Klinik Healthcare Solutions Oy) [41], which are the 2 most-used symptom checkers in Finnish public primary care. Omaolo has been actively used since 2019, while Klinik Access, which is also used internationally, has been in use since 2015. Both services are designed to assist patients in obtaining appropriate care. Users answer a set of questions regarding their symptoms, following which the symptom checkers use artificial intelligence to assess the urgency of care. If necessary, the services guide patients to contact emergency care services.

The Omaolo symptom checker comprises 15 specialized symptom checkers tailored for different types of symptoms, along with a generic symptom checker. Each symptom checker prompts the user with a specific set of questions and subsequently recommends the next steps they should take. Additionally, if the user provides their home municipality, the
service displays recommended actions specific to the area, offers contact details, and may even facilitate direct contact with health care professionals if deemed necessary. The Omaolo symptom checker served as the primary COVID-19 symptom checker in Finland, enabling users to schedule appointments for COVID-19 tests. Consequently, its user base experienced a significant surge [23].

The Klinik Access symptom checker enables users to initially select the part of the body where their main symptoms are located. Subsequently, it prompts for more specific symptoms. The responses can then be forwarded to the medical staff responsible for the patient’s care before their appointment, ensuring the patient is directed to the appropriate type of health care professional. The primary distinction between these services lies in their user interface (UI): Klinik Access features a more visual UI with a list of clickable symptoms, whereas Omaolo presents users with multiple-choice questions describing the symptoms. Henceforth, the Omaolo service will be denoted as service A, and Klinik Access will be referred to as service B. It is important to note that both services are classified as medical devices and must adhere to specific safety requirements, such as repetitive questions, which may impact usability.

**Sampling Strategy**

Purposive sampling [42] was used to recruit participants, who were sourced through personal contacts and various associations representing the targeted user groups. These associations included initiatives such as the Selkäisäistä meille, which focuses on enhancing cognitive accessibility, and Väyläry, which is dedicated to improving the employment opportunities of individuals with intellectual disability. It is important to note that the test facilitator did not have a close personal relationship with the participants, such as being a friend or family member, during any of the test sessions. A total of 13 participants were recruited to partake in the study. Notably, an evaluation of sample sizes within the field of human-computer interaction has indicated that 12 is the most common sample size for usability studies [43].

**Ethical Approval**

The study received approval from the ethical review board of Aalto University (D/902/03.04/2021). Each participant provided informed consent by signing a consent form after confirming their understanding of the study’s purpose and how their information would be handled. Reporting has been conducted in such a manner that individual participants cannot be identified.

**Data Collection Methods**

**Overview**

The main methods used in this study were thinking aloud, observations, questionnaires, and semistructured interviews. Before the actual tests, a pilot test was conducted to identify any potential inconsistencies and to ensure that the questions and instructions were comprehensible. Minor adjustments to the test setup were made based on the findings from the pilot test.

**Test Procedure**

An overview of the test sessions is presented in Figure 2.

**Figure 2.** An overview of the usability test sessions with older adults, mildly intellectually disabled individuals, and nonnative speakers (N=13). Half of the participants started with Service A and the other half with Service B.

Each participant tested both services, and the order of service usage was counterbalanced. During the testing phase, participants were presented with 2 symptom vignettes, each providing a brief description of the symptoms they were instructed to imagine having. These vignettes were used 1 at a time. Participants were then asked to open the service and imagine they had the symptoms described in the first vignette, aiming to determine how they should proceed. The vignettes and mode of distribution between the participants are presented in Multimedia Appendix 1.
After using the first service, participants were instructed to take the second vignette and attempt to use the service again. However, if the first part of the test had exceeded 40 minutes, the second vignette was omitted for the first service to prevent the overall test time from exceeding 90 minutes. Following their interaction with each service, participants were asked to evaluate the respective service.

After testing the first service, participants were instructed to open the second service and follow the same procedure. Upon completion of both testing phases, participants were asked to compare the 2 systems and select the one they preferred.

Data Collection Instruments

Test Sessions

The test sessions were conducted via the Microsoft Teams videoconferencing platform, which facilitates screen sharing, screen recording, and voice recording functionalities. The decision to conduct remote testing was primarily influenced by the COVID-19 pandemic situation, but it also aligned well with the nature of the tests, as the services being evaluated were online. Participants used their personal computers to access the services during the testing sessions.

Symptom Vignettes

To streamline the usability test and eliminate the necessity for participants to input their personal medical information into the services, each participant was provided with 2 standardized clinical vignettes featuring predefined symptoms. These vignettes were selected from a list compiled by Semigran et al [44], encompassing a total of 6 conditions with varying severity levels. The selection included conditions with different severity levels to account for the fact that individuals may use symptom checkers in both urgent and nonurgent situations [45].

In line with the recommendations provided by Semigran et al [44], the selected vignettes encompassed 3 categories of triage urgency: conditions necessitating emergency care, conditions warranting nonemergency care, and conditions deemed unnecessary for medical visits, thus manageable with self-care. Moreover, we opted for conditions commonly observed within the age group under study to ensure relevance. These conditions encompass ailments such as acute bronchitis, back pain, and meningitis. To ensure clarity and relevance to the participants, the selected conditions were translated from English to Finnish and simplified. The English versions of the vignettes used can be found in Multimedia Appendix 2.

Background Questionnaires

Before the actual test session, participants were requested to complete a brief background survey and the health literacy survey HLS-EU-Q16 [46]. The background information collected were the participant’s gender; age; the frequency of doctor visits in the preceding 2 years; the number of doctor-diagnosed medical conditions; their previous usage frequency of digital health care services; and their frequency of digital device usage, such as smartphones or computers. These questions aimed to ascertain whether participants met the study’s target demographic criteria in terms of age and their ability to independently use electronic devices such as computers. The health literacy survey provided insights into participants’ understanding of health-related topics.

Interview and Questionnaire

After interacting with each service, participants were asked to evaluate the tested services. This involved administering an SUS questionnaire [47] to gauge the perceived usability of the system, as well as posing 4 interview questions:

- Would you use the service again in the future?
- Were the summary and the instructions about what to do next clear enough?
- Would you actually follow the instructions given?
- Given the option, would you use the service using your phone?

Data Processing and Analysis

The test sessions were recorded using Microsoft Teams. The voice recordings of the initial 8 tests were transcribed in full, while for the remaining 5 tests, notes were taken from the recordings, and user comments were documented to streamline the process. An experienced researcher could identify the issues encountered by users as well as their comments without requiring a complete transcription. The notes and transcriptions underwent anonymization. Qualitative content analysis was used in this study. Using the notes and recordings, all usability issues were identified and compiled. This encompassed problems mentioned by participants as well as those observed during testing or evident from the recordings. The identified usability problems were coded and categorized based on their similarities. When new problems were identified, they were compared with existing ones, and if deemed similar, they were grouped under the same code. Eventually, these groups were consolidated under higher-level descriptive categories. Furthermore, user comments were collected to bolster the analysis and reporting process.

The background questionnaires were analyzed by aggregating the responses to obtain an overview of participant characteristics. Additionally, the health literacy surveys were analyzed according to the guidelines [46] to determine the groups to which participants belonged. The SUS questionnaires were analyzed by computing the SUS scores as per the guidelines [47], resulting in scores of up to 100 points, which were then compared with the general score.

To ensure the quality and trustworthiness of the study, a senior researcher (the second author) supervised the entire research process and provided support for the analysis work. Two other researchers (the first author and the master’s thesis worker) conducted the actual tests and analyzed the data. Therefore, a total of 3 researchers participated in the process, ensuring that data gathering and processing proceeded appropriately.

Results

Overview

The subsequent sections present the principal findings of the study. We commence with an overview of the participants’ characteristics, followed by an examination of the identified
usability issues. Finally, we present additional findings that emphasize the characteristics of these user groups.

**Test Participants**
A total of 13 individuals participated in the study in Finland. Among them, 4 were individuals with mild intellectual disability, 4 were older adults (aged 75-79 years), and 5 were nonnative Finnish speakers. Therefore, all test users potentially encountered cognitive accessibility challenges with the services. The background characteristics of the participants are detailed in Multimedia Appendix 3.

The HLS-EU-Q16 questionnaire results were calculated in accordance with the guidelines [46], with each participant receiving a score corresponding to a quartile representing their health literacy level. The results were computed only for participants who responded to at least 80% of the questions, as recommended in the guidelines [46]. The questionnaire includes an “I don’t know” answer option, which was interpreted as the question not being answered. Consequently, the results of 2 of the nonnative participants were excluded, as they chose this answering option too frequently. Figure 3 depicts the distribution of health literacy among the 3 groups.

**Usability Problems**

**Cognitive Accessibility Issues**
The study identified a total of 65 usability problems with the 2 systems. Specifically, 36 usability problems were discovered with service A, while 29 problems were identified with service B. These issues occurred across 99 and 91 individual user instances, respectively. The problems were classified into 14 usability problem categories. A comprehensive list of the usability problem categories is provided in Multimedia Appendix 4. For the purpose of this discussion, we will focus on issues related to cognitive accessibility, primarily concerning terminology, text volume, and UI clarity.

**Terminology-Related Issues**
The most prevalent issues were associated with terminology and answering options. eHealth services frequently incorporate specialized language and specific terminology, posing challenges for users with cognitive limitations. Nearly all users encountered confusion with certain terms or inadvertently mixed them up with similar ones. Furthermore, lengthy words and extensive blocks of text, such as lengthy paragraphs, presented challenges, a sentiment that was also echoed during the interviews. Users with cognitive restrictions often encounter challenges when confronted with long words and extensive passages of text.

As one user commented,

*It takes time to go through all the texts.* [ID10, nonnative]

Related issues were reported and commented on by users across all user groups. In addition to contributing to usability problems, these issues slowed down the usage of the services and occasionally led users to select incorrect symptoms.

**Issues Related to the Clarity of the UI**
Another area where users encountered difficulties was with the visibility of information and the lack of clarity in the UI. It is crucial for the most important information and elements of the UI to be clearly visible, facilitating easy comprehension for users. Additionally, problems arose when users’ attention was diverted to unimportant features. These issues are especially pronounced among user groups with cognitive difficulties, as they require additional attention to comprehend the content and must focus more intently. Furthermore, some users found the input methods challenging; initially, they struggled to discern the type of information required for input in a field and how the inputting should be performed.

The most prevalent individual usability problems we identified regarding the logic and functionality of the UI, observed across all 3 user groups, are detailed in **Textbox 1**.
These individual usability problems highlight issues with how information is presented to users, with clarity being particularly emphasized among this user group. In some instances, the selection or input options were unclear, and the services featured lengthy lists of symptoms.

Clarity was a recurring theme in several test sessions. As one user commented:

...if you think about this in real life, if you have a fever and you’re doing this and you start to scroll all these selection choices and you’re evaluating which one would fit best, the options are quite broad, so it might be quite difficult to do in practice... [ID7, user with mild intellectual disability]

Similarly, one user suggested:

I don’t know you could kind of put those in order like one row and another row, these are quite...your eye kind of jumps, but otherwise those are clear. [ID2, older adult]

One user preferred the structure of service A and, again, referred to the clarity with which the information is presented:

Well, I prefer Service A because it was maybe better organized, there was one thing and one question and then one answer. After this, the next question and so on. In the other one [Service B], you had to read all the small boxes and look for your symptom. [...] [ID8, user with mild intellectual disability]

Well, maybe what is the most [difficult], this one had so many small boxes that at least for me, it was difficult to find my own symptom, the one I needed to select from there. So, if I wanted to know what fit me, I had to read through them all and then, since they are not in any order, they just are there, I had to read them all, to see if I could find the one I have at the moment. [ID8, user with mild intellectual disability]

It is worth noting that the symptoms were arranged in alphabetical order; however, the layout was such that users did not realize this ordering method had been used.

Differences Between the User Groups

Some differences between the user groups were evident, although the majority of the usability problems were consistent across all user groups. Nonnative Finnish speakers found the service to be particularly slow to use, often taking an extended period to read the texts. One user commented regarding service B that:

Reading and writing text is not easy for an immigrant. When you can click on an item it is easy, you don’t have to write. [ID13, nonnative]

The older adults did not encounter as many issues with longer texts. Instead, they faced more challenges in understanding the logic of the services and remembering to scroll down to view all the provided information. However, this scrolling also frustrated some nonnative users; as one user commented:

And again, we’re scrolling, this is terrible! [ID12, nonnative]

The task completion times were also measured and presented for the initial tasks of both services. As depicted in Table 1, aside from the older adults, there were no significant differences in the completion times between the services. However, for the older adults, service B, which featured more clickable elements to choose from, appeared to be quicker to use. Nonnative speakers took the longest time to complete the tasks, primarily because they often needed to translate some of the terms used in the services. Three of the users used an online translator (eg, Google Translator), and at times, users asked the facilitator about specific terms. Overall, the task completion times were
quite lengthy, suggesting that these user groups require ample time to use these services effectively.

**Table 1.** Average task completion times (first task) for both services. For older adults there was a clear difference in favor of service B; for the 2 other groups service A got a slightly better time.

<table>
<thead>
<tr>
<th>Task completion times</th>
<th>Service A, hh:mm:ss</th>
<th>Service B, hh:mm:ss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Older adults</td>
<td>0:14:30</td>
<td>0:08:00</td>
</tr>
<tr>
<td>User with mild intellectual disability</td>
<td>0:14:54</td>
<td>0:16:00</td>
</tr>
<tr>
<td>Nonnatives</td>
<td>0:15:46</td>
<td>0:18:16</td>
</tr>
</tbody>
</table>

*hh:mm:ss: hours:minutes:seconds.*

For the few users who had the opportunity to test the services twice, the second time was generally much faster than the first, indicating good learnability. As one user mentioned:

*Now I know that I need to select this and not the other, which I didn’t know previously.* [ID8, user with mild intellectual disability]

The SUS scores are provided in Multimedia Appendix 5, illustrating how participants evaluated the usability of the services. The SUS score ranges from 0 to 100 points. It has been assessed for numerous services, and according to Bangor et al [48], a satisfactory SUS score is above 70, with superior products typically scoring 80 or higher. However, it is important to note that the interpretation of SUS scores can vary depending on the type of product and its development phase. When evaluating the SUS scores of the tested services, which are predominantly below 75, it is evident that the perceived usability was not considered very good, except for nonnative Finnish speakers, as their scores hovered around 80.

From the interviews, we found that older adults tended to prefer computers over mobile devices when using the symptom checkers, whereas nonnative speakers mostly preferred mobile devices. The preference among users with mild intellectual disability was evenly divided. Nonetheless, the advantage of this type of online symptom checker was evident, as all participants expressed willingness to use the services again. The nonnative participants particularly valued a service that enabled them to input information at their own pace, as opposed to speaking on the phone. However, their preference for the service they would use was fairly evenly split, with no clear consensus: 7 participants favored service A, while 6 participants favored service B.

**Discussion**

**Principal Findings**

Testing for cognitive accessibility with 2 symptom checkers revealed that older adults, individuals with mild intellectual disability, and nonnative speakers may encounter numerous challenges when using the services. Primarily, problems arise concerning the terminology used. This highlights the need for greater emphasis on ensuring that the vocabulary used in the health sector, while specialized, remains understandable to a broad audience when services are intended for universal use. Furthermore, complications arose from the intricate structure and layout of the services. The significance of simplifying services, minimizing lengthy lists, and using more understandable terminology was highlighted in nearly all the test sessions. Implementing these improvements to the services would likely benefit a broader range of users [5].

There were distinct differences observed among the 3 user groups. Primarily, nonnative speakers assigned notably higher usability ratings to the services compared with the other 2 groups. One possible reason for this could be their overall satisfaction with the existence of such services, which enable them to seek help for their health issues without having to converse over the phone in a language that is not their native tongue.

One notable distinction between the user groups pertained to their preference for using either a computer or a mobile device. It was evident that older adults favored using computers, likely because of their larger screens and the familiarity that older adults have with them. Conversely, most nonnative Finnish speakers showed a preference for mobile devices, with some noting that they solely rely on their mobile devices and do not even own a computer. This preference may be influenced in part by financial constraints, which limit the number of devices a person can afford. Additionally, in our sample, older adults encountered fewer difficulties with processing long pieces of text compared with the other groups.

The promotion of online symptom checkers as a means to decrease unnecessary clinic visits [13] underscores the importance of ensuring they do not inadvertently increase contact with health care staff. Therefore, greater attention should be directed toward enhancing the cognitive accessibility of these tools, thereby enabling a wider range of users to use them effectively. In this study, users’ incomplete understanding of the questions or answer options led them to select additional symptoms, resulting in more serious care recommendations and advising users to seek emergency health care.

In ensuring the cognitive accessibility of eHealth services, it is imperative to involve vulnerable groups in testing. Testing with vulnerable groups provides valuable insights. First, it emphasizes the need for well-planned test sessions with a manageable number of tasks. This approach ensures that participants can fully engage and provide meaningful feedback without being overwhelmed. All of these groups required considerable time to complete the test tasks, with most participants unable to finish both planned tasks with either service. Moreover, they necessitated more detailed instructions and support during the test sessions, as many participants within these groups were not at ease with using eHealth services.
Based on the findings of this study and as supported by the broader universal design literature [5], several design guidelines can be outlined. Foremost among these is the emphasis on clarity. (1) The options provided to the user should be clear and understandable. The user should understand what the differences between different options are and what actions are available for them. (2) It should be made clear to the user where they should be focusing on. This is particularly important in services that contain a lot of information and options. (3) Long or uncommon words and difficult compound words should be avoided. This is especially relevant in health-related terminology, as the user might not understand the special terms and might confuse different terms. (4) Navigating the services should be easy and effortless. The user should be presented with as few options as possible, and excessive scrolling should be minimized. This is because the user may inadvertently overlook relevant information.

**Limitations**

There are, naturally, some limitations to this study. First, the sample size of 13 participants was rather small, albeit quite typical for this type of qualitative study [43]. However, given the diverse nature of the user group and potential challenges related to cognitive accessibility, a more diverse participant pool could have been beneficial. Specifically, a wider age range of older adults could have been tested, considering their versatility as a group. Additionally, nonnative Finnish speakers could have been recruited from a more geographically diverse range of countries of origin. Moreover, testing should involve other diverse human characteristics, such as neurodiversity (including conditions such as attention-deficit/hyperactivity disorder, attention-deficit disorder, and various forms of autism). Given society’s rapid transition toward digitalized services, it is crucial to broaden the scope to include other groups at risk of digital exclusion.

Another limitation of this study is its focus on only 2 online symptom checkers. While the range of available online symptom checkers is already extensive, it is important to include testing of other eHealth services designed for use by all citizens. Additionally, this study only examines a limited list of symptoms and assesses usage on a 1-time or 2-time basis.

In conclusion, we recommend conducting testing with a more diverse user group, with a specific focus on accessibility and cognitive accessibility. Additionally, adopting a broader test setup that encompasses a wider range of symptoms and includes other eHealth services intended for broad usage would be beneficial.

**Comparison With Prior Work**

Usability issues were efficiently identified during testing with special user groups. In a study by Liu et al [36], which involved 350 participants, similar problems were discovered with service A as found in our study. The authors observed comparable challenges related to understanding questions and terminology, along with a need to enhance the visual layout and instructions for users. However, a notable disparity was observed in completion times: their participants completed the symptom checkers in an average of 4 minutes and 9 seconds, whereas users in our study required, on average, 3 times longer. In addition to uncovering issues that notably impact cognitive accessibility, our study identified similar usability problems as other assessments. Furthermore, as highlighted by Jormanainen et al [23], the same service was used over 1.5 million times for COVID-19 evaluation, suggesting its successful use by a vast number of users. Moreover, challenges with terminology have been recognized in other services [20].

This study has concentrated on cognitive accessibility with 3 distinct user groups. Comparable user groups have been used in other studies that center on eHealth services [1,29,30]. Upon comparing our findings with these studies, we observe that the necessity for clearer language and terminology, along with the clarity of the service, has previously been recognized through interviews and focus groups [1,29]. Our study provides more nuanced insights into how these issues manifest in practical usage.

**Conclusions**

In this study, we conducted a qualitative usability evaluation of 2 online symptom checkers, with a particular emphasis on the cognitive accessibility of the services. The evaluation targeted potentially vulnerable groups at risk of digital exclusion. Three distinct user groups participated in the tests: older adults, individuals with mild intellectual disabilities, and nonnative Finnish speakers. Our findings revealed that these groups encountered numerous difficulties with the tested services, particularly concerning their clarity and the language/terminology used. Furthermore, when testing with these groups, several key points must be considered: test sessions should be meticulously planned, instructions need to be clear, sessions should not be overly prolonged, and sufficient time must be allocated for each task.

In general, we found that testing with vulnerable groups was both useful and efficient. The rate of usability problems identified was notably high compared with the number of participants, and these issues were readily uncovered. These user groups encountered similar challenges related to information processing. It is imperative to provide them with better support through services that are clear, presenting less information and fewer options at once, and incorporating fewer long and complex words and selection lists. Additionally, following the principles of universal design, the proposed improvements are such that they will also benefit a more general user group. Therefore, we highly recommend testing with potentially vulnerable groups and, furthermore, expanding the user groups to include a representation of a broader variety of cognitive characteristics and challenges.
Acknowledgments

This research was conducted as part of the “Toward a Socially Inclusive Digital Society: Transforming Service Culture” project (DigiIN) and was funded by the Strategic Research Council at the Academy of Finland (grant numbers 352501and 352503) and NordForsk (project 100477). We thank the service providers DigiFinland Oy and Klinik Healthcare Solutions for allowing us to test the services, and Selkeistä meille initiative (a joint project between The Finnish Association on Intellectual and Developmental Disabilities [FAIDD] and Kehitysvammatuki 57 ry) and Välär ry for helping us to contact the test users. Finally, we thank Felipe González Carceller for his excellent contribution to the research.

Data Availability

The data sets generated during or analyzed during this study are not publicly available due to the sensitive nature of the data but the numeric data are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

The distribution of the symptoms and services.
[DOCX File, 17 KB - humanfactors_v11i1e45275_app1.docx ]

Multimedia Appendix 2

Symptom vignettes used.
[DOCX File, 16 KB - humanfactors_v11i1e45275_app2.docx ]

Multimedia Appendix 3

The background information of the participants. All the participants used digital services multiple times a day. MID: mildly intellectually disabled.
[DOCX File, 15 KB - humanfactors_v11i1e45275_app3.docx ]

Multimedia Appendix 4

Usability problem categories.
[DOCX File, 15 KB - humanfactors_v11i1e45275_app4.docx ]

Multimedia Appendix 5

Average SUS scores for both services. For older adults the two services got the same results, for the two other groups Service A got a slightly better score.
[DOCX File, 13 KB - humanfactors_v11i1e45275_app5.docx ]

References


40. Omaolo. DigiFinland Oy. URL: https://omaolo.fi [accessed 2022-07-10]


Abbreviations

SUS: System Usability Scale
UI: user interface
The Role of Coherent Robot Behavior and Embodiment in Emotion Perception and Recognition During Human-Robot Interaction: Experimental Study

Laura Fiorini1,2, PhD; Grazia D'Onofrio3, PhD; Alessandra Sorrentino1, PhD; Federica Gabriella Cornacchia Loizzo2, MSc; Sergio Russo4, MSc; Filomena Ciccone3, PsyD; Francesco Giuliani3, MSc; Daniele Sancarlo5, MD; Filippo Cavallo1,2, PhD

1Department of Industrial Engineering, University of Florence, Firenze, Italy
2The BioRobotics Institute, Scuola Superiore Sant'Anna, Pontedera (Pisa), Italy
3Clinical Psychology Service, Health Department, Fondazione IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo (Foggia), Italy
4Innovation & Research Unit, Fondazione IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo (Foggia), Italy
5Complex Unit of Geriatrics, Department of Medical Sciences, Fondazione IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo (Foggia), Italy

Corresponding Author:
Laura Fiorini, PhD
Department of Industrial Engineering
University of Florence
Via Santa Marta 3
Firenze, 50139
Italy
Phone: 39 0552758663
Email: laura.fiorini@unifi.it

Abstract

Background: Social robots are becoming increasingly important as companions in our daily lives. Consequently, humans expect to interact with them using the same mental models applied to human-human interactions, including the use of cospeech gestures. Research efforts have been devoted to understanding users’ needs and developing robot’s behavioral models that can perceive the user state and properly plan a reaction. Despite the efforts made, some challenges regarding the effect of robot embodiment and behavior in the perception of emotions remain open.

Objective: The aim of this study is dual. First, it aims to assess the role of the robot’s cospeech gestures and embodiment in the user’s perceived emotions in terms of valence (stimulus pleasantness), arousal (intensity of evoked emotion), and dominance (degree of control exerted by the stimulus). Second, it aims to evaluate the robot’s accuracy in identifying positive, negative, and neutral emotions displayed by interacting humans using 3 supervised machine learning algorithms: support vector machine, random forest, and K-nearest neighbor.

Methods: Pepper robot was used to elicit the 3 emotions in humans using a set of 60 images retrieved from a standardized database. In particular, 2 experimental conditions for emotion elicitation were performed with Pepper robot: with a static behavior or with a robot that expresses coherent (COH) cospeech behavior. Furthermore, to evaluate the role of the robot embodiment, the third elicitation was performed by asking the participant to interact with a PC, where a graphical interface showed the same images. Each participant was requested to undergo only 1 of the 3 experimental conditions.

Results: A total of 60 participants were recruited for this study, 20 for each experimental condition for a total of 3600 interactions. The results showed significant differences ($P<.05$) in valence, arousal, and dominance when stimulated with the Pepper robot behaving COH with respect to the PC condition, thus underlying the importance of the robot’s nonverbal communication and embodiment. A higher valence score was obtained for the elicitation of the robot (COH and robot with static behavior) with respect to the PC. For emotion recognition, the K-nearest neighbor classifiers achieved the best accuracy results. In particular, the COH modality achieved the highest level of accuracy (0.97) when compared with the static behavior and PC elicitations (0.88 and 0.94, respectively).
Conclusions: The results suggest that the use of multimodal communication channels, such as cospeech and visual channels, as in the COH modality, may improve the recognition accuracy of the user’s emotional state and can reinforce the perceived emotion. Future studies should investigate the effect of age, culture, and cognitive profile on the emotion perception and recognition going beyond the limitation of this work.

(JMIR Hum Factors 2024;11:e45494) doi:10.2196/45494

KEYWORDS
social robot; emotion recognition; human emotion perception; human-robot interaction; robot cospeech gestures evaluation

Introduction

Background

During the last decade, there has been increasing interest in research on socially assistive robotics aimed at realizing intelligent robotic solutions for health care and social assistance. We experience an evolution of social robot applications; indeed, they moved from the role of concierge and helper [1] toward the role of companion and therapist [2,3]. Social robots have the potential to contribute to the greater good of society; indeed, it has been demonstrated that they can support everyday life as companions and the health care system from logistics to assistance and rehabilitation. Thinking to include social robots in the care chain, they can be used to reduce stress, anxiety, and pain in children [4]; they can be integrated into conventional behavioral and cognitive therapies for both children and adults who struggle with social anxiety [5]; or they can be used to promote mental health [6]. A review by Hung et al [7] showed evidence that Paro robots can reduce negative emotions in patients, promoting a positive mood and improving social engagement. Rossi et al [8] demonstrated that social robots are effective in decreasing stress in children accessing the emergency room. As the complexity of the robot task increases, social robots are required to perform more complex perceptual, cognitive, and interactive functionalities. This is the case in long-term interactions in which robots and users should establish meaningful communication, emotional awareness, and reliable engagement.

In this context, the human-robot interaction (HRI) field has become crucial, and it is now compelling to better understand how humans perceive, interact with, or accept these machines in social and real contexts. Researchers are also debating on defining the factors that can influence the perceived social capabilities and intelligence of a robot [9,10]. De Graaf et al [11] highlighted the significance of the robot’s social capability, emphasizing the importance of 2-way interaction where a robot is expected to respond to humans in a social manner. In addition, De Graaf et al [11] underlined that a social robot should also display thoughts and feelings and should be socially aware of the environment, among other issues. When a robot failed to perform this 2-way interaction, people were disappointed and experienced a sense of dissonance. In other words, when interacting with a social robot, especially a humanoid robot, we expect to use the same mental structure and social rules that guide us in human-human communication, expecting empathetic interaction because they are perceived as social actors [12].

From a roboticist or engineering point of view, these concepts are translated into the design and development of behavioral models that can guarantee an efficient and reliable 2-way interaction [13,14]; they should perceive and show emotions (and social norms) and thus be understood by humans with whom they are interacting. The key challenge in this field is to provide robots with cognitive and affective capabilities, developing architectures that allow them to establish empathetic relationships with users, which can foster long-term and meaningful interactions. From an implementation perspective, the design and the deployment of a socially capable social robot comprises 2 essential parts. The first is devoted to designing and implementing a consistent and congruent emotional behavioral architecture that makes the robot react or act to the environment (ie., display thoughts and feelings). The capabilities of a user to understand the emotions displayed by a robot have been explored in different settings [15,16]. Examples of actions can include the expression of congruent cues such as facial expressions [17], changes in the color of the eyes, movement of the upper limbs [16,18], or smart navigation strategies [19]. In contrast, the other part is more focused on the robot’s perception of the user’s emotional response to these behaviors [20], with special attention to contextualizing its action and reaction according to the living contexts and habits or preferences of the person with whom it is interacting (ie, being socially aware of the environment) [21].

Related Work on Emotion and Social Robots

The ability of a robot to perceive the nonverbal cues of the user, which convey user emotion and intent, plays a key role in the development of social robots capable of performing meaningful interactions [22,23]. In this sense, humans’ gaze, body posture, cospeech gestures, and facial expressions play a leading role in defining the context of the interaction, helping the robot to correctly classify the experience, and associating it with informative content [21]. The development of such abilities, for a researcher in the field of robotics, translates into the use of multimodal sensor modality and the implementation of several complex algorithms to endow robots with different cognitive and social capabilities. The visual modality is the most commonly used [24] because it can detect nonverbal behaviors that are representative of the emotional state of users without requiring them to wear any external sensor. Alternatively, wearable sensors [25] can be used, also using a multimodal approach, to overcome the problems related to occlusion and low light. Other algorithms or modules were implemented to perform multiperson tracking [26], speech recognition [27,28], and automatic engagement detection [29]. A recent review paper [24] provides a deep insight into the most used methods and approaches.
For the showemotion part, robots must exploit several channels (ie, auditory, visual, cospeech, and gestures) and mechanisms (eg, body posture, facial expressions, vocal prosody, touch, and gaze) to communicate their “internal emotional status” and intentions authentically and clearly [30]. Thus, the capabilities of a user to understand the emotions displayed by a robot have been explored in several settings [31]. Over the last few years, several attempts have been made using both video-simulated robots and real robots. Guo et al [20] showed participants 5 different emotions using the humanoid robot called Alpha2, and they were asked to rate the perceived emotion using the Self-Assessment Manikin questionnaire (SAM; only valence and arousal dimensions) [32]. In contrast, Barchard et al [33] conducted a web-based study to evaluate the perception of a robot’s social intelligence by showing videos of robot interactions. However, the embodiment and the appearance of social robots play important roles in the perception of the robot; therefore, video-based elicitation could introduce some bias in the analysis of perceived emotion. This is why other research has relied on investigating the emotion perceived during a real HRI. This is the case of Bagheri et al [34], who asked participants to watch 6 performances of America’s Got Talent Show on Pepper’s tablet that are expected to evoke the 6 basic emotions. Rossi et al [35] and Staffa et al [36] relied on movie trailers to evoke emotions. However, they used nonstandard videos, making it challenging to identify the target emotion in a recognized and standardized manner, as the elicited emotion through the video clip is not known a priori, and consequently, it is difficult to define the role of the robot (and its embodiment) in the elicitation process.

Research groups have recently begun to study the effects of multimodal channels on communication. Studies conducted with embodied conversational agents showed that incongruent emotional stimuli (eg, auditory and visual stimuli) can result in adverse consequences on user rating; conversely, congruent stimuli can facilitate the recognition of emotions [37]. Other researchers have also studied the role of nonverbal behavioral cues while interacting with robots. Movie clips showing coherent and incoherent robot behaviors are often used to elicit emotional responses from users with respect to those induced by movie clips [15,16,18,35]. For instance, Rossi et al [16] investigated how an incoherent nonverbal robot’s behavior with respect to the presented emotion can produce a type of humorous effect. Tsiouri et al [18] investigated how contextual incongruence (ie, a robot’s reaction conflicts with the socioemotional context) can confuse the observers, decreasing the accuracy of the perceived emotion. Nevertheless, such a cospeech robot’s behavior was used in addition to a nonstandard method of emotion elicitation, as previously remarked; thus, it is not easy to understand the role of the robot’s behavior with respect to the emotional context. Therefore, it is important to understand how the robot’s nonverbal behavior might shape the human perception of the showed emotion elicited through standard emotionally labeled visual data sets and, at the same time, observe the robot’s emotion recognition accuracy rate. Although previous studies have shown a correlation between the robot’s nonverbal action and perceived emotion, there is a lack of use of standard elicitation modalities.

Therefore, in this work, we present the results of 3 experimental sessions to observe the performance of the robot in recognizing users’ emotions as well as to investigate the difference (if any) in eliciting emotions in humans when using a social robot (with or without coherent behavior) rather than a PC. We plan to use a standard data set of pictures, namely, the International Affective Picture System (IAPS) [38], to elicit emotions in users. Particularly, the robot will use a multimodal behavior (ie, head movements, vocal reinforcement, and body gestures) to interact with the participants while showing the graphical emotions by establishing social binding, whereas the PC will provide emotion elicitation only through a graphical interface. The 2 graphical interfaces have been designed to provide the same information to the user but using different communication channels. In this context, the aim of this work is dual. First, it aims to investigate the increase in the user’s emotional perception during the interaction with a robot with respect to a PC (Figure 1, blue arrow). In particular, this work investigates the role of the robot’s coherent nonverbal behavior in emotion perception by consequently assessing the impact of robot embodiment and, eventually, its coherent behavior. Robot nonverbal cues are manipulated with respect to a mapping between the main associated emotion and cospeech gestures that can be generated on the robot. At the end of each interaction, the participants were asked to self-assess their perceived emotions. In this study, we used the emotion classification proposed by Russel et al [39], which relies on 3 variables, namely, valence, arousal, and dominance. Valence describes the degree to which a stimulus causes a positive or negative emotion, arousal refers to the intensity or level of energy invested in the emotion, and dominance reflects the extent of perceived control over the emotional response when facing the stimulus. The collected answers were analyzed to answer the following research questions (RQs):

1. RQ1: Emotion elicited through a humanoid robot interacting with coherent emotional behavior is rated higher than emotions elicited by a web application in terms of emotional valence, arousal, and dominance.
2. RQ2: There are significant differences in terms of emotional valence, arousal, and dominance between a robot showing coherent behavior rather than a robot that it is not moving at all (static condition).
3. RQ3: The embodiment of the humanoid robot will not affect the emotion perception compared with the web application.

Second, this study aims to assess the accuracy of the robot in recognizing the elicited emotion in the participants (Figure 1, yellow arrow). The ability to infer and interpret emotions plays a key role in establishing intuitive and engaging HRIs. On the one hand, a robot endowed with emotion recognition skills can adapt its behavior based on the detected user emotion [22]. On the other hand, a robot expressing recognizable emotions positively influences the evaluation of its capabilities [40]. In particular, features related to facial expressions were extracted, preprocessed, and analyzed with 3 supervised machine learning techniques to verify the following RQ:

1. RQ4—There is no difference in the robot emotion recognition accuracy despite the elicitation modalities (robot or web application).
Figure 1. Two-way interaction proposed in this study. To improve the human-robot interaction, the robot should perceive the user’s behavior (yellow arrow) and plan appropriate action (blue arrow).

In our previous studies [41,42], we evaluated the perceived acceptance and the recognition rate of having a robot that acts coherently and incoherently despite the standard emotion showed with respect to the standard elicitation modality. In contrast, in this study, we focus only on coherent behavior by comparing it with a standard web application that runs on a PC. In addition, instead of focusing on evaluating how the robot’s acceptance is modulated according to the elicitation modality, we focused on the perceived emotion evoked.

Methods

Instrumentation
The instrumentation is composed of the following elements: (1) a Pepper robot (Aldebaran, United Robotics Group) or a PC, (2) the RoboMate (Behaviour Labs) interface for cospeech gestures, (3) a custom interface that contains pictures from the IAPS for eliciting emotion, and (4) an external camera placed on Pepper to record the participants’ emotions during the interaction. Pepper is a humanoid robot that is widely used for experimentation in socially assistive robotics. It is 120 cm tall, weighs 28 kg, and has 20 df, including 1 head, 2 arms, and 1 wheeled base. In addition, it has a tablet on the front.
coherent behavior was managed through the RoboMate interface [43] to animate Pepper, when necessary, selecting among the behaviors classified as “positive social stimulus” or “negative social stimulus.” The selected stimulus was modeled by a psychologist using 3 modalities: body gestures (upper limb and head), gaze, and sound. IAPS is a database of images devoted to eliciting standardized emotions [44]. It was developed by the Center for Emotion and Attention at the University of Florida. This database is commonly used in psychological studies on emotions and attention. Each image in the data set is labeled with the corresponding emotion, thus enabling researchers to properly select the stimulus. In this study, 60 images were selected from the team of psychologists of the hospital “Casa Sollievo della Sofferenza.” According to the IAPS valence dimension, 21 of the selected images were rated as positive, 19 as negative, and 20 as neutral. A customized web-based interface was developed to standardize the emotional stimulation when using 2 different communication channels (a robot and a PC).

Experimental Setup

A psychologist welcomed the participant, briefly explaining the experimental setup, including how to use the evaluation tool. It is important to emphasize that the participant was not aware of the real objective of the experimentation, thus avoiding interference with the experience. To properly investigate the RQs, each participant underwent 1 of the following elicitation modalities.

1. Static (STA) behavior: Pepper robot has its arms along the body in a neutral position (Figure 2A). Pepper’s face was looking at the participant but without any animacy. Pepper displayed IAPS images on its tablet through the customized web application.

2. Coherent (COH) behavior: Similar to the STA condition, the IAPS images were shown on Pepper’s tablet. Using the RoboMate application, the psychologist assigned a coherent behavior to Pepper with the shown images. In particular, the psychologist can choose and combine 3 modalities for elicited emotions: body gesture (upper limb and head), gaze, and sound, which are available on the RoboMate application (Figure 2B). For example, in the case of positive emotion, Pepper’s gestures were chosen to look friendly; it should look to the user direction, and the voice gave positive reinforcements.

3. PC: For this experimental condition, we used a PC instead of the Pepper robot. Participants were asked to evaluate the images shown on a PC through the customized web application.

Figure 2. Experimental setup. The participants were interacting with Pepper robot during the experimentation. (A) Participants were asked to sit in front of the robot and watch the images on its tablet. (B) If the participant belonged to the coherent elicitation modality group, the Pepper robot would move its arms, eyes, and head.

The participant was asked to sit in front of the technology (ie, Pepper robot or PC). If the user interacts with Pepper, Pepper is placed 0.5 to 0.6 m far from the user (ie, personal distances [45]); in the case of interaction with the PC, the user is requested to sit and interact with the computer as he or she will commonly do.

Each stimulus was shown for 7 seconds, and at the end, the participant was asked to fill out the SAM [32], as adapted in the study by Gatti et al [46] directly on the robot or on the computer after each picture. SAM is an emotion assessment tool that uses graphic scales, depicting cartoon characters expressing 3 emotional elements (valence, arousal, and dominance). Each participant was asked to rate the domains by selecting an image that corresponded to a score between 1 and 9. A picture of the interface is presented in Multimedia Appendix 1.

At the end of the experimental session, each participant completed 60 SAM questionnaires. The psychologist was present during the test, and she or he was ready to intervene in case of necessity. All the tests were performed at the “Casa Sollievo della Sofferenza” research hospital.

Ethical Considerations

The approval of the study for experiments using human participants was obtained from the local Ethics Committee on Human Experimentation (register code 3038/01DG). All
participants signed an informed consent form before participating in this study, and pictured participants provided written informed consent to allow their image to be published. The data were pseudoanonymized and stored on a GDPR-compliant server.

Participants
Participants were recruited from July 2020 to February 2021 from employees and staff of the “Casa Sollievo della Sofferenza” research hospital located in Apulia (San Giovanni Rotondo, Foggia) using convenience sampling. Participants were excluded if they had a hearing or visual impairment. Recruited participants were then randomly assigned to undergo 1 of the 3 experimental conditions (ie, STA, COH, and PC). Sociodemographic information (age, education, and sex) was collected to verify the similarities between the groups.

Data Analysis

Overview
Owing to the sample size of each cohort, the nonparametric statistic was used, particularly the Kruskal-Wallis test and chi-square test, to investigate significant differences between participants’ groups in terms of age, sex, and educational level. The significance level was set at \( P = .05 \). The following paragraphs describe the analysis performed on the SAM questionnaires and the data collected from camera sensors.

Emotion Perception Analysis
A total of 60 SAM questionnaires were collected for each participant. The average values of the valence, arousal, and dominance domains were computed for each selected image of each group of elicitation modality (ie, STA, COH, and PC). Differences were analyzed with the Kruskal-Wallis test \( (P<.05) \) and post hoc evaluated with the Mann-Whitney \( U \) test (with Bonferroni correction) used to identify between which pair of elicitation modes the difference has occurred.

Emotion Recognition Analysis
Data from the camera were processed and examined offline. The recordings were initially analyzed [47] to ensure that only the frames featuring the face of the person performing the test were included in the study. Then the recordings were segmented, providing short videos that corresponded to the user’s reaction to each image proposed, totaling 60 videos per user. The OpenFace toolkit [48] was used to extract 150 features related to gaze and facial expression from each video as well as the quality (ie, confidence) of the extracted features. The data were filtered according to the confidence score (frames with a confidence score <0.90 were discarded). The data were then labeled based on the IAPS-defined emotions (ie, positive, negative, and neutral). Data were normalized and selected. Only features with a correlation coefficient of <0.85 were picked from the initial data set, avoiding those with a high correlation coefficient (which may represent redundant information). The data of the merged data set were then separated into sub-data sets (one for each participant), and emotion classification was performed using the selected features. In this study, we rely on state-of-the-art methods used for emotion recognition [24] to facilitate a comparison with other works. The 3 supervised classifiers used are support vector machine (SVM), random forest (RF), and K-nearest neighbor (KNN). To classify the data by participant, a 10-fold cross-validation procedure was applied, and the outputs were organized in a confusion matrix. The classification performance was assessed in terms of accuracy, precision, recall, and F-measure [49]. The calculations were computed using MATLAB 2020a. More details on emotion recognition analysis are available in Multimedia Appendix 2 [24,47-49].

Results

Description of the Participant Cohort
A total of 60 participants were involved in this study, 20 for each modality, resulting in 3600 interactions with technologies. In total, 3 participants were excluded from the analysis of perceived emotion because not all SAM evaluations were correctly saved after each elicitation. In case of missing SAM values, these ratings were removed from the analysis of average values. Finally, 57 participants were included in these subgroups of analyses linked to RQ1, RQ2, and RQ3. Regarding the recognition of emotion using machine learning techniques (linked to RQ4), a total of 53 participants were included in the analysis. A total of 7 participants were excluded because of technical problems related to the quality of the recorded images. The statistical tests did not indicate any difference between the 3 participant cohorts regarding age, sex, and educational level. The participant demographics and educational analyses are reported in Multimedia Appendix 3.

Participants’ Perceived Emotion Results
The results underline significant differences \( (P<.001) \) in the perceived emotions according to the different elicitation modalities, except for the arousal elicited with the positive images (Figure 3). The median and IQR values are fully reported in Multimedia Appendix 4. As for valence, the robot with coherent behavior elicited significant differences \( (P<.001) \) and higher values in terms of valence, arousal, and dominance domains compared with the other 2 modalities for negative and neutral emotions. In terms of negative valence, the participants perceived fewer negative emotions with the coherent robot than with the other 2 modalities. For positive valence, elicitation with the web application is significantly different from that with the robot \( (P<.001) \).
Regarding arousal, the coherent robot was rated higher than the other 2 modalities, but there were significant differences ($P<.001$) only for negative and neutral emotions, whereas for positive arousal, the results, depicted in Figure 3, highlight only a trend. All the $P$ values are reported in Multimedia Appendix 4.

The participants stimulated using the robot rated significantly higher dominance across all 3 emotions rather than the cohort that used the PC in the test. As for positive elicitation, we found significant differences ($P<.001$) between the cohort stimulated with the PC and those stimulated with the robot (ie, static behavior and coherent behavior). Indeed, the participants rated the emotions (in terms of valence and arousal) elicited by the robot more than the ones elicited using the PC. All $P$ values are reported in Multimedia Appendix 4.

**Robot’s Emotion Recognition Results**

Because of technical issues 1848 frames pertaining to the PC modality were removed from the analysis during the preprocessing. At the end, the total number of samples included in this study was 296,677 for the STA modality, 228,170 for the COH modality, and 103,758 for the PC modality. The number of columns in each data set corresponded to the number of features selected using the correlation analysis method. The following features were selected (Figure 4):

1. The x-, y-, and z-coordinates of the eye gaze direction vector for eye 0 (3 features).
2. The z-coordinate of the eye gaze direction vector for eye 1 (1 feature).
3. The x- and y-coordinates of the location of the landmark 8 (the leftmost in the image) of the eye 0 (2 features).

The 53 data sets were fed into 3 classifiers (SVM, RF, and KNN) [24]. The data sets were uniformly distributed across the 3 groups, as presented in Table 1.
Figure 4. Selected features. (A) Face and (B) eye landmarks extracted with OpenFace software. The landmark 8 in panel B was chosen after the feature selection.

Table 1. Distribution of data set instances.

<table>
<thead>
<tr>
<th>Group</th>
<th>Neutral, n (%)</th>
<th>Negative, n (%)</th>
<th>Positive, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static (n=296,677)</td>
<td>98,499 (33.2)</td>
<td>94,257 (31.77)</td>
<td>103,992 (35.03)</td>
</tr>
<tr>
<td>Coherent (n=228,170)</td>
<td>83,077 (36.41)</td>
<td>74,383 (32.6)</td>
<td>70,710 (30.99)</td>
</tr>
<tr>
<td>PC (n=103,758)</td>
<td>36,492 (35.17)</td>
<td>32,072 (30.91)</td>
<td>35,195 (33.92)</td>
</tr>
</tbody>
</table>

Accuracy, precision, F-measure, and recall were calculated as the mean values from the participants in the same experimental cohort. According to the findings, the KNN classifier offers the best classification results, with an accuracy of up to 0.88 for STA behavior, 0.97 for COH, and 0.94 for PC. The SVM classifiers, in contrast, had the lowest results (accuracy of up to 0.57, 0.67, and 0.68 for STA, COH, and PC, respectively); hence, they were excluded from further research. Compared with the RF classifier, the KNN classifier has the best F-measure (>0.88).

Table 2 presents the complete results for the KNN and RF classifiers, including the accuracy, F-measure, precision, and recall for each group. According to the overall trend, the COH modality achieves a high level of accuracy when compared with the STA and PC elicitations. In terms of the other indicators, the COH was better with the KNN classifier and slightly worse with the RF classifier when it came to elicitation with the PC.

Table 2. Performance of K-nearest neighbor (KNN) and random forest (RF) classifiers.a

<table>
<thead>
<tr>
<th>Group</th>
<th>Accuracy</th>
<th>Precision</th>
<th>F-measure</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KNN</td>
<td>RF</td>
<td>KNN</td>
<td>RF</td>
</tr>
<tr>
<td>Static</td>
<td>0.88</td>
<td>0.65</td>
<td>0.88</td>
<td>0.65</td>
</tr>
<tr>
<td>Coherent</td>
<td>0.97</td>
<td>0.73</td>
<td>0.96</td>
<td>0.72</td>
</tr>
<tr>
<td>PC</td>
<td>0.94</td>
<td>0.74</td>
<td>0.94</td>
<td>0.74</td>
</tr>
</tbody>
</table>

aMean values are used to calculate the results.

Confusion matrices (Figure 5) for the 3 elicitation modalities were generated to investigate the performance of the classifiers in recognizing the 3 selected emotions. The positive emotion was often better identified, whereas the negative emotion was the least recognized. When the user is stimulated with the robot with coherent modality and the PC, the RF classifier performs better than the KNN classifiers in distinguishing emotions. The KNN classifier appeared to perform better in the static modality than in the other 2.
**Discussion**

**Principal Findings**

The results confirm RQ1 (“A humanoid robot interacting with coherent emotional behavior is rated higher in terms of emotional valence, arousal, and dominance compared to the web application”) because the COH robot is rated significantly higher for all SAM dimensions (except positive arousal) with respect to the PC condition (Figure 3). However, it is worth noting that when speaking of negative elicitation, receiving a higher rating of valence means that the stimulus with the COH condition was perceived less negatively than the ones elicited with the others. RQ2 ("There are significant differences in terms of emotional valence, arousal, and dominance between the static robot compared to the robot that shows movement") is confirmed for the 3 dimensions for negative and neutral emotions (Figure 3). It is worth noting that these results confirm that the robot’s movements cause the negative emotion to be perceived as less negative (STA valence median value=3.32; COH valence median value=5.13). As for the positive emotion, there were no significant differences, which could suggest that the robot’s behavior per se did not affect the perception of the positive emotion.

The presented results did not confirm the RQ3 ("The embodiment of humanoid robot will not affect the emotion perception compared to the web application") for all elicited emotion and SAM constructs. Indeed, there were no significant differences between the STA and the PC elicitation for valence and arousal measured during negative and neutral elicitation (Figure 3). Conversely, COH and STA differed significantly from PC in terms of positive elicitation. These results suggest that robot embodiment per se has a role in the perception of dominance associated with negative and neutral emotions with respect to a standard web interface. On the contrary, as for the
positive emotion, embodiment seems to play a key role because both COH and STA elicitions differ from the web application in terms of valence and dominance. The ability to recognize user emotions is a fundamental step in the development of socially aware robots (RQ4). The emotions were recognized with an average accuracy >0.88 over the 3 elicitation conditions. In addition, the amount of gaze also depends on the interpersonal dynamics between the partner and their personalities and on the intent of using gaze to communicate their internal state. Therefore, it is important to measure it during interactions. As shown in Table 2, the accuracy of COH stimulation was higher than that of the other 2 methods. In addition, the results in the confusion matrices were aligned with the perceived emotion (Figure 5). According to the SAM results, the valence ratings for positive elicitation elicited with PC were significantly different from the other 2 with lower median values. This trend is reflected in the confusion matrices obtained using RF classifiers.

Comparison With Prior Work

Previous qualitative studies have pointed out how incoherent behavior can generate hilarious reactions in humans [16]. The presented results suggest that we can observe something similar, even if the stimulus is coherent. It appears that the robot’s behavior somewhat distracts from perceiving negative emotions, even if the behavior is aligned with the shown emotion. In addition, as confirmation, positive emotion was perceived significantly more positively than the PC modality, suggesting that robot movements make the robot more positive. Consequently, these results suggest that it is important to tailor the reaction of the robot appropriately to elicit a specific emotion. Indeed, if we need to stimulate—for a certain reason—negatively the users, we need to reduce the robot’s body expression because they can decrease the perception of negative emotions. Alternatively, if we need to provide positive feedback to users, the combined actions of both verbal and nonverbal communication can be used.

A previous study [36] compared robots and web applications that focused on investigating preferences and acceptance, and they did not find any significant deviation in the quantitative results. In contrast, in this study, we focus on human emotion perception, and this perception seems to be influenced or biased by the emotion itself and the robot’s movement. This finding highlights the significance of not just robot embodiment but also its co-speech gestures in designing social agents, particularly when evaluating all dimensions of emotions. Methodologically, the presented findings carry significant implications for the design of experimental protocols. Evaluating HRI cannot rely solely on videos, as they overlook the importance of physical interaction. In the literature, some papers [33] provide a user impression without direct interaction with a robot; the collected results can be biased because the participant missed the contribution in the perception related to embodiment. Take, for instance, the scenario where you are testing a new game application or software on a tablet meant for eventual integration into a robot. Particularly when assessing emotions, it’s crucial to approach the generalization of results with caution. In this sense, the result could be altered because the emotions elicited could not be directly applicable when interacting with an embodied agent.

The results obtained for the STA robots with the KNN and RF classifiers were slightly improved with respect to the results obtained in our previous work [42] (average accuracy was equal to 0.85 with KNN and 0.98 with RF), where we used them in combination with encoders. It is also worth noting that after the feature selection process, only the features related to gaze were retained in the analysis. Gaze is extremely important in managing interpersonal interaction and also during human-robot conversation; indeed, it can be correlated with user engagement during conversation or mutual tasks [50,51].

Limitations of the Study

The limitations of this study were mainly related to the cohort of recruited participants. First, both cognitive and cultural backgrounds are factors that can influence the perception of emotions [52]. Some neurological pathologies (eg, Parkinson disease) can affect facial expressions, whereas others can affect body gestures and language (eg, autism spectrum disorders and apathy); consequently, emotion recognition accuracy in such cases can change. The RQs do not focus on investigating their role in emotion perception; consequently, we recruited cohorts of people comparable for cultural background and cognitive status to limit the impact of these factors. The second limitation of this study refers to how the emotion is evaluated; in this study, we evaluated each SAM dimension separately. The third limitation of this study relies on the supervised machine learning techniques used. In this study, we rely on standard supervised methods because our main RQs are not focused on learning methods; therefore, we apply the most used techniques.

Future Directions

In this context, by applying the findings and implications of this paper in the health care context, we can conclude that it is important to tailor the reaction of the robot properly; indeed, if we need to stimulate—for a certain clinical reason—the users negatively, we need to reduce the robot’s body expression because they can decrease the perception of negative emotions. Alternatively, if we need to give positive feedback to the users, for instance, during an exercise, we can use the combined action of both verbal and nonverbal communication. To overcome the limitations of this study, future research can be planned to extend the study to include a different group of participants with some cognitive and physical disorders and different cultural backgrounds to evaluate the effect of these factors on emotion perceptions. Future studies should also investigate whether there are differences in combining valence-arousal domains, as proposed in other studies [16,53]. Finally, the data could be analyzed using also deep learning and reinforcement learning techniques.

Conclusions

This study aimed to investigate the role of robot embodiment and its behavior in emotion perception and recognition using a standard elicitation model. In total, 4 RQs were investigated to understand how the robot’s nonverbal behavior might shape the human perception of the showed emotion elicited through a standard data set and, at the same time, to observe the robot’s
emotion recognition accuracy rate. This study presents an
experimental setup in which 60 participants were asked to
interact with 2 embodied agents (ie, a robot or tablet) that acted
as emotion facilitators by showing them 60 standard pictures.
The results underline the good recognition accuracy of the
perception modules of the robot. Indeed, we can correctly
classify the valence of the emotion (ie, positive, neutral, and
negative) with an accuracy of up to 0.97 in the best case.
According to the results, robot embodiment affects the
perception of dominance significantly compared with web
applications, which means that participants’ emotions were less
controlled when they were interacting with an embodied agent.

Acknowledgments
The authors would like to thank all the people involved in the study. This study was funded by “An adapted behavioral robot
model with advanced cognitive/physical interaction capabilities for assessment and rehabilitation of neurodegenerative diseases
(DESTINI)” founded by Unione Europea—NextGenerationEU (CUP: B55F21007810001).

Data Availability
The data sets generated and analyzed during this study are not publicly available because of the scope of the consent signed by
the patient participating in the study but are in part (no video recordings) available from the corresponding author on reasonable
request.

Authors’ Contributions
The conceptualization was done by LF, GDO, and F Cavallo. Data curation was conducted by GDO, F Ciccone and AS. LF,
FGCL, and AS were responsible for the data analysis. LF acquired the funding. The methodology was developed by GDO, LF,
F Cavallo, and FG. The investigation was carried out by GDO and F Ciccone. AS and SR handled the software. DS, FG, and F
Cavallo provided supervision. LF was responsible for the original draft of writing, while all authors contributed to the writing,
review, and editing.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Self-Assessment Manikin questionnaire.
[PDF File (Adobe PDF File), 134 KB - humanfactors_v11i1e45494_app1.pdf ]

Multimedia Appendix 2
Emotion recognition analysis.
[PDF File (Adobe PDF File), 194 KB - humanfactors_v11i1e45494_app2.pdf ]

Multimedia Appendix 3
Participants’ description.
[PDF File (Adobe PDF File), 120 KB - humanfactors_v11i1e45494_app3.pdf ]

Multimedia Appendix 4
Median and IQR values computed for each elicited emotion.
[PDF File (Adobe PDF File), 143 KB - humanfactors_v11i1e45494_app4.pdf ]

References
doi: 10.2196/jmir.9460 [Medline: 30249588]
10.1007/s12369-019-00568-1]
10.1007/s12369-015-0284-0]


Abbreviations

COH: coherent
HRI: human-robot interaction
IAPS: International Affective Picture System
KNN: K-nearest neighbor
RF: random forest
RQ: research question
SAM: Self-Assessment Manikin questionnaire
STA: static
SVM: support vector machine
Original Paper

Physical Therapists’ Acceptance of a Wearable, Fabric-Based Sensor System (Motion Tape) for Use in Clinical Practice: Qualitative Focus Group Study

Audrey Lee, MS; Patricia Dionicio, MPH; Emilia Farcas, PhD; Job Godino, PhD; Kevin Patrick, MD, MS; Elijah Wyckoff, MS; Kenneth J Loh, PhD; Sara Gombatto, PT, PhD

Corresponding Author:
Sara Gombatto, PT, PhD
School of Exercise & Nutritional Sciences
College of Health & Human Services
San Diego State University
5500 Campanile Drive
San Diego, CA, 92182-7251
United States
Phone: 1 619 594 3288
Email: sgombatto@sdsu.edu

Abstract

Background: Low back pain (LBP) is a costly global health condition that affects individuals of all ages and genders. Physical therapy (PT) is a commonly used and effective intervention for the management of LBP and incorporates movement assessment and therapeutic exercise. A newly developed wearable, fabric-based sensor system, Motion Tape, uses novel sensing and data modeling to measure lumbar spine movements unobtrusively and thus offers potential benefits when used in conjunction with PT. However, physical therapists’ acceptance of Motion Tape remains unexplored.

Objective: The primary aim of this research study was to evaluate physical therapists’ acceptance of Motion Tape to be used for the management of LBP. The secondary aim was to explore physical therapists’ recommendations for future device development.

Methods: Licensed physical therapists from the American Physical Therapy Association Academy of Leadership Technology Special Interest Group participated in this study. Overall, 2 focus groups (FGs; N=8) were conducted, in which participants were presented with Motion Tape samples and examples of app data output on a poster. Informed by the Technology Acceptance Model, we conducted semistructured FGs and explored the wearability, usefulness, and ease of use of Motion Tape for PT management of LBP. FG data were transcribed and analyzed using rapid qualitative analysis.

Results: Regarding wearability, participants perceived that Motion Tape would be able to adhere for several days, with some variability owing to external factors. Feedback was positive for the low-profile and universal fit, but discomfort owing to wires and potential friction with clothing was of concern. Other concerns included difficulty with self-application and potential skin sensitivity. Regarding usefulness, participants expressed that Motion Tape would enhance the efficiency and specificity of assessments and treatment. Regarding ease of use, participants stated that the app would be easy, but data management and challenges with interpretation were of concern. Physical therapists provided several recommendations for future design improvements including having a wireless system or removable wires, customizable sizes for the tape, and output including range of motion data and summary graphs and adding app features that consider patient input and context.

Conclusions: Several themes related to Motion Tape’s wearability, usefulness, and ease of use were identified. Overall, physical therapists expressed acceptance of Motion Tape’s potential for assessing and monitoring low back posture and movement, both...
within and outside clinical settings. Participants expressed that Motion Tape would be a valuable tool for the personalized treatment of LBP but highlighted several future improvements needed for Motion Tape to be used in practice.

(JMIR Hum Factors 2024;11:e55246) doi:10.2196/55246

KEYWORDS
low back pain; physical therapy; physical therapist; wearable sensor; technology acceptance model; motion tape; kinesiology tape

Introduction

Prevalence and Impact of Low Back Pain
Low back pain (LBP) is one of the world’s leading causes of disability [1-3]. In 2019, there were approximately 568.4 million prevalent cases, 223.5 million incident cases, and 63.7 million cases of years lived with disability owing to LBP reported globally [4]. LBP affects all ages and genders, but its prevalence increases with age, peaking at the age of approximately 45 to 54 years [4]. Approximately 70% to 85% of adults are expected to experience at least 1 episode of LBP in their lifetime [5]. Once predisposed to LBP, individuals are twice as likely to experience recurrent episodes of LBP [6]. Annually, LBP in the United States results in 149 million missed work days [7]. The total costs of LBP worldwide amount to approximately US $100 billion a year, with two-thirds of this amount owing to lost wages and decreased work productivity [8].

Treatment of LBP With Physical Therapy
Physical therapy (PT) is a common, effective, evidence-based treatment for LBP [9,10]. Specifically, active interventions including exercises prescribed by a physical therapist are effective for prevention and treatment of LBP [11,12]. During an initial examination, a physical therapist can identify musculoskeletal and neuromuscular impairments associated with the LBP problem by conducting assessment of the patient’s posture and movement. Then, the physical therapist and patient can work together to promote strength, stability, and mobility with in-clinic sessions and an assigned home exercise program with the goal of decreasing pain and disability [10,13]. Monitoring the patient’s posture and movement can provide a basis for determining individualized factors associated with the LBP problem, which can then be addressed through targeted interventions.

Incorporation of Technology in PT
Whether at home or at work, specific movement patterns that are performed repeatedly have been identified as a significant risk factor for the development and persistence of LBP [2,14,15]. These movement patterns of the low back region can be characterized by evaluating the angle, velocity, and acceleration [16] and can assist in LBP diagnosis, treatment, and prevention. There are several approaches to monitoring spine posture and movement. Generally, when conducting a PT examination, clinicians visually monitor posture and movement or use tools that measure the range of movement such as goniometers or inclinometers [17], but an alternative approach is to use technology to help better quantify the objective measures of spine posture and movement and offer potential benefits such as remote monitoring [16,18,19] while the patient is away from the clinic.

Technologies for Monitoring LBP
To date, existing technologies used to measure spine posture and movement in research and practice include optical motion capture, inertial measurement units (IMUs), and other wearable sensors [20-22]. Despite the variety of systems available, they generally present ≥1 limitation. Optical motion capture systems offer great precision and accuracy in monitoring human movement. However, their applications are limited owing to space needs, cost, and level of expertise needed. IMUs are portable devices that measure metrics such as acceleration and orientation [23] and include a variety of wearable sensors such as accelerometers, gyroscopes, and magnetometers, making them ideal for collecting data in a free-living environment. However, when used for monitoring human movement, IMUs have several limitations including decreased accuracy and precision for measuring slow movements [24,25], difficulty with measuring the axial plane movement accurately, inability to account for the multisegmented nature of the spine [26], and the need for multiple IMUs to triangulate posture and movement of a segment that can be cumbersome to the wearer [27].

Motion Tape
Owing to the limitations of existing sensor systems for measuring spine posture and movement, there is a need to explore new sensor innovations to address this issue. Ideally, such an approach would be wearable, unobtrusive, and usable in a clinical environment during PT sessions and in a person’s natural environment to support home-based care. Another desired requirement would be high accuracy while collecting posture and movement data for a prolonged period.

Motion Tape, developed by Loh and Lin [28], is a disposable, self-adhesive skin-strain sensor system made using graphene nanosheets coated onto commercially available kinesiology tape (also known as K-Tape) [29-33]. Motion Tape has piezoresistive properties based on the deformation of the integrated graphene nanosheets in the tape that makes it sensitive to strain [33]. In previous studies, Motion Tape has demonstrated stable performance under cyclic strains [33,34]. In addition, the Motion Tape sensor system has been tested on human participants [33,34], displaying accuracy in measuring skin strains and angles across biceps, knees, shoulders, wrists, and various other body regions when compared with IMUs and skin strains estimated using optical motion capture systems [35]. Overall, Motion Tape offers noninvasive, comfortable, and practical skin-strain measurements and can comprehensively capture complex movements and muscle engagement, especially when applied as a network of sensors [35].
Motion Tape for a Low Back Use Case

When used for a low back use case, Motion Tape provides a means to capture the lumbar spine’s multisegmental nature and multiplanar movements [36]. Motion Tape’s low-profile and stretchable nature allows it to be worn throughout the day for all human shapes and sizes, and it could be suitable for use in an individual’s natural environment with minimal interference to their daily activities. Motion Tape provides unique sensing streams that can be used in machine learning and artificial intelligence models to optimize inferences related to the management of LBP. Specifically, Motion Tape for a low back use case can address several key issues in a physical therapist’s management of LBP, including the following: expanding on the level of detail available during the clinical assessment of posture and movement, assessing spinal posture and movement in a free-living environment, use for the promotion of engagement and adherence with and precise performance of a prescribed home exercise program, and using the patient’s response to treatment to make informed decisions for future treatment or other patients [37]. Although there are several potential benefits that Motion Tape may add to personalized health care for LBP, the acceptability of Motion Tape among physical therapists has yet to be assessed.

Physical Therapists’ Acceptance of Motion Tape

The success of this device is dependent on user acceptance or one’s belief that the device will help them perform their work better (ie, perceived usefulness) and that the device’s performance benefits outweigh the effort of using the device (ie, perceived ease of use) [38]. Thus, it is vital to understand physical therapists’ perspectives about Motion Tape and their willingness to use it in their practice, to inform future developments and improvement of the technology.

Problem Statement

The primary aim of this research study was to evaluate physical therapists’ acceptance of Motion Tape for the management of LBP. The secondary purpose was to explore physical therapists’ current needs and recommendations regarding future development of Motion Tape.

Methods

Device Description and Stage of Development

In this study, licensed physical therapists evaluated a prototype of Motion Tape and examples of data streams from the app for a low back use case. The Motion Tape samples evaluated in this study included the Motion Tape sensor system with conductive wire leads connected to both sides of the sample (Figure 1).

Study Design

This exploratory, qualitative study was designed to explore physical therapists’ acceptance of Motion Tape to provide a basis for future device development (Figure 2). The study was conducted from a constructivist point of view, with the goal of gaining insightful accounts and narrations of clinicians’ lived experiences with technology and patients, rather than identifying an absolute truth [39]. We used semistructured focus groups (FGs) that incorporated human factor considerations to uncover real-world needs and obstacles and to ensure that the development of the sensor system can be informed by real-world PT clinical needs.
Theoretical Framework and Constructs

The Technology Acceptance Model (TAM) framework was used in this study to assess two determinants of user acceptance of or willingness to use a technology: (1) perceived usefulness and (2) perceived ease of use [38,40]. An additional factor of wearability was also assessed to examine physical therapists’ perceptions about patient-centered issues that would affect whether the device would be worn [41]. Recommendations for future improvements were also investigated to collect insight into data, device, and app developments that clinicians would like to see for Motion Tape.

**Perceived usefulness** was defined as the degree to which the use of Motion Tape would enhance the physical therapists’ management of LBP [39-42], and this was assessed using the following constructs: (1) productivity, (2) effectiveness, (3) ability to make their job easy, and (4) benefits to PT treatment and recovery. Perceived ease of use was defined as the degree to which the use of Motion Tape would be effortless when used for managing LBP [39-42], and this was assessed using the following constructs: (1) how easy it would be for physical therapists to learn how to use it, (2) what level of instruction would physical therapists need to use it, and (3) how clear and understandable Motion Tape was in its current state. Wearability was defined as the degree to which Motion Tape would fit well and be comfortable for patients to wear on their back [42], and this was assessed based on (1) adhesion, (2) fit, (3) feel, and (4) how comfortable physical therapists would feel about applying and prescribing Motion Tape.

Participants and Setting

This study was conducted at the American Physical Therapy Association’s (APTA’s) Combined Sections Meeting (San Diego, California) on February 24, 2023. Participants were recruited by sending study information via email to physical therapists who were members of the APTA Academy of Leadership Technology Special Interest Group. Members were also offered an opportunity to participate when they attended the Technology Special Interest Group in-person meeting at the APTA Combined Sections Meeting. Individuals were included in this study if they were a licensed physical therapist and were excluded from participating if they were unable to respond to questions in English. In total, 8 physical therapists were eligible and agreed to participate in 2 FGs of 4 clinicians each. A sample size of 8 people, in 2 FGs, was considered sufficient for this qualitative study to provide adequate variability and data saturation [43] and to provide a basis for device improvement. In addition, after data from the 2 FGs were collected and analyzed, the data were deemed saturated (ie, no new themes or codes were generated) and no further FGs were needed.

Ethical Considerations

The study protocol was considered to be exempt from ethics approval by the San Diego State University institutional review board. Each participant provided written consent before participating.

FG Methods

An FG guide (Multimedia Appendix 1) was used to lead the group’s discussion. The FG guide was developed by investigators (AL, PD, and SG) to be semistructured with open-ended questions to explore the participants’ perspectives about the usefulness, usability, and wearability of Motion Tape and to collect insight into future improvements for the sensors and data visualization (Textbox 1). A template of the FG guide was piloted with a Doctor of Physical Therapy student and a physical therapist at San Diego State University to ensure credibility [44]. General domains for each construct were prespecified to correspond with each interview question. Domains were defined based on the TAM framework and included perceived usefulness and perceived ease of use. An additional domain of wearability also was assessed.
Textbox 1. Guiding questions from the focus group guide.

**Perceived wearability (W)**
- How secure do you think the Motion Tape adhesive will be? (W-adhesion)
- To what degree do you think these sensors would fit your patients’ anatomy (ie, their low back)? (W-fit)
- To what degree do you think your patients would feel the sensors on their back? (W-feel)
- How do you predict the Motion Tape Sensors would feel when being removed? (W-feel)

**Perceived usefulness (U)**
- To what degree would the usage of Motion Tape sensors affect how quickly you can assess your patient’s posture, movement, or exercise performance? (U-efficiency)
- How effective do you think the Motion Tape sensors will be to capture valid data on your patients in the clinic? (U-effectiveness)
- How effective do you think the Motion Tape will be to capture valid data on your patients in their daily routine and normal environment? (U-effectiveness)
- To what degree would the usage of Motion Tape sensors affect the level of difficulty of your job as a clinician/physical therapist? (U-make job easier)
- What features, if any, would make the Motion Tape more useful to you? (U-useful)

**Perceived ease of use (EU)**
- How easy do you think it would be to learn how to use Motion Tape? (EU-easy to learn)
- How comfortable would you feel prescribing Motion Tape to a patient to monitor their movements at home? (EU-comfort in usage)
- What level of knowledge do you think a clinician/PT would need to use the Motion Tape? (EU-clear and understandable)
- How easy/difficult do you think it would be for a clinician/PT to apply the Motion Tape to the patient’s back? (EU-easy to use)
- What features, if any, would make the Motion Tape easier for you to use? (EU-easy to use)

FGs were conducted by AL (a female Master of Science student investigator) and PD (a female PhD student investigator). Reflexivity was maintained by the research team by discussing assumptions and biases that may influence how the clinicians responded to the FG moderators, who were not licensed physical therapists. As SG is a licensed physical therapist and member of APTA, she was able to provide valuable insight during the development of the interview guide, analysis, and interpretation to ensure credibility of the findings [44].
**Data Processing and Analysis**

All FG audio data were downloaded to a HIPAA (Health Insurance Portability and Accountability Act)-compliant laboratory server, accessible only to the research staff, and removed from the digital voice recorder. The recordings were then transcribed, first using computer-based transcription (Word; Microsoft Corp). An investigator then checked and verified each transcription by listening to the original audio and reviewing and correcting the computer-based transcription.

Considering the need for timely feedback in the sensor development process, we adopted a rapid qualitative analysis (RQA) approach to explore themes regarding the acceptability and wearability of Motion Tape [45]. RQA was conducted by 3 investigators to assess the FG responses effectively and efficiently and to identify major themes. Codes and themes for RQA were deductively developed based on the TAM framework and the study objective [41]. We then used an inductive approach to generate RQA codes and themes, allowing for quick sorting of FG dialogue.

To ensure rigor and consistency of the method, a constant comparative approach with investigator triangulation was used at each stage [46]. First, the 3 investigators independently completed a summary report for each FG, with quotes and relevant topics under the respective themes and codes. Once the individual coding and summary reports for both FGs were completed, the investigators consolidated them into a combined rapid analysis summary report for each FG, unifying themes and reconciling discrepancies by consensus through discussion.

The summary reports for each FG were then transferred into a matrix in which each row was a participant quote and each column was a domain. From this matrix, investigators identified the underlying themes and subthemes between the 2 FGs.

**Results**

**Overview**

In total, 8 physical therapists (n=5, 63% men and n=3, 38% women), with a mean age of 47.5 (SD 5.6) years participated in this study. Participants reported obtaining PT degrees ranging from a bachelor’s degree to a Doctorate in Physical Therapy and had, on average, 20 (SD 8.5) years of clinical practice experience, and most reported practicing in an outpatient orthopedic setting. Of the 8 participants, 5 (63%) reported having advanced doctoral degrees (3/5, 60% PhD; 2/5, 40% EdD).

The qualitative results from the FGs were organized using the TAM for the acceptance of Motion Tape [38,40-42]. Data were organized based on the 3 main domains relevant to user acceptability and wearability of Motion Tape [48].

Subthemes were further designated using positive, negative, and neutral valences. Positive valence indicates that the FG participants perceived the Motion Tape attribute as positive. Negative valence indicates that the FG participants perceived the attribute as negative. Neutral valence indicates that the FG participants perceived the attribute as neither positive nor negative.

---

**Figure 3.** Poster of Motion Tape placement and app data output for a low back use case. (A) The laboratory setup with 6 pieces of Motion Tape and several optical motion capture markers on anatomical landmarks of the lumbar spine. (B) The graphs display the following: (1) blue—the normalized strain data for extension, captured by the 6 Motion Tapes, and (2) purple—the kinematics for extension in degrees, captured by the optical motion capture system (reference standard). (C) The normalized strain data for right and left lateral bending obtained from the 6 Motion Tapes. (D) The normalized strain data for right and left rotation obtained from the 6 Motion Tapes.
Textbox 2. Themes (n=3), subthemes (n=21), and valences of user acceptance of Motion Tape.

<table>
<thead>
<tr>
<th>Theme 1: perceived wearability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>• Motion Tape has a small, universal fit.</td>
</tr>
<tr>
<td>• The feeling of Motion Tape on the skin would decrease over time.</td>
</tr>
<tr>
<td>Negative</td>
</tr>
<tr>
<td>• Patients may feel Motion Tape’s wires snagging or sensors rubbing on clothes.</td>
</tr>
<tr>
<td>• Motion Tape does not consider people with skin sensitivities.</td>
</tr>
<tr>
<td>Neutral</td>
</tr>
<tr>
<td>• Motion Tape adheres for 3-4 days but may adhere less owing to external factors.</td>
</tr>
<tr>
<td>• The feeling of Motion Tape being removed depends on the physical therapist.</td>
</tr>
<tr>
<td>Theme 2: perceived usefulness</td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>• Motion Tape could increase specificity of physical therapy management of low back pain (LBP).</td>
</tr>
<tr>
<td>• Motion Tape could be effective for the diagnosis, management, and monitoring of low back pain (LBP).</td>
</tr>
<tr>
<td>• The feeling of Motion Tape and the awareness of Motion Tape monitoring would increase adherence to a home exercise program.</td>
</tr>
<tr>
<td>• Motion Tape would be beneficial in telerehabilitation and hybrid sessions.</td>
</tr>
<tr>
<td>• Motion Tape could increase the physical therapist’s awareness of the pain source.</td>
</tr>
<tr>
<td>Negative</td>
</tr>
<tr>
<td>• Motion Tape brings legal concerns with data responsibility.</td>
</tr>
<tr>
<td>• Motion Tape’s reliability could be affected by external factors.</td>
</tr>
<tr>
<td>Neutral</td>
</tr>
<tr>
<td>• Motion Tape could increase the efficiency of assessments, but set up could take more time.</td>
</tr>
<tr>
<td>Theme 3: perceived ease of use</td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>• Motion Tape would be easy for a physical therapist to apply.</td>
</tr>
<tr>
<td>Negative</td>
</tr>
<tr>
<td>• Motion Tape has a lot of data to sift through.</td>
</tr>
<tr>
<td>• Motion Tape data are hard to interpret in their current state.</td>
</tr>
<tr>
<td>• The self-application of Motion Tape would be difficult.</td>
</tr>
<tr>
<td>• Motion Tape is designed for single use.</td>
</tr>
<tr>
<td>Neutral</td>
</tr>
<tr>
<td>• The prescription of Motion Tape is subjective to many factors.</td>
</tr>
<tr>
<td>• The user interface would dictate how much knowledge would be needed to use Motion Tape.</td>
</tr>
</tbody>
</table>

**Domain 1: Perceived Wearability**

Regarding perceived *wearability*, all physical therapists were familiar with commercially available kinesiology tape. Thus, their thoughts about perceived wearability reflected their experience with kinesiology tape. For example, the physical therapists expected Motion Tape to last about 3 to 4 days. A physical therapist mentioned the following:

> Oh, I’ve used the K-Tape for four days before it started peeling off. Sometimes it lasts more than five days actually. Three to four days I think is average. [FG1]
However, some physical therapists clarified that the longevity of Motion Tape’s adhesion depends on several factors. For example, 2 of the physical therapists expressed the following:

> How secure it is depends on a lot of factors, like moisture on the skin. It depends on not just moisture, but how clean your skin is and how much hair is on the skin. [FG2]

Some of them, specifically on the low back, tend to have more oily skin, and that depreciates the life of the tape. [FG1]

Regarding the fit aspect of wearability, physical therapists also believed that Motion Tape’s size was sufficiently small to be universal to the wearer and the placement location. They expressed the following:

> In my experience with tapes like this, it fits most of the clientele that I’ve worked with, both inpatient and long-term post-acute. [FG1]

If it was that little strip, I think it would be great to use anywhere. [FG2]

Regarding the feel aspect of wearability, generally, physical therapists felt that patients would feel Motion Tape at first when applied but would become less aware over time until the tape starts to peel off:

> They’d know that they’re there, and they’d probably become less aware of over time. [FG1]

However, physical therapists generally felt that with Motion Tape’s current design, patients would feel the wires snagging or the sensors rubbing on clothing. A physical therapist explained it as follows:

> So contraptions with wires will always have that uncomfortable feeling. Always. But if you go the wireless route, then probably after two days, the patient will be more comfortable until the tape starts peeling off. However, what I’m wearing right now, something that goes above my PSIS, if I go to the bathroom or do something, I’m going to, it’s gonna move around, it might get pulled on it by my clothes. [FG1]

When removing Motion Tape, physical therapists said that patient feelings about the removal process would be quite variable. Some physical therapists felt that it was subjective to how the therapist removed Motion Tape and how much hair or oil the individual has on their skin. A physical therapist explained it as follows:

> I’m just thinking of whoever is taking it off. You know, like, it depends on you, like, some people just rip. And some people are just gentle. So subjective. So it depends on the training of the therapist and concern if they’re empathetic to our patients. [FG1]

The physical therapists mentioned some wearability concerns during the FGs. A concern was about how patients with skin sensitivities would be able to use Motion Tape. A physical therapist asked the following:

> For those with skin allergy. Can you put an under wrap under this? [FG1]

### Domain 2: Perceived Usefulness

Physical therapists expressed mixed feelings about whether Motion Tape would increase their efficiency with assessments of lumbar spine posture and movement. Some expressed that if all they had to do was apply the tape, then there would be increased efficiency:

> If it’s easy to objectively document, by understanding the graph, I think it’s a night and day difference versus getting into the goniometers and doing manual assessment. Instead, you put on the tape, ask the patient to rotate their trunk, lean forward, reach forward, extend their back. And then if I have it digitally by email or direct messaging, it would save a bunch of time. [FG1]

However, others felt that it would reduce efficiency. A physical therapist explained the following:

> Regarding the speed of assessment, I would be a little doubtful. I think by the time that you took this and you put it on the patient, you hooked up all the wires to it, you did the calibration, if you need to do a calibration, it might take just as long as doing an assessment. I would have concerns around the accuracy of this, to give you a number, an accurate range of motion, particularly for things like rotation. But if the data was convincing that everyone, if it was validated for everyone that gave you an accurate number, I think it could improve the quality of assessment. [FG1]

A physical therapist felt that for the in-clinic assessments, Motion Tape would improve specificity:

> I don’t feel like it [Motion Tape] would improve speed, it would improve specificity. [FG2]

Physical therapists also mentioned that they could envision Motion Tape as a useful tool for self-management and remote monitoring when used in combination with in-clinic PT. A physical therapist mentioned that the ability to monitor patients outside the clinic would be very meaningful:

> That’s the best place to actually observe them, their normal environment. If they’re in therapy, they’re being observed, coached, cued by a skilled clinician. Their performance is definitely going to be different. So if they’re at home, and we’re able to monitor them at home, I think the treatment will be more, and your adjustment and progression will be more meaningful. [FG1]

Some physical therapists suggested that having patients wear these sensors would increase their awareness of being monitored and thus increase engagement with and adherence to the home exercise program:

> I think that what it has to offer is improving...adherence with our programs. I think that’s your potential. [FG1]
When you tell someone, I’m looking at your posture right now, you change [gesturing to posture]. If they think you are watching, they’ll do better. [FG1]

Physical therapists expected patients to have a phenomenal experience with Motion Tape when used in a hybrid setting:

I think to his point that if it’s applied properly in the clinics, it’s hybridized, and you can take a call, and there’s no technical involvement on the patient side, and all they do is open up the app, they’d have this really phenomenal experience. [FG2]

Specifically, several physical therapists expressed that Motion Tape would help with the identification of postures and movements in free-living environments that provoke pain, allowing for more meaningful interventions:

I think for it to be very useful. It would have to compare with the app where you’ve got user input as to what’s going on...where he’s got these flags and the data that was pain here, pain here, pain here, and you can look, you know, to the periods of time before that. [FG2]

Some physical therapists did have some concerns about the usefulness of Motion Tape. A physical therapist expressed legal concerns regarding data responsibility:

As long as you collect data, someone’s then responsible for it. So who’s going to look at it? What’s the liability then that person takes on by having that information?...if something goes wrong, and the therapist hasn’t looked at the data, I’d like to know, are they liable? [FG1]

Another concern was knowing what external factors affect Motion Tape’s signal and data reliability, mentioning that the use of Motion Tape in practice was “gonna depend on the reliability of the data” (FG2).

Several physical therapists felt that there were a variety of variables that might affect the reliability of the signal or data. They expressed the following:

And what other factors affect them, the sensors, as far as humidity, water, other environmental factors that might affect it? You know, what if they have a compression garment around the trunk, for example, does that affect the sensors? [FG2]

Whether, getting it wet and getting so some things on it changes the conductivity, and therefore the calibration over time. [FG2]

You get variability in the readings based on amount of tension that people put on it when they applied it. [FG2]

Whether that’s different from person to person because of different makeup morphologies. [FG1]

Domain 3: Perceived Ease of Use

Physical therapists felt that it would be easy to apply Motion Tape, given their background knowledge in human anatomy. A physical therapist stated the following:

You would need to know basic clinical knowledge of the application for where to look for the muscles, you know, right. So, they need to be clinician to have knowledge of the body. [FG2]

When asked whether they would feel comfortable using Motion Tape with their patients, there were mixed responses among physical therapists. Some mentioned that it would depend on “cost and buy-in” (FG2) or how it was going to be “incorporated into the plan of care” (FG1). A physical therapist even explained the variability as follows:

Depends on the situation, honestly. I mean, I have some families that I’ll show them how to do the application. And I’ll see them three weeks later, and they’ve reapplied four times and done it great. And then I’ve seen others that I’m like, “Oh, no! This is nope.” [FG2]

There were also several concerns about the ease of use. Some physical therapists felt that they would have challenges with ease of use, specifically regarding interpretation of the data:

I think in its current form, easy to apply. Hard to interpret. [FG1]

It depends on the interface and how much it interprets the points. The tape will be easy, but it’s all the other pieces. [FG2]

Additional concerns about the ease of use included that the amount of data presented was excessive and the type of data displayed was difficult to interpret. The physical therapists expressed a desire to see the range of motion displayed in degrees rather than resistance in ohms:

I think I’m probably realistically just correlated with what they report has been painful. Because I don’t know that I’ve ever been so interested in all of that. Like, it might be too much data. For a patient, like I don’t necessarily need to know their range of motion during every single activity, I need to know when it is relevant to them. And when it is impacting whatever condition they’re here for. [FG1]

And again, I think for a clinician, it’s going to have to be meaningful data. It’s gonna have be Range of motion data not ohms. [FG1]

So then, conceivably, would it be helpful instead of giving you normalized strain,...if they could interpret it, would convert this over to degrees of rotation and flexibility? [FG2]

If you could get range of motion kind of information, I think that would be great. [FG2]

Another concern was about how challenging the self-application would be for patients:

How are people actually going to apply this on their own, someone that doesn’t know how? [FG2]

Finally, another concern was that Motion Tape is a single-use product. A physical therapist explained the challenge of a single-use product as follows:
Okay, now how about waste? So it’s like a single use thing? Now I’m gonna throw in a whole planet into this is single usage. Or can you reapply? [FG2]

Future Recommendations

Future recommendations from the physical therapists were organized into 3 categories (Textbox 3): data, physical features, and app features.

Textbox 3. Themes (n=3) and subthemes of future recommendations for Motion Tape.

<table>
<thead>
<tr>
<th>Theme 1: Data recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Motion Tape data should be easy to read at a glance.</td>
</tr>
<tr>
<td>• Motion tape data should account for differing patient morphology.</td>
</tr>
<tr>
<td>• Physical therapists should be aware of factors that affect Motion Tape data.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Theme 2: Physical feature recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Motion Tape should be made wireless or with removable wires.</td>
</tr>
<tr>
<td>• Motion Tape should be reusable.</td>
</tr>
<tr>
<td>• Motion Tape should be customizable in length.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Theme 3: App feature recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Motion Tape app should include BMI input.</td>
</tr>
<tr>
<td>• Motion Tape app should include input for a patient’s change in activity.</td>
</tr>
<tr>
<td>• Motion Tape app should allow flagging events.</td>
</tr>
<tr>
<td>• Motion Tape app should include comparative data.</td>
</tr>
</tbody>
</table>

Regarding the data recommendations, physical therapists expressed that data should be summarized in the form of an at-a-glance graph with 1 overall meaningful number, reflecting the range of motion. They would also like to know how the data change from person to person owing to morphology and how external factors (water, application stretch variability, and skin movement) affect the data. Additional data that would be useful for their job included comparative data, graphs with a color scale, and information about muscle activation. Participants in an FG expressed the following:

Take a baseline and have them rotate from that position and determined by the volume strain, whether they are tension either degrees, or even if it’s yellow, green, yellow, red, like if they’re moving within if they can’t pinpoint it specifically, but you know, within a range, would that be helpful? [FG2]

I think even just having comparative data would be helpful, right? Because, you know, I keep telling my students, “Don’t tell me, ‘I want to increase range 10 degrees.’” Because that doesn’t tell me, “Can they walk?” Right? But, “Are they doing better now than they were doing when we started?” That’s useful. So even if we get baseline data that could be translated into amount of motion and then follow up data that says, “hey, it’s more, it’s more fluid, it’s better, it’s whatever.” I think that can be really useful. Now I know the payers are gonna want, how much rotation did you get? How much lateral flexion did you get? [FG2]

And I think beyond the range of motion, I work in neuro. I think just like muscle activation would be interesting, you know, like, how much activation did you get today, for example, versus six weeks ago, post stroke or, you know, spinal cord or something? I think that would be really interesting just to see the firing muscle activation. And on the flip side, and I don’t know if that’s possible, but looking at specificity. Could that be something to monitor changes in specimen specificity? Post- X Y & Z intervention, right? That could also be interesting. So it’s not really about range of motion, we’re also activity known as firing or not? [FG2]

Regarding physical feature recommendations, physical therapists wanted a way to mitigate the wires, either by moving to a wireless system or making them removable. Physical therapists were also concerned about the limited stretchability of the short pieces of tape, as it would not be long enough in length for typical kinesiology tape use, and recommended making the length customizable to the physical therapists’ needs. Physical therapists were also concerned about Motion Tape’s single-use design and were curious about whether it could be reusable to reduce waste:

Again, I’m thinking like, in the future, no wires, you’ve got a strip of graphene that you could customize length to, with those couple millimeters around the edge. And if we wanted a whole length, we cut whole lengths. And if we want segments, we can cut segments. And it feeds the data to the app somehow tailor it to someone’s body. [FG2]

So you can imagine that maybe something like this could be a roll of tape. Yeah, the width of duct tape. And there’s actually two pieces on this roll. There’s one section, that’s the conductive piece, that you can cut it to length, and then next to it there are maybe
there’s a wire section, that’s conductive tape that you can pull off and put on the ends of whatever you choose. So you get one roll of tape. And then one of them is the is the graphene is the other piece that you tear off to the appropriate length is the conductive tape that connects it to the box. And then it’s a solution, you can customize length and you have your conductive piece and then your graphene. [FG2]

Regarding future app feature recommendations, physical therapists expressed a need for the capability to input factors such as BMI, activity changes, “flags” for events, and changes in pain to help label, compare, or contextualize the data.

**Discussion**

**Overview**

There is a gap in the research between rehabilitation device development and evaluation of clinicians’ acceptance of such devices. Most existing studies have considered patient or user satisfaction [47,48], whereas others that consider the clinician’s perspective have not specifically evaluated sensors for measuring spine posture and movement [49,50]. In this study, several themes relating to physical therapists’ perspectives about Motion Tape’s wearability, usefulness, and ease of use for a low back use case were identified.

**Domain 1: Perceived Wearability**

One of the most common challenges for wearable sensors is ensuring that they are unobtrusive to the wearer’s natural movement and environment [39]. The small form and fit of Motion Tape was considered by physical therapists to be ideal for a wearable sensor. However, similar to previous studies, the wires in the current design were considered to be not ideal [37]. Studies have shown that wireless technologies tend to be more widely used in many fields, especially in the field of wearable devices for health care [51]. Thus, a future iteration of Motion Tape without wires would be considered optimal. On the basis of feedback obtained from physical therapists, wearability for people with skin sensitivities also should be considered. Previous studies have shown that skin irritation is the most common concern when using kinesiology tape for extended periods of time [52,53]. Thus, future studies should explore whether a medium or substrate can be used under Motion Tape to mitigate skin irritation, possibly as an extension of recent research that integrated Motion Tape with elastic fabric for respiration monitoring [54].

**Domain 2: Perceived Usefulness**

There were mixed feelings among physical therapist participants about how efficient Motion Tape would be in the clinic. Overall, most physical therapists felt that Motion Tape would increase the specificity of their assessments, a characteristic that has been shown to be beneficial for LBP diagnosis and treatment [55]. Furthermore, Motion Tape’s ability to monitor the patient’s movements remotely was considered beneficial, as this feature may increase adherence to home exercise programs, which is an important component of effective treatment for LBP [56,57].

**Domain 3: Perceived Ease of Use**

On the basis of physical therapists’ perspectives, Motion Tape would be easy to apply, but data would be difficult to interpret. Creating a device that is easy to use and understand is crucial because it predicts consumer use behavior [38,41]. Recommendations included presenting the data in units that physical therapists are more familiar with (ie, degrees of range of motion) and creating an app that requires minimal time for the physical therapists to use. These changes may promote increased device use and acceptance in PT.

**Future Recommendations**

On the basis of clinician feedback, Motion Tape appears to be a promising new technology that could be used for monitoring lumbar spine posture and movement in the management of patients with LBP. Future device development will be needed to address clinician recommendations obtained from this study in the domains of wearability and ease of use. In addition, future studies will be needed to validate Motion Tape in laboratory, clinical, and free-living environments and to investigate patient acceptance of Motion Tape.

**Limitations**

A limitation of this study is that participants were physical therapists who were part of a Technology Special Interest Group and are likely to be more receptive to using technology in practice. Thus, this study’s results regarding Motion Tape’s acceptability may be biased in favor of Motion Tape’s ease of use, usefulness, and wearability. Future studies should also assess the acceptability of Motion Tape for clinicians who do not regularly use technology in their practice. Another limitation is that the physical therapists were not presented with active samples of Motion Tape with live data streams in the app. Instead, participants were given inactive samples of Motion Tape and presented with a poster with examples of app data streams. Future studies should provide an opportunity for physical therapists to apply Motion Tape to a person and use it with the app interface. Finally, there was a potential for investigator bias in the interpretation of the results, as several investigators of this study are actively working on the development of this device. However, 2 of the 3 investigators who conducted data analysis were outside the primary research team.

**Conclusions**

Physical therapists expressed overall acceptance of Motion Tape for its potential to monitor and assess low back posture and movement, both within and outside clinical settings. Physical therapist participants expressed that Motion Tape would be a valuable tool for personalized treatment of LBP but highlighted several future improvements needed for Motion Tape to be used in practice.
Acknowledgments

The authors would like to thank Nicolette Jaghab for her assistance with qualitative data analysis. This study is part of the Multi-Sensor Adaptive Data Analytics for Physical Therapy project and was supported by US National Science Foundation (award IIS-2205093).

Conflicts of Interest

KJL is a cofounder of JAK Labs Inc, a company that may potentially benefit from the study results. JAK Labs intends to commercialize Motion Tape for the physical therapy and rehabilitation market, among other markets. The terms of this arrangement have been reviewed and approved by the University of California San Diego in accordance with its conflicts of interest policies.

Multimedia Appendix 1

Focus group guide.

References


Abbreviations
- APTA: American Physical Therapy Association
- FG: focus group
- HIPAA: Health Insurance Portability and Accountability Act
- IMU: inertial measurement unit
- LBP: low back pain
- PT: physical therapy
- RQA: rapid qualitative analysis
- TAM: Technology Acceptance Model
Usability of an App for Medical History Taking in General Practice From the Patients’ Perspective: Cross-Sectional Study

Klara Albrink¹; Dominik Schröder¹, MSc; Carla Joos¹; Frank Müller¹, MSc, Dr med; Eva Maria Noack¹, PhD
Department of General Practice, University Medical Center Göttingen, Göttingen, Germany

Corresponding Author:
Eva Maria Noack, PhD
Department of General Practice
University Medical Center Göttingen
Humboldtallee 38
Göttingen, 37073
Germany
Phone: 49 551 39 65745
Email: evamaria.noack@med.uni-goettingen.de

Abstract

Background: A future shortage of physicians, especially in general practice, will result in an increasing workload for health care providers as a whole. Therefore, it is important to optimize patient-encounter processes to increase time efficiency related to visits. Utilizing digital tools to record patients’ medical histories prior to a consultation offers great potential to achieve this goal. The collected information can be stored into the practice’s electronic medical record, allowing for the general practitioner to review structured information of the patients’ complaints and related medical history beforehand, thereby saving time during the encounter. However, the low usability of new digital developments in this setting often hinders implementation.

Objective: The aim of this study was to evaluate the usability of an app designed for medical history taking in general practice to capture the patients’ perspective.

Methods: Between November 2021 and January 2022, we recruited 406 patients with acute complaints in one out-of-hour urgent care and seven general practice clinics. These study participants used the app during their waiting time and subsequently assessed its usability by completing the System Usability Scale (SUS), a robust and well-established 10-question survey measuring the perceived usability of products and technologies. Additionally, we collected general participant information, including age, sex, media usage, health literacy, and native language. Descriptive and inferential statistics were applied to identify patient characteristics associated with low or high SUS scores.

Results: We analyzed data from 397 patients (56.7% female, 43.3% male). The mean total SUS score was 77.8 points; 54.4% (216/397) of participants had SUS scores of 80 points or higher, indicating high usability of the app. In a multiple linear regression predicting SUS score, male sex and higher age (65 years or older) were significantly negatively associated with the SUS score. Conversely, a higher health literacy score and German as the native language were significantly positively associated with the SUS score.

Conclusions: Usability testing based on the SUS anticipates successful implementation of the app. However, not all patients will easily adapt to utilizing the app, as exemplified by the participants of older age in this study who reported lower perceived usability. Further research should examine these groups of people, identify the exact problems in operating such an app, and provide targeted solutions.

Trial Registration: German Clinical Trials Register World Health Organization Trial Registration Data Set DRKS00026659; https://trialsearch.who.int/Trial2.aspx?TrialID=DRKS00026659

(JMIR Hum Factors 2024;11:e47755) doi:10.2196/47755

KEYWORDS
digitization; application software; usability; mHealth; history of present illness; medical history taking
Introduction

As in many countries, demographic change is becoming evident in the German health care system, resulting in more complex, multimorbid patients [1] and a shortage of physicians [2]. Moreover, the proportion of older people in the population is rising steadily [3] and people tend to use medical services at a higher rate as they increase in age [4]. In Germany, one group that is particularly affected by this development are general practitioners (GPs) who are the first point of contact for patients requiring medical care and serve as the “gatekeepers” in the German health care system [5]. Approximately 80% of the German population aged 18 years and older are treated by a GP at least once a year [6]. A considerable number of GPs will retire in the upcoming years, resulting in 11,000 GP vacancies expected by 2035. These vacancies will disproportionately impact structurally weak and rural areas [7,8]. Without a sufficient workforce to replace the retired GPs and meet the greater demand for physicians, remaining GP workloads are expected to increase significantly within the next decade [9]. These developments challenge the German health care system at various levels and require attention to address the following key issues: future financing, improving allocation of resources, ensuring access to care, increasing efficiency and effectiveness of health care provision, and strengthened collaboration between providers [10].

To streamline patient care in the upcoming years, it is of importance to optimize patient-encounter processes to increase time efficiency related to visits. In this respect, digital tools offer great potential to support GPs in patient management, documentation workload, and the collection of medical history before consultation.

Digital tools designed to collect patients’ medical history can ensure that information is always collected and documented thoroughly in a structured manner and with consistent quality. As many conditions can be diagnosed via a thorough medical history [11,12], these tools can be helpful in maintaining quality of care when time constraints may lead to an otherwise superficial medical history.

As part of our project titled “Digitally assisted information acquisition before medical consultation” (DASI), we developed an app for medical history taking in general practice settings. The app is used by the patient prior to the medical consultation, which could be either in the waiting area or at home. The collected information can be stored in the practice’s electronic medical record and eventually be transferred to the individual electronic patient file, which statutory health insurers in Germany have been entitled to use since 2021 [13]. In the electronic patient file, patient data such as medical reports, X-rays, immunization records, and other medical data can be stored and shared among medical providers involved in the care of a particular patient by using the telematics infrastructure [14].

One advantage of the app is that the GP can review structured information of the patients’ complaints and related medical history before the encounter. This is particularly helpful for patients that are unknown to the provider, those with many complaints, or those who have a comprehensive medical history.

These situations are especially prevalent in out-of-hour urgent care practices. Furthermore, the tool might help patients to reflect on their conditions and enable them to better address their needs when seeing the provider. In this way, we expect that the limited consultation time can be used more efficiently.

Despite Germany’s progress in digitalization within the health sector, concerns remain about the limited usability of new digital tools, hindering their full implementation. More than half of German practices see low usability as a strong obstacle to digitalization [15]. The evaluation of a digital tool’s perceived usability is of special interest as it is a key determinant of performance for end users. Therefore, the aim of this study was to assess the usability of the app from the patients’ perspective and to identify features in need of improvement. The broader aim is to ensure that the app is suitable for implementation in everyday practice, considering that GPs treat a broad range of patients of all ages and various educational and cultural backgrounds [16].

Methods

Study Design and Recruitment

This was a cross-sectional study conducted in Germany in one out-of-hour urgent care practice and seven GP practices to assess the usability of an app designed for medical history taking in general practice settings.

Software and Hardware

The app was developed to take a medical history based on general medical complaints directly from the patients. While there are no international standards for the composition of a standardized patient history, this app was developed based on guidelines and health literature by medical experts from aidminutes GmbH (Hamburg/Buchholz in der Nordheide, Germany). For this study, the content and query structure were further refined for primary care (general practice and out-of-hour practices) by aidminutes GmbH in collaboration with experienced researchers from the Department of General Practice at the University Medical Center Göttingen, Germany. The app was designed to be used by patients in the waiting room before they see the doctor. Patients select one or several complaints and are then guided through a symptom-related questionnaire. In the sense of a branching logic, the app is adaptive to patient responses, which trigger further specific questions about the selected key complaints (eg, how and when a symptom started). Patients are also asked about preexisting conditions, previous treatments and surgeries, current medication, living habits, and chronic conditions in the family history. Information such as biological sex, height, weight, age, as well as the subjectively perceived severity of the complaints are inquired from all patients. More details can be found in the published study protocol [17].

The app was designed to be intuitive for the user such that no prior knowledge or any kind of instruction for its use is necessary. The user interface was designed to be simple to follow and only one question is asked per screen. As the app is operated in the waiting area, sound and video output of an earlier version [18] was omitted due to data protection. The questions
are phrased in plain language; medical terminology is avoided or otherwise explained. The questions are substantially comprised by single-choice or multiple-choice questions that can be answered by tapping but also include several data fields (for age, height, and weight) and slider-type questions (Figure 1). The color scheme was designed to ensure reading accessibility for patients who may be color blind. A zoom function can be used for users who may experience visual impairment.

Figure 1. Screenshots of the app for medical history taking in general practice showing different types of questions: (A) single-choice question; (B) multiple-choice question; (C) hybrid question (i.e., patients can either select several options or negate all of them); (D) slider for questions including a ranking between items (depicted here as “How sick do you feel?”); (E, F) data entry field (here: “Please enter your age”); and (G, H) selection of a body region on a figure (depicted in figure: “Please mark on the figure where you are suffering from the problems”).

As this is a web-based app, it relies on a permanent internet connection. For this study, the app ran on an iPad Mini 4 (Apple Inc, Cupertino, CA, USA) held in an upright position. Tablets were equipped with haptically and visually inconspicuous cases (dark grey polyurethane leather outside and microfiber inside).

Setting
In Germany, GP practices aim at providing preventive, acute, and rehabilitative health care with long-lasting patient-doctor relationships. Out-of-hours urgent care practices provide urgent medical care for acute but not life-threatening cases when other practices are closed. Urgent care practices are often staffed with doctors of various specialties and an established relationship of care between the patients and doctors is not common. These aspects can lead to challenges in efficiently obtaining an accurate medical history and identifying serious health problems. Although the app was designed for general practice, it is also suitable to be used in out-of-hour urgent care practices.

Data Collection
The recruitment of patients was carried out by three study nurses and took place from November 22, 2021, to January 12, 2022. Patients were approached by study nurses in the waiting room of the respective practices before seeing their GP.

Patients meeting the following criteria were eligible to participate in the study: (1) seeking care in a participating practice because of acute somatic and/or psychological complaints, (2) at least 18 years old, and (3) consenting to participate in the study. Patients meeting the following exclusion criteria could not participate in the study: (1) younger than 18 years old (legal minor), (2) patients in an apparent emergency, (3) patients who required immediate medical treatment, and (4) patients who were unable to provide consent.

After the study nurses obtained written informed consent, a tablet on which the app was run on was handed over to the study participants. Participants used the app to report their medical history without an introduction on how to navigate the app. Once finished, they were asked to answer questions on personally perceived usability, media usage, and further
sociodemographic data, which were digitally attached to the medical history–taking document. The study nurse in charge was present to observe any problems study participants may have had with using the app and was available to answer questions about the app’s content and usability if specifically requested. Data were collected in an anonymized format without any personal information (eg, name or address) linking the results to each study participant. More detailed information on the data collection can be found in the study protocol [17].

**Ethical Considerations**

The Medical Ethics Committee of the University Medical Center Göttingen approved the study (approval number 26/3/21). A written informed consent form was collected from all patients before their inclusion in the study. Participating in the study was voluntary for patients. Patients could withdraw from participation without giving a reason at any time before they had completed the survey. Subsequently, their data could no longer be deleted because it could not be traced back to the individual.

**Measures**

The main outcome “usability” was measured using the System Usability Scale (SUS) [19], a commonly used instrument for this purpose [20]. The SUS was developed based on Standard ISO 9241-11 [21], in which usability is measured by the three main attributes of “effectiveness,” “efficiency,” and “satisfaction” [22,23]. Compared to other instruments, the SUS offers several advantages: (1) it can be analyzed quickly, (2) it is relatively easy to understand by academics from other disciplines [24], (3) it contains only 10 statements for easy completion, and (4) it can be used to evaluate almost any type of user interface [25]. We used the translated and validated German version of the SUS [26] and modified the statements to suit our purpose (see Multimedia Appendix 1).

The SUS consists of 10 statements (Table 1), where statements 1, 3, 5, 7, and 9 are positively connoted and statements 2, 4, 6, 8, and 10 are negatively connoted [19]. The scores for these statements are therefore inverted when calculating the sum. The raters decide on the extent to which they agree or disagree to these statements on a 5-point Likert scale ranging from 0 (strongly disagree) to 4 (strongly agree). The final sum score is multiplied by 2.5, resulting in a score range of 0-100 with higher scores indicating better usability [19].

Lewis and Sauro [27] developed a curved grading scale for SUS scores by comparing more than 200 industrial usability studies and using the percentile ranges, resulting in grades “C” (scores of 62.7-72.5), “B” (scores of 72.6-78.8), and “A” (scores of 78.9-100). As a SUS score of 80 proves an above-average user experience, it has become a common industrial goal. This threshold was therefore used for interpreting our results.

**Covariates**

Consultations in general practice are attended by patients of different ages and educational as well as cultural backgrounds, who have a different quantities of digital interactions in everyday life. To determine whether these factors have an influence on the personally perceived usability, we surveyed age, sex, media usage, health literacy, and native language. Information about age and sex were part of the app–taken medical history. In addition to the SUS, we asked patients about which digital media tools were available to them in everyday life (possible answers: cell phone/smartphone, computer/laptop/notebook, tablet, television, none, and others; multiple answers were possible) and how many hours a day they used digital media (possible answers: 0≤1, 1≤2, 2≤3, 3≤4, or 4 or more hours). We asked three questions concerning health literacy as a proxy for education attainment, given that educational achievement is the central determinant of health literacy [28]. Questions covering the three aspects of finding/accessing, evaluating/appraising, and understanding health-related information and content were derived from the European health literacy survey [29,30] adapted for the German language (HLS-GER 2 [31]). The HLS-GER 2 uses a predefined 4-point Likert scale.

**Table 1. Items of the System Usability Scale (SUS) [19].**

<table>
<thead>
<tr>
<th>Items</th>
<th>English version of the statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUS 1</td>
<td>I think that I would like to use this system frequently</td>
</tr>
<tr>
<td>SUS 2a</td>
<td>I found the system unnecessarily complex</td>
</tr>
<tr>
<td>SUS 3</td>
<td>I thought the system was easy to use</td>
</tr>
<tr>
<td>SUS 4a</td>
<td>I think that I would need the support of a technical person to be able to use this system</td>
</tr>
<tr>
<td>SUS 5</td>
<td>I found the various functions in this system were well integrated</td>
</tr>
<tr>
<td>SUS 6a</td>
<td>I thought that there was too much inconsistency in this system</td>
</tr>
<tr>
<td>SUS 7</td>
<td>I would imagine that most people would learn to use this system very quickly</td>
</tr>
<tr>
<td>SUS 8a</td>
<td>I found the system very cumbersome to use</td>
</tr>
<tr>
<td>SUS 9</td>
<td>I felt very confident using the system</td>
</tr>
<tr>
<td>SUS 10a</td>
<td>I needed to learn a lot of things before I could get going with this system</td>
</tr>
</tbody>
</table>

aThe scores of negatively connoted SUS items were inverted when calculating the sum.
**Statistical Analysis**

Data from the app were saved into a database and subsequently exported to a tab separated format for further analyses. Participants with two or more missing values of the SUS questionnaire were excluded from statistical analysis. In the case of one missing SUS response, we substituted the missing value with a neutral score of 2, as this method has been used with the SUS in previous research [32].

Sociodemographic data are presented as number and percentage of patients for each categorical data point. Mean and SD were utilized for interval or ratio-scaled data, which has become a common industrial goal. Sociodemographic data were compared between participants with SUS scores <80 and ≥80 using the Fisher exact test for 2×2 tables or the Fisher-Freeman-Halton test for categorical variables and the Wilcoxon rank-sum test for continuous variables. A multiple linear regression was conducted using sex, age, native German language, health literacy score, media usage duration per day, sickness level of the participants, and number of stated complaints in the app as independent variables and the SUS score as the dependent variable. Additionally, the individual SUS items were compared according to sex, age (<65 years vs ≥65 years), German native language, and tablet usage with the Wilcoxon rank-sum test. Data are visually presented as boxplots and radar charts. All analyses were carried out using R (4.1.3 under a GNU license) with the packages fmsb [33], psych [34], tidyr [35], dplyr [36], and ggplot2 [37].

**Results**

**Patient Characteristics**

We aimed to include approximately 400 patients for this study. This target was set to be able to form subgroups and to ensure that all types of patient complaints were included in our sample, including those selected on a limited basis. In total, individual data from 397 participants were included, with 5 participants having one missing SUS item. Figure 2 shows the flowchart of included patients and Table 2 shows the patients’ characteristics.

![Flowchart of patient inclusion in the cross-sectional study capturing patients’ perceived usability of the app. SUS: System Usability Scale.](https://humanfactors.jmir.org/2024/1/e47755)
Table 2. Characteristics of the participants of the cross-sectional study capturing patients’ perceived usability of the app (N=397).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>225 (56.7)</td>
</tr>
<tr>
<td>Male</td>
<td>172 (43.3)</td>
</tr>
<tr>
<td><strong>Age (years), median (IQR)</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>35.0 (25.0)</td>
</tr>
<tr>
<td><strong>Age group (years), n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;30</td>
<td>152 (38.3)</td>
</tr>
<tr>
<td>30-65</td>
<td>223 (55.4)</td>
</tr>
<tr>
<td>65+</td>
<td>22 (6.3)</td>
</tr>
<tr>
<td><strong>Native language German, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>328 (82.6)</td>
</tr>
<tr>
<td><strong>Devices used regularly(^a), n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>Smartphone</td>
<td>389 (98.0)</td>
</tr>
<tr>
<td>Tablet</td>
<td>210 (52.9)</td>
</tr>
<tr>
<td>Computer/notebook</td>
<td>310 (78.1)</td>
</tr>
<tr>
<td>Television</td>
<td>296 (74.6)</td>
</tr>
<tr>
<td><strong>Media usage duration per day (hours), n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;2</td>
<td>89 (22.4)</td>
</tr>
<tr>
<td>2-4</td>
<td>174 (43.8)</td>
</tr>
<tr>
<td>&gt;4</td>
<td>134 (33.8)</td>
</tr>
<tr>
<td><strong>Self-assessed health literacy, median (IQR)(^b)</strong></td>
<td></td>
</tr>
<tr>
<td>Understanding doctor</td>
<td>2.0 (0.0)</td>
</tr>
<tr>
<td>Search and understand health information</td>
<td>2.0 (1.0)</td>
</tr>
<tr>
<td>Evaluate health information</td>
<td>1.0 (1.0)</td>
</tr>
<tr>
<td><strong>“How sick do you feel?”(^c), n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>I don’t feel sick</td>
<td>32 (8.1)</td>
</tr>
<tr>
<td>Just a little</td>
<td>70 (17.6)</td>
</tr>
<tr>
<td>Fairly</td>
<td>226 (56.9)</td>
</tr>
<tr>
<td>Very</td>
<td>61 (15.4)</td>
</tr>
<tr>
<td>Unbearably</td>
<td>6 (1.5)</td>
</tr>
</tbody>
</table>

\(^a\) Multiple selection possible.

\(^b\) Measured on a 4-point (0-3) Likert-scale (higher scores indicate higher health literacy levels).

\(^c\) Perceived severity of acute complaint.

**Usability for All Participants**

We found a mean total SUS score of 77.8 points, with 54.4% (216/397) of participants having SUS scores of 80 points or higher, indicating high usability of the app overall. Figure 3 shows boxplots of the individual items in which the scores were calculated for each statement. Irrespective of a positive or negative connotation, a higher score indicates a better result. The maximum score that can be achieved for each item is 10.
Usability Stratified by Sociodemographic Factors

We divided the sample into two groups with the cutoff at a SUS score of 80. Participants with a SUS score of at least 80 were significantly younger, reported higher levels of technology device usage, and higher levels of self-assessed health literacy compared to participants with a SUS score below 80 (Table 3).
Table 3. Sociodemographic variables of study participants stratified by System Usability Scale (SUS) score.

<table>
<thead>
<tr>
<th>Variable</th>
<th>SUS score&lt;80 (n=181)</th>
<th>SUS score≥80 (n=216)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex, n (%)</td>
<td></td>
<td></td>
<td>.09a</td>
</tr>
<tr>
<td>Female</td>
<td>94 (51.9)</td>
<td>131 (60.6)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>87 (48.1)</td>
<td>85 (39.4)</td>
<td></td>
</tr>
<tr>
<td>Age (years), median (IQR)</td>
<td>38.0 (26.0)</td>
<td>32.5 (22.0)</td>
<td>.002b</td>
</tr>
<tr>
<td>Age group (years), n (%)</td>
<td></td>
<td></td>
<td>.003c</td>
</tr>
<tr>
<td>&lt;30</td>
<td>55 (30.3)</td>
<td>97 (44.9)</td>
<td></td>
</tr>
<tr>
<td>30-65</td>
<td>111 (60.2)</td>
<td>112 (51.4)</td>
<td></td>
</tr>
<tr>
<td>65+</td>
<td>15 (9.4)</td>
<td>7 (3.7)</td>
<td></td>
</tr>
<tr>
<td>Native language German, n (%)</td>
<td>142 (78.5)</td>
<td>186 (86.1)</td>
<td>.05a</td>
</tr>
<tr>
<td>Devices used regularlyd, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smartphone</td>
<td>174 (96.1)</td>
<td>215 (99.5)</td>
<td>.03a</td>
</tr>
<tr>
<td>Tablet</td>
<td>84 (46.4)</td>
<td>126 (58.3)</td>
<td>.02a</td>
</tr>
<tr>
<td>Computer/notebook</td>
<td>136 (75.1)</td>
<td>174 (80.6)</td>
<td>.22a</td>
</tr>
<tr>
<td>Television</td>
<td>125 (69.1)</td>
<td>171 (79.2)</td>
<td>.03a</td>
</tr>
<tr>
<td>Media usage duration per day (hours), n (%)</td>
<td></td>
<td></td>
<td>.08c</td>
</tr>
<tr>
<td>&lt;2</td>
<td>48 (26.5)</td>
<td>41 (19.0)</td>
<td></td>
</tr>
<tr>
<td>2-4</td>
<td>69 (38.1)</td>
<td>105 (48.6)</td>
<td></td>
</tr>
<tr>
<td>&gt;4</td>
<td>64 (35.4)</td>
<td>70 (32.4)</td>
<td></td>
</tr>
<tr>
<td>Self-assessed health literacy, median (IQR)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Understanding doctor</td>
<td>2.0 (0.0)</td>
<td>2.0 (1.0)</td>
<td>.16b</td>
</tr>
<tr>
<td>Search and understand health information</td>
<td>2.0 (0.0)</td>
<td>2.0 (1.0)</td>
<td>.01b</td>
</tr>
<tr>
<td>Assess confidence of health information</td>
<td>1.0 (1.0)</td>
<td>1.0 (1.0)</td>
<td>.19b</td>
</tr>
<tr>
<td>&quot;How sick do you feel?&quot;, n (%)</td>
<td></td>
<td></td>
<td>.35c</td>
</tr>
<tr>
<td>I don't feel sick</td>
<td>11 (6.1)</td>
<td>21 (9.7)</td>
<td></td>
</tr>
<tr>
<td>Just a little</td>
<td>30 (16.6)</td>
<td>40 (18.5)</td>
<td></td>
</tr>
<tr>
<td>Fairly</td>
<td>102 (56.4)</td>
<td>124 (57.4)</td>
<td></td>
</tr>
<tr>
<td>Very</td>
<td>34 (18.8)</td>
<td>27 (12.5)</td>
<td></td>
</tr>
<tr>
<td>Unbearably</td>
<td>3 (1.7)</td>
<td>3 (1.4)</td>
<td></td>
</tr>
</tbody>
</table>

a Fisher exact test.
b Wilcoxon rank-sum test.
c Fisher-Freeman-Halton test.
d Multiple selection possible.
e Perceived severity of acute complaint.

A multiple linear regression predicting the SUS score was conducted, including sex, age, native German language, health literacy score, media usage duration per day, sickness level of the participants, and number of stated complaints in the app as independent variables (see Table 4). Age, sex, health literacy score, and German native language were significantly associated with SUS score. A higher age ($t_{385}=3.30$, $P=.001$) and male sex ($t_{385}=1.98$, $P=.05$) were negatively associated with SUS score, whereby a higher health literacy score ($t_{385}=2.83$, $P=.01$) and German as a native language ($t_{385}=2.51$, $P=.01$) were positively associated with SUS score.
Table 4. Multiple linear regression predicting the System Usability Scale sum score.

<table>
<thead>
<tr>
<th>Variable</th>
<th>β (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male sex (reference=female)</td>
<td>−3.21 (−6.40 to −0.02)</td>
<td>.05</td>
</tr>
<tr>
<td>age (per year)</td>
<td>−.17 (−.27 to −.07)</td>
<td>.001</td>
</tr>
<tr>
<td>German not native language (reference=yes)</td>
<td>−5.39 (−9.61 to −1.17)</td>
<td>.01</td>
</tr>
<tr>
<td>Does not use tablet (reference=yes)</td>
<td>−1.44 (−4.66 to 1.79)</td>
<td>.38</td>
</tr>
<tr>
<td><strong>Average daily media usage (reference=&lt;2 h)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-4 h</td>
<td>3.21 (−.95 to 7.37)</td>
<td>.13</td>
</tr>
<tr>
<td>&gt;4 h</td>
<td>.01 (−4.46 to 4.48)</td>
<td>.99</td>
</tr>
<tr>
<td>Health literacy score (scale 0-9)**</td>
<td>1.48 (.45 to 2.51)</td>
<td>.01</td>
</tr>
<tr>
<td>How sick do you feel? (score 1-5)**</td>
<td>−1.05 (−2.92 to .82)</td>
<td>.27</td>
</tr>
<tr>
<td>Number of stated complaints (1-11)</td>
<td>−.98 (−2.11 to .15)</td>
<td>.09</td>
</tr>
</tbody>
</table>

**a**Higher scores indicate a higher level of health literacy.  
**b**Perceived severity of acute complaint; higher values indicate a higher level of discomfort.

Differences in Individual Items of the SUS

Stratified according to sex, age, native language, and tablet usage (see Figure 4), significant differences were detected in SUS items 2 (“unnecessarily complex”), 4 (“need technical support”), 7 (“learn to use quickly”), 8 (“cumbersome to use”), and 10 (“needed to learn a lot”).

In comparing female and male respondents, all statements were rated more positively by female participants, except for items 1 (“would use frequently”) and 4 (“need technical support”). Female participants also scored significantly higher than male participants on items 2 (“unnecessarily complex”) (mean 7.82 vs 7.11; P=.02), 7 (“learn to use quickly”) (mean 8.11 vs 7.53; P=.05), 8 (“cumbersome to use”) (mean 8.57 vs 8.08; P=.05), and 10 (“needed to learn a lot”) (mean 9.14 vs 8.63; P=.03).

Respondents aged 65 years and older scored significantly higher on items 2 (“unnecessarily complex”) (mean 7.58 vs 6.50; P=.04), 4 (“need technical support”) (mean 8.72 vs 6.20; P<.001), and 10 (“needed to learn a lot”) (mean 9.05 vs 7.; P<.001) compared to their counterparts.

German native language speakers scored significantly higher on items 4 (“need technical support”) (mean 8.73 vs 7.75; P=.001), 8 (“cumbersome to use”) (mean 8.62 vs 7.10; P<.001), and 10 (“needed to learn a lot”) (mean 9.11 vs 8.01; P<.001) relative to nonnative speakers.

Lastly, patients who regularly use a tablet had significantly higher SUS scores on items 4 (“need technical support”) (mean 8.95 vs 8.11; P<.001) and 10 (“needed to learn a lot”) (mean 9.18 vs 8.64; P=.02) in comparison to those of participants who reported reduced levels of tablet use.
Figure 4. Mean System Usability Scale (SUS) score items (see Table 1) stratified according to sex (A), age (B), German as native language (C), and current tablet usage (D). bold: \( P < .05 \) (Wilcoxon rank-sum test).

**Discussion**

**Principal Findings**

In this study, we evaluated the usability of an app in taking medical histories in general practice directly from patients using the SUS [19].

The app achieved a mean SUS score of 77.9, which corresponds to a B+ grade on the curved grading scale [27] and represents a “better” product that does not necessarily need improvement [25]. Other medical devices, even those widely used at home, have lower SUS scores. Kortum and Peres [38] assessed the usability of home health care devices among students, thus representing relatively young, healthy, and well-educated participants. SUS scores for these devices ranged from 65 for...
an epinephrine injector to 67 for a pregnancy test kit and 81 for a thermometer, even with previous experience using these devices.

To ensure patients can be active participants in the digital medical history-taking process, the app must be easy and intuitive to use without technical introduction or support. This importance is reflected in items 3 (“easy to use”), 4 (“need technical support”), 7 (“learn to use quickly”), and 10 (“needed to learn a lot”). Mean values between 7.8 and 8.9 for these items indicate that intuitive use has been successfully addressed in the development of our app. Item 1, assessing the frequency of app usage, scored the lowest (6.2), which can be explained by the app’s implementation solely in a medical setting and not utilized regularly in leisure time. As such, this finding is the least meaningful for our purpose.

In a pilot study by Melms et al [39], a self-completed tablet-based digital questionnaire designed for collecting medical histories in an emergency department was found to score high with respect to perceived usability. The design and content were similar to those of our app; however, their questions were only based on the SUS, which does not allow direct comparison. Other comparable instruments, although also for emergency departments, have been tested for usability in pilot studies using self-developed satisfaction surveys [40,41], a single question, and researcher or staff documentation of a patient’s need for assistance [42]. In these studies, patients were mostly satisfied with the self-administered medical history-taking tools and reported good ease of use. Taken together, these results give hope that it is possible to design a medical history app that is perceived as user-friendly.

Nonetheless, obstacles to implementing a digital tool in general practice settings can be multifaceted. Surprisingly, we found that sex was significantly associated with usability; female participants had significantly higher SUS scores than male participants. The fact that men scored higher than women for item 4 (“need technical support”) suggests that men felt more confident than women with using the app. Previous studies demonstrated that men tend to report overconfidence in their abilities, especially in fields with a male connotation [43], which computer science certainly represents [44]. Therefore, it is unclear whether men really would have needed less help or whether they overestimated themselves in their technical skills.

Our study suggests that older people are more likely to have difficulties with the handling of such an app. This aligns with a study showing that from the retirement age of 65 years, digital media use among the German population begins to decline dramatically [45] and a positive attitude toward digitalization decreases with increasing age [46]. Older age has a negative impact on the broad usability score given to a user interface [25]. To that end, this study cannot definitively conclude if the older participants of this study actually perceived the app to be easier and more intuitive to use than younger participants; however, several items were rated significantly lower by older participants, indicating difficulties with the handling of the app. Since GP’s are consulted predominantly by older people [47], further research should focus on app testing with older patients to obtain specific feedback, including suggestions for improvement.

Having learned German as a native language was positively associated with a higher SUS score, although only patients with sufficient German language proficiency were included in the study [17]. This could be due to two different reasons: despite the app’s plain language, it is possible that some of the medical history questions or SUS items were not understood properly.

Daily media use was not associated with the SUS score, which suggests that the app is designed to also ensure that people with limited digital experience do not feel overstrained with its operation.

Limitations
Despite our efforts, this study comes with several limitations. The number of older participants (ie, aged 65 years and above) was relatively low in comparison to their constituents in GP practice settings [47]. One potential reason could be a more pronounced skepticism toward digital tools in older generations, leading to an increase in refusal for participation in the study among older patients. However, as no screening lists were maintained, this is mere speculation. A screening list should be obtained in future studies to be able to characterize individuals who declined participation. Another consideration is that people with lower levels of digital media literacy use may not have agreed to participate in the study.

Data collection was performed during the SARS-CoV-2 pandemic, which may have disproportionately impacted study participants as certain patient groups may have avoided seeing a doctor or were more likely to refuse to participate in the study to avoid unnecessary contact. This could have included especially vulnerable groups such as older people or those with multimorbid conditions.

The Likert scale of the SUS questions shown with clickable singular dots was replaced by a slider on December 8, 2021. In the dot-based representation, it was compulsory to make an entry before continuing, whereas the slider was automatically set to the neutral center and could be shifted. This may have led to incorrectly rated items. For example, this may have occurred in instances of the internet faltering or the patient having double-clicked without noticing. Since there were repeated questions about the word “Inkonsistenzen” (SUS item 1), we replaced it by the more common synonym “Unstimmigkeiten” (English translation: discrepancies). Furthermore, hardware as well as the operating system may have influenced the evaluation of the personally perceived usability of a system [48]. For this study, iPad Minis with the iOS operating system were exclusively used. Therefore, possible differences in assessment related to the operating system and hardware are not part of this study.

The SUS is able to classify the usability of a system but is unable to identify specific usability issues nor capture the usability of the system in its entirety. For a more in-depth usability evaluation, different methods could be used (eg, interviews and observations). During data collection, staff were able to observe usability problems. In their observations, multiple-choice, single-choice, and hybrid questions as well as the slider did not...
appear to cause any difficulties. In contrast, problems concerning the handling of the app arose when participants were required to input free-text entries (eg, age, height, weight). Further, some study participants were unclear on how to open and close the on-screen keyboard. Some participants also did not understand that the figure on which a pain or an injury could be assigned to a body region (see Figure 1E) could be rotated by clicking on an icon at the bottom left of the screen. This means that, for example, back pain may have been falsely reported as abdominal pain. Lastly, an unstable internet connection arose during data collection, which caused the app to be unresponsive intermittently. These factors may have influenced the SUS score.

Conclusion

The app examined in this study for medical history taking passes the usability test based on the SUS and appears to function on par with other digital tools that have become well-integrated in our everyday lives. However, not all people adapted equally well to the app. For successful implementation, all end users, regardless of age, technical affinity, health literacy, or preferred language, must be able to use such a tool. Only if that is attained, providing practical digital solutions can contribute to the efficient and effective delivery of health care services. Therefore, further research should focus on the identification of causes for difficulties of using the app as well as finding appropriate solutions.

Acknowledgments

We would like to thank Julie Ngo, MD for proofreading. We sincerely thank all patients for their participation. We are very grateful to the physicians and their teams who allowed data collection in their practices despite the SARS-CoV-2 pandemic and the extraordinary burden that came along for health care workers. We thank the team at aidminutes GmbH for their commitment and support. This research is funded by the German Innovation Fund (funding number 01VSF19050) of the Federal Joint Committee (G-BA). The funders have not influenced the design of this study and did not play any role during its implementation (eg, data collection and analysis, interpretation, and publication of the results).

Data Availability

The data sets used and analyzed during the study are not publicly available due to the decision of the research ethics board but can be obtained from the authors upon reasonable request within a data sharing agreement.

Authors' Contributions

KA wrote the original draft. KA and EMN were mainly responsible for writing the manuscript. KA, EMN, and DS conceived the study design and analyzed the data. FM and CJ revised the manuscript. All authors have read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Original and German versions of the System Usability Scale (SUS) and the items customized for this study.

References


33. Package ‘fsmb’. CRAN R project. 2023. URL: https://cran.r-project.org/web/packages/fsmb/fsmb.pdf [accessed 2023-12-29]
35. tidyR: Tidy Messy Data. URL: https://tidyr.tidyverse.org/ [accessed 2022-05-19]
36. dplyr: A Grammar of Data Manipulation. URL: https://dplyr.tidyverse.org/ [accessed 2023-12-29]

Abbreviations

- **DASI**: digitally assisted information acquisition before medical consultation
- **GP**: general practitioner
- **HLS-GER 2**: European health literacy survey adapted for the German language
- **SUS**: System Usability Scale

---

**Edited by A Kashnirak; submitted 31.03.23; peer-reviewed by A Simmenroth, R Marshall; comments to author 27.08.23; revised version received 10.10.23; accepted 20.11.23; published 05.01.24.**

**Please cite as:**

Albrink K, Schröder D, Joos C, Müller F, Noack EM

Usability of an App for Medical History Taking in General Practice From the Patients’ Perspective: Cross-Sectional Study

JMIR Hum Factors 2024;11:e47755

URL: https://humanfactors.imij.org/2024/11/e47755
doi: 10.2196/47755
PMID:36180798
Original Paper

A Novel Continuous Real-Time Vital Signs Viewer for Intensive Care Units: Design and Evaluation Study

Shiming Yang, PhD; Samuel Galvagno, MD, PhD; Neeraj Badjatia, MD; Deborah Stein, MD; William Teeter, MD; Thomas Scalea, MD; Stacy Shackelford, MD; Raymond Fang, MD; Catriona Miller, PhD; Peter Hu, PhD; VS viewer study group

1Department of Anesthesiology, University of Maryland School of Medicine, Baltimore, MD, United States
2Department of Neurology, University of Maryland School of Medicine, Baltimore, MD, United States
3Department of Surgery, University of Maryland School of Medicine, Baltimore, MD, United States
4Emergency Medicine, University of Maryland School of Medicine, Baltimore, MD, United States
5United States Air Force Academy, Colorado Springs, CO, United States
6Department of Surgery, Johns Hopkins University School of Medicine, Baltimore, MD, United States
7See Acknowledgments

Corresponding Author:
Shiming Yang, PhD
Department of Anesthesiology
University of Maryland School of Medicine
11 S Paca St, LL01
Baltimore, MD, 21201
United States
Phone: 1 4103284179
Email: syang@som.umaryland.edu

Abstract

Background: Clinicians working in intensive care units (ICUs) are immersed in a cacophony of alarms and a relentless onslaught of data. Within this frenetic environment, clinicians make high-stakes decisions using many data sources and are often oversaturated with information of varying quality. Traditional bedside monitors only depict static vital signs data, and these data are not easily viewable remotely. Clinicians must rely on separate nursing charts—handwritten or electric—to review physiological patterns, including signs of potential clinical deterioration. An automated physiological data viewer has been developed to provide at-a-glance summaries and to assist with prioritizing care for multiple patients who are critically ill.

Objective: This study aims to evaluate a novel vital signs viewer system in a level 1 trauma center by subjectively assessing the viewer’s utility in a high-volume ICU setting.

Methods: ICU attendings were surveyed during morning rounds. Physicians were asked to conduct rounds normally, using data reported from nurse charts and briefs from fellows to inform their clinical decisions. After the physician finished their assessment and plan for the patient, they were asked to complete a questionnaire. Following completion of the questionnaire, the viewer was presented to ICU physicians on a tablet personal computer that displayed the patient’s physiologic data (ie, shock index, blood pressure, heart rate, temperature, respiratory rate, and pulse oximetry), summarized for up to 72 hours. After examining the viewer, ICU physicians completed a postview questionnaire. In both questionnaires, the physicians were asked questions regarding the patient’s stability, status, and need for a higher or lower level of care. A hierarchical clustering analysis was used to group participating ICU physicians and assess their general reception of the viewer.

Results: A total of 908 anonymous surveys were collected from 28 ICU physicians from February 2015 to June 2017. Regarding physicians’ perception of whether the viewer enhanced the ability to assess multiple patients in the ICU, 5% (45/908) strongly agreed, 56.6% (514/908) agreed, 35.3% (321/908) were neutral, 2.9% (26/908) disagreed, and 0.2% (2/908) strongly disagreed.

Conclusions: Morning rounds in a trauma center ICU are conducted in a busy environment with many data sources. This study demonstrates that organized physiologic data and visual assessment can improve situation awareness, assist clinicians with recognizing changes in patient status, and prioritize care.

(JMIR Hum Factors 2024;11:e46030) doi:10.2196/46030

https://humanfactors.jmir.org/2024/11/e46030

JMIR Hum Factors 2024 | vol. 11 | e46030 | p.136
(page number not for citation purposes)
KEYWORDS
clinical decision-making; health information technology; intensive care units; patient care prioritization; physiological monitoring; visualization; vital signs

Introduction
Clinicians working in intensive care units (ICUs) must be able to see, understand, and respond quickly to the complex and ever-changing clinical environment of the ICU. They need to be able to collect, analyze, and interpret what is happening and what it means [1]. Situational awareness is essential for ICU clinicians to provide safe and effective care to their patients. When clinicians have good situational awareness, they are better able to identify and respond to changes in their patients’ condition and to coordinate care with other members of the health care team. However, clinicians are immersed in a cacophony of alarms and a relentless onslaught of data. Within this frenetic environment, clinicians make high-stakes decisions using multiple data sources and are often oversaturated with information of varying quality. Within this environment, clinicians make high-stakes decisions using multiple data sources and are often oversaturated with information of varying quality. While modern hospitals are equipped with bedside monitors collecting various physiological data in a real-time, continuous, and automated way, these data are not always easily accessible remotely or available to be viewed as a continuous trend [2].

During the design of the VS viewer for ICU, our goal was to create a novel physiological data displayer that can reduce ICU clinicians’ workload, enhance clinical decision-making, and improve communication in a noisy and confined ICU environment. To achieve the goal, we considered the factors of usability and patient safety, which can be closely related in this application. For usability, current bedside monitors often suffer from insufficient time windows to display physiological trends, a lack of clear indications of patients’ physiological status, and a lack of overview of multiple patients for prioritizing [4]. To enhance the clinicians’ efficiency while maximizing patients’ safety, we adopted the following design strategies: First, the viewer should reduce the information overload for clinicians to access patients’ physiological data, current or past, individual or group [8-10]. Second, it should be compatible with the existing patient monitor system so that clinicians can reuse their existing knowledge about the monitor, which may increase the acceptance of the VS viewer [8]. Third, in the user experience design, the viewer should place the user in control [11]. It should use simple colors and graphs to convey efficient information while still providing detailed data for advanced users to access with simple operations [12]. Fourth, the viewer should have reasonable reliability for patients’ safety. Redundance was introduced in the design for key components in the system, such as the data collection, database, and web server [7].

The VS viewer adopted a client-server architecture. The server handles 2 types of clients: the bedside monitors and the users. It receives and persists in real-time physiologic data that are processed data adds an additional burden on ICU clinicians who work in a dynamic environment with voluminous decision-making requirements. Traditional bedside monitors only show a single patient’s instantaneous (static) vital signs (VS) data, limiting the clinician’s scope to view a patient’s physiological trajectory within a clinically meaningful period of time. Clinicians must rely on separate nursing charts—handwritten or electronic—to review a patient’s physiological status. Moreover, auditory alarms often cause “alarm fatigue” instead of increasing situational awareness [3]. Many bedside monitors only display 1 or 2 patients’ information; the ability to view an entire unit or ward allows a clinician to prioritize attention to those in most need of critical care support [4]. Improved visualization of patient information may help clinicians cope with information overload in critical care settings by improving situational awareness and supporting clinical decision-making [5].

An automated physiological data-organizing and information-summary system that presents aggregated information from multiple data sources while providing at-a-glance summaries of clinical data can assist ICU clinicians with prioritizing care for multiple patients.

Developed initially for use in aircraft transporting multiple patients who are critically ill, this VS viewer has 2 outcomes of direct and important clinical applicability. First, the VS viewer can provide clinicians with the capability to monitor individual patient trends, improving overall decision-making. Since patients in the ICU require multiple life support treatments to ensure ideal long-term outcomes, improved display of VS patterns could improve patient assessment and clinical decision-making. Second, the VS viewer system allows remote monitoring of groups of patients through a display that provides clinicians with the ability to quickly identify patients in need of rapid intervention. The objective of this work is to evaluate the use of a VS viewer in ICUs at a high-volume level 1 trauma center. We hypothesized that clinicians would subjectively report improved situational awareness and enhanced ability to make clinical decisions with the use of a VS viewer.

Methods

Data and System Design

In the ICUs of the University of Maryland Medical System, GE Marquette Solar 7000/8000 (General Electric) patient VS monitors are networked to provide a collection of real-time patient VS data streams. Each patient monitor collects real-time 240 Hz waveforms and 0.5 Hz trend data, which are transferred through the secure intranet to a dedicated BedMaster server (Excel Medical Electronics) and archived [6]. To increase the system’s availability and reliability, a triple-redundant design was used, in which 3 BedMaster servers were used in parallel to collect data from all bed units [7]. Physiological data collected through this system, when they are displayed on the GE Marquette monitor, include electrocardiographic, photoplethysmographic, carbon dioxide, arterial blood pressure, and intracranial pressure (ICP), among others. Trends include heart rate (HR), respiratory rate, temperature, oxygen saturation, end-tidal carbon dioxide, and ICP, among many others. This information provides continuous VS data that relays important physiological information regarding brain perfusion, cardiac stability, overall tissue perfusion, and respiratory status.
transmitted from the bedside monitors. A database records each bedside VS value, bed name, and timestamp. The server also responds to users’ requests for viewing data within a given time frame. To continuously present the latest data to the user with low latency, the VS viewer uses the asynchronous Javascript and XML technique to pull the most recent data from the database every minute [13]. Such a method allows the VS viewer to automatically redraw all VS trajectories without refreshing the entire viewing page.

The VS viewer provides a rich interface for data monitoring, exploration, and recording. Data are depicted according to each clinical area of operations, such as the trauma resuscitation unit or emergency department, operating room, computed tomography suite, and individual critical care units. Figure 1 demonstrates the grouping of bed units. On the left panel, a list of all groups can be used as a shortcut to bed units. Selecting a specific unit, a default 24-hour view is displaced for shock index (SI=HR/systolic blood pressure), HR, systolic blood pressure, ICP, cerebral perfusion pressure, brain trauma index, and end-tidal carbon dioxide concentration. If ICP data are not collected, the space is used to plot the next available VS, optimizing the view.

When a bed is selected, a page for this bed (unit view) is displayed. Figure 2 demonstrates the structure of the information. The page is partitioned into multiple areas for navigation, viewing, and tools. Its center is assigned for presenting the selected patient’s physiologic data in a time frame (up to 72 hours). VS trajectories are stacked vertically in order of predefined importance. The bottom is reserved for plotting bar segments of all VS that summarize the colored warnings without showing the value changes. This provides a summary of all available VS trends in a condensed space, which could be used to view the physiological stability of the patient over time. To provide an at-a-glance view of other rooms in this group, the left panel lists all the rooms in the current group and updates their VS trajectories in real time. The color-coded warning in the thumbnails enhances situation awareness even when the users are focusing on 1 patient.

The VS viewer has additional diagnostic tools. For example, SI is a commonly used blood transfusion diagnosis tool [14]. The VS viewer adds a 2D SI diagram to show a changing trajectory (Figure 3). To present the temporal information, a heat map is plotted, ranging from blue (cold) to red (warm); blue colors represent past events, whereas red colors represent current data trends. Similarly, the brain trauma index (which is ICP or cerebral perfusion pressure) can also be visualized in the 2D plot [15].

Figure 1. The vital sign viewer in the “group” mode, with a default 24-hour display.
Figure 2. Vital sign (VS) viewer in the “unit” mode, with default 24-hour display. Labeled area 1: navigation menu to other room groups. Area 2: title information for room name, current time, and the next update time. Area 3: user portal. Area 4: list of beds in the same group with their current VS thumbnails. Area 5: the main area to display selected room VS trajectories and the summarization with color-coded patterns. Area 6: diagnostic tools for 2D scatter plots of shock index (=heart rate/systolic blood pressure) and brain trauma index (=intracranial pressure/cerebral perfusion pressure). Area 7: functional buttons for selecting various time ranges for viewing.

Figure 3. An example 2D shock index plot. The colored scatter plot shows the change in shock index (heart rate/systolic blood pressure) from past (blue) to recent (red), thereby depicting a 3-day change in worsening shock index.
Clinical Thresholds

Colored warnings are an effective means to gain a clinician’s attention and may be more effective than audible alarms, especially in a noisy, busy, and confined environment [16]. In the VS viewer, VS trajectories with colors may be viewed to highlight the sections where the VS are outside of normal clinical thresholds. For example, too low or too high HR segments are displayed differently from normal HR. Clinical thresholds for VSs were developed after surveying 47 clinicians (24 medical doctors, 18 registered nurses, and 5 respiratory therapists). Among them, 36 clinicians were from the University of Maryland, Baltimore, and 11 from the University of Cincinnati. After the survey was completed, a team of clinicians met to review the results to reach a consensus on the viewer’s opinion of their visual appearance. Multimedia Appendix 1 summarizes the optional threshold distributions for some important VS. Based on these threshold values, a consensus set of color-coded cutoffs was determined (Multimedia Appendix 2). These values were set as fixed parameters under consideration of a simplified and consistent user interface.

Survey Design

Clinicians who were scheduled to work in the ICU or on the trauma teams were contacted and trained on how to use the VS viewer. Once trained, ICU and team clinicians were asked to participate in the study. Clinicians were surveyed anonymously from Tuesday to Friday and were asked to conduct rounds normally, using data reported from nurse charts and briefs from fellows to inform their clinical decisions. None of those clinicians participated in the design of the VS viewer. A total of 2 questionnaires were designed to collect clinicians’ opinions about a patient’s condition and satisfaction with the VS viewer. Clinicians were given a preview survey upon their assessment and formulation of their plan for each patient after traditional rounds and before accessing the viewer. Immediately following the completion of the pre-view survey, the VS viewer was presented to the clinicians on a tablet, displaying the patient’s past physiologic data visualized and summarized for up to 72 hours. After reviewing the viewer for up to 1 minute, clinicians completed the postview questionnaire. In both questionnaires, the clinicians answered questions regarding the patient’s stability, status, and need for a higher or lower level of care. In the post-view questionnaire, clinicians were also asked if they intentionally planned to implement any of the following interventions after seeing the viewer: (1) changing any current medications, (2) ordering additional medications, (3) ordering additional diagnostic tests, (4) changing ventilation settings, (5) ordering additional labs, (6) physically reexamining this patient, (7) providing fluid bolus, or (8) providing a blood transfusion.

Statistical Methods

A participant’s perceiving of the VS viewer’s usefulness is represented by a vector consisting of the percentage of the 5 categories (strongly agree, agree, neutral, disagree, and strongly disagree) that he or she assigned to the question “the viewer enhanced my understanding of the patient’s condition.” We used the Ward method, a hierarchical clustering method, with Manhattan distance to group the participants based on their ratings to the question “the viewer enhanced my understanding of the patient’s condition” [17,18]. Between those clusters, we compared the participants’ opinion changes on the patients’ conditions in 7 questions (Table 1) before and after using the viewer. The chi-square test was used to compare percentage differences.

Table 1. The number of opinion changes for 7 questions (Q1-Q7) before and after seeing the viewer, with respect to the 5 clustered user types.

<table>
<thead>
<tr>
<th>Questionsa</th>
<th>Total changes, n (%</th>
<th>Unique participants, n</th>
<th>C1, n</th>
<th>C2, n</th>
<th>C3, n</th>
<th>C4, n</th>
<th>C5, n</th>
<th>Like (C1 and C2), n (%)</th>
<th>Dislike (C3, C4, and C5), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>129 (14.2)</td>
<td>16 (66.7)</td>
<td>46</td>
<td>31</td>
<td>10</td>
<td>42</td>
<td>0</td>
<td>77 (59.7)</td>
<td>52 (40.3)</td>
</tr>
<tr>
<td>Q2</td>
<td>112 (12.3)</td>
<td>15 (62.5)</td>
<td>38</td>
<td>34</td>
<td>8</td>
<td>32</td>
<td>0</td>
<td>72 (64.5)</td>
<td>40 (35.7)</td>
</tr>
<tr>
<td>Q3-6</td>
<td>145 (16)</td>
<td>18 (75)</td>
<td>58</td>
<td>54</td>
<td>3</td>
<td>30</td>
<td>0</td>
<td>112 (77.2)</td>
<td>33 (22.8)</td>
</tr>
<tr>
<td>Q7</td>
<td>92 (10.1)</td>
<td>17 (70.8)</td>
<td>20</td>
<td>32</td>
<td>9</td>
<td>31</td>
<td>0</td>
<td>52 (56.5)</td>
<td>40 (43.5)</td>
</tr>
</tbody>
</table>

Please refer to Textbox 1 for the question.
Textbox 1. Questions.

- Q1: Having reviewed the last 24 hours of information during rounds and before and after seeing the 24-hour viewer, do they feel that in the past 24 hours the patient has shown evidence of (a) infection, (b) hemodynamic instability, (c) uncontrolled bleeding, or (d) respiratory deterioration?
- Q2: Over the past 24 hours, has the patient’s condition (a) improved significantly, (b) improved slightly, (c) unchanged, (d) deteriorated slightly, or (e) deteriorated significantly?
- Q3: Can the patient be transferred to a lower level of care?
- Q4: Can the patient be transferred to a higher level of care?
- Q5: Does the patient have a traumatic brain injury?
- Q6: Did the patient have intracranial pressure problems in the past 24 hours?
- Q7: Due to the viewer, do they plan for any changes in interventions, including (a) changing any current medications, (b) ordering additional medications, (c) ordering additional diagnostic tests, (d) changing ventilation settings, (e) ordering additional labs, (f) physically reexamining this patient, (g) providing a fluid bolus, or (h) providing a blood transfusion?

Note: These are the questions referenced in Table 1.

Ethical Considerations

The study has been approved by the institutional review board of the University of Maryland School of Medicine (HP-00063086).

Results

Survey Collection

From February 2017 to June 2017, the survey team followed clinicians who agreed to take the surveys. A total of 908 surveys were collected from 24 participants with unbalanced proportions. Among the 908 rounds, 48 (5%) were patients who were newly admitted, and 860 (95%) were not. When asked if the VS viewer enhanced their understanding of the patient’s condition, clinicians strongly agreed 45 (5%) times, agreed 514 (56.6%) times, were neutral 321 (35.4%) times, disagreed 26 (2.9%) times, and strongly disagreed 2 (0.2%) times. Figure 4 lists the total surveys each participant contributed and the proportions of ratings on whether the viewer enhanced his or her understanding of the patient’s condition during a round.

Results show that physicians’ clinical assessments and plans could be influenced by viewing the VS viewer for 1 minute or less, indicated by a “yes” answer to at least 1 of the 8 questions (Q7 in the survey). Of the 908 rounds, a total of 92 (10.1%) rounds had at least 1 “yes” as planning on some changes to the interventions. The most common change was (Q1) changing current medications (36/908, 4%). The next most common changes were (Q6) physically reexamining the patient (31/908, 3.4%), (Q2) ordering additional medications (20/908, 2.2%), and (Q7) providing a fluid bolus (20/908, 2.2%).

We used the Ward method with Manhattan distance to group the participants based on their ratings to the question “the viewer enhanced my understanding of the patient’s condition” [17]. For example, 1 participant contributed 62 surveys and rated 2 “strongly agree,” 22 “agree,” 32 “neutral,” 5 “disagree,” and 1 “strongly disagree.” The vector of percentages (0.03, 0.35, 0.52, 0.08, and 0.02) represents the overall rating that this participant had about the viewer. The 24 participants were clustered into 5 groups, as shown in Figure 5. The 5 groups correspond to the participants who are mostly in favor (C1) of the viewer to those least in favor (C5). There are 6 in C1, 6 in C2, 3 in C3, 7 in C4, and 2 in C5, which shows a very balanced grouping, with half of the participants in the C1 and C2 groups and the other half in the other 3 clusters. This shows that the sampled rounds were done by participants with almost similar proportions of different attitudes toward the viewer. In other words, the survey team sampled the rounds randomly enough so that the collected data were not biased by participants with certain preexisting feelings about the viewer.


Figure 4. Distribution of each participant’s rating on if the viewer enhanced his or her understanding of the patient’s condition during a round.

Figure 5. Clusters of participants with similar feelings about the viewer. In total, 24 unique participants are grouped into 5 groups, corresponding to “strongly favor,” “favor,” “neutral,” “dislike,” and “strongly dislike.”.

Comparisons
We analyzed the opinion changes before and after seeing the viewer, regarding the patient’s stability, status, and need for a higher or lower level of care. Instead of summarizing the total changes in opinions, we compared them with respect to the clusters of user types. The participants who were “neutral” (C3) or “strongly dislike” (C5) had low numbers of opinion changes.
for all 7 questions. Those who were in clusters C1, C2, and C3 had more numbers of opinion changes (Table 1). For simplicity, we can further group the participants into 2 types: those who liked the VS viewer (C1 and C2) and those who disliked it (C3, C4, and C5). The clinicians who liked the VS viewer had a higher rate of changed opinions than those who disliked the VS viewer regarding Q1 to Q6 (Q1: 59.7% vs 40.3%, Q2: 64.3% vs 35.7%, and Q3-6: 77.2% vs 22.8%). When asked if they planned for any changes for interventions (Q7), there was no significant difference between the 2 major groups of clinicians (56.5% vs 43.5%, P=.10).

Discussion

Principal Results

With the development of sensor and computing technologies, vast amounts of high-quality, continuous electronic data, including VS, alarms, and clinical interventions, are collected at the bedside. Those data have the potential to provide an unprecedented view of dynamic physiologic responses to injury, illness, and treatments. Therefore, data gathered from bedside could assist clinicians in care planning and decision support. However, massive amounts of data that are not well organized or presented still create a barrier for clinicians making full use of them in a busy resuscitation or intensive care environment. Bedside monitors often only display instantaneous readings or a short strip of recent physiologic VS for diagnosis. Clinicians need to rely on separate nursing charts, handwritten or electronic, to review a patient’s developing conditions. The VS viewer, which automates physiological data by displaying clear color-coded trends, presents aggregated information from multiple data sources, provides at-a-glance summaries of clinical data, and assists with the prioritization of care for multiple patients.

The use of the VS viewer was subjectively assessed with 908 observations from clinicians working in ICUs at a high-volume level 1 trauma center. Clinicians generally perceived the use of the VS viewer favorably, as evidenced by survey data. The VS viewer was originally developed for the United States Air Force Critical Care Air Transport Teams [19,20]. Critical Care Air Transport Teams transport up to 3 patients who are critically ill in the back of the aircraft, allowing trauma surgeons to perform far-forward damage control surgery, knowing that these patients could be quickly transported rearward with full support. This rapid transport of complex patients with multisystem trauma, shock, burns, and respiratory failure who are in hemodynamic flux requires continual resuscitation, stabilization, advanced care, and life-saving interventions during air transport; however, currently available advanced ICU monitoring systems suitable for the needs of such patients were developed for use in stable, hospital-based settings, not in the crowded, noisy, vibrating, and sometimes frankly jetting environment of air evacuation or long-distance air transport. The noise levels, confined space, limited access to patients, vibration, and overall limited patient visibility make using a VS viewer advantageous in such a setting. Such technology can also be valuable in enhancing emergency medical personnel’s decision-making for initial triage. While traditional VS are useful in guiding prehospital care and triage, they represent isolated points in time, and trends and fluctuations in vitals may not be apparent.

In this study, we set the clinical thresholds for colored warnings to be uniform across all beds. This was to make the user interface simplified and more consistent during a survey. Additionally, a set of predefined thresholds from a group of experienced clinicians could be a useful out-of-the-box feature when the VS viewer is deployed in the field. That said, the clinical thresholds could be personalized for each bed. For example, if the bedside monitor allows alarm threshold settings, such settings could be used as the colored warning thresholds in the VS viewer for each bed.

The VS viewer has expanded from ICUs to trauma resuscitation units, operating rooms, neuro ICUs, and pediatric ICUs at the University of Maryland Medical Center. In 2020, during the COVID-19 pandemic, it was deployed to monitor 150 beds in biocombinated units to reduce the risk of infection and improve efficiency for clinicians in treating their patients.

Innovations

The VS viewer is a multipatient physiological monitor. To the best of our knowledge, we could not find any articles that describe a viewer system with a similar design. In a comprehensive review by Waller et al [5], a total of 17 information displays in ICU settings were designed for specific disease states or body systems, such as cerebral perfusion monitoring for individual patients or monitoring for arterial blood gas trends. The novel user interface presented in this study was designed with the aim of conveying information more efficiently to ICU clinicians in a noisy, confined, and busy environment. It uses color-coded warnings to indicate a patient’s status and highlight data that needs attention. The side panel provides a peek at the physiological status of other patients, which can help clinicians keep an eye on other patients even if their attention is focused on a single patient. It uses advanced web front-end techniques to hide large quantities of data behind simple line charts and reveal them when needed.

Clinical Impact

The use of the VS viewer can have several possible influences on clinical assessment and plans. It can help clinicians quickly recognize critical changes in the patient’s physiologic status and provide early interventions to prevent further deterioration. The VS viewer can potentially improve patient outcomes by providing clinicians with a concise overview of key information, reducing cognitive load and errors, and improving compliance with evidence-based safety guidelines [12]. It may also help to improve communication efficiency within the ICU team by providing easy access to a shared platform of patient longitudinal data. It can reduce the workload of the ICU team by automating routine tasks such as extracting data from nursing charts.

To prioritize care in high-volume ICUs, intensive care clinicians must be able to rapidly identify physiologic events and the need for intervention. The VS viewer can help organize a large amount of data in a busy, noisy ICU environment where close monitoring of patients who are critically ill is essential to detect potentially harmful physiologic trends. The presentation of data with temporal, color-coded patterns, and the ability of the
VS viewer to provide at-a-glance data for entire units is advantageous for clinicians working in high-volume ICUs.

The color-coded patterns may reduce the “alarm fatigue” issue in noisy ICUs. The noise burden is common in modern physiologic monitoring systems, and has been recognized as a critical patient safety concern in the hospital care setting [21-23]. In noisy environments, such as ICUs, helicopter transportation, or aeromedical evacuation, loud and continuous alarms could reduce their specificity in getting clinicians’ responses. Another issue with audible alarms is that they are transient and cannot be replayed once they are gone. While the visual alert patterns could show the longitudinal patterns of physiologic change.

**Related Work**

The VS viewer with organized and easy-access information could be part of the effort to build the smart ICU or the tele-ICU. The concepts of smart ICU and tele-ICU aim to maximize the use of bedside clinical expertise in assessing and treating patients by providing integrated monitoring and actionable information [24-26]. A survey study of 86 ICU staff in a German university hospital summarized that health providers expect ICU monitoring could be improved by reducing false alarms, using wireless sensors and mobile devices, preparing for the use of AI, and enhancing the digital literacy of ICU staff [27,28]. The VS viewer could be used in both centralized and decentralized architectures of tele-ICU for extending coverage and facilitating patient transfer between hospitals because of its flexible configuration of grouping ICU beds virtually [29]. By making essential clinical information available remotely, the VS viewer allows clinicians to provide care plans when on-site support is infeasible or limited [30,31]. It may potentially reduce exposure to contagious diseases and, hence, increase patient safety.

With continuous physiologic data and other clinical information, the VS viewer has the ability to process real-time data into predictive algorithms, which is also desired for tele-ICU [30]. Beyond being a plain display, the VS viewer could embed risk-prediction algorithms that use continuous VS as inputs and may promote more efficient interventions to reduce ICU risk [31]. For example, ICU mortality prediction [32,33], secondary insults after severe brain traumatic injury [34], needs for transfusion [35,36], and neurologic decline in the ICU [37] are reported to have good predictive performances by using variables derived from continuous VS. We have also shown that using risk scores calculated from continuously measured VS, patients requiring endovascular resuscitative interventions can be identified with high accuracy [38]. Moreover, the VS viewer could serve as a platform for predictive model diagnosis by providing clinicians with explainable artificial intelligence [39]. With patient VS data, we can use the Shapley Additive exPlanations algorithm to calculate each variable’s contribution to the prediction result [40]. Therefore, the clinicians would know not only the prediction but also the contribution of each variable to the prediction. Such information may help clinicians make more personalized care plans.

**Limitations**

There are limitations to this work that are worth noting. We collected data from a large number of ICU clinicians compared to trauma team clinicians. Trauma team clinicians are surgeons responsible for the same patient throughout the entire length of stay, regardless of the acuity of the patient. ICU clinicians are intensivists and are only responsible for patients in the ICU. Hence, disparities between both groups of clinicians are inevitable, as each group has different clinical perspectives and patient workloads. As occurs in nearly all survey work, response rates and receptiveness to the surveys varied. Some clinicians were more amenable to being surveyed compared to others. In the collected forms, there were more surveys from some clinicians than from others. To reduce this potential bias, we clustered the participants based on their overall rating on each round, from which we estimated each participant’s a priori attitude toward using this viewer. The results show that there was a balanced “favoring” and “non-favoring” of using this viewer.

We only evaluated the viewer based on clinicians’ satisfaction and efficiency (potential changes in interventions before and after seeing the viewer). In future studies, randomized controlled trials can be designed to analyze the viewer’s impact on patients’ outcomes and safety [12].

**Conclusions**

We designed, implemented, and evaluated an automated physiologic data organizer and visualization platform. It provides at-a-glance summaries and assists with prioritizing care for multiple patients. The VS viewer demonstrates a method to assemble large quantities of data from multiple sources and represents trends in each patient’s condition with simple color codes, greatly improving situational awareness. It has the potential to be used in en route care, hospitals with multiple branches, and understaffed hospitals in remote areas. The survey shows that organized physiologic data and visual assessment could assist clinicians in recognizing changes in patient status and prioritizing care.

**Acknowledgments**

This work was partially funded by USAF 8650-14-2-6D19 and W81XWH-17-C-0034.

The authors thank the VS viewer study group: Colin Mackenzie, MD, Sarah Wade, BS, Lauren Hartsky, BS, Thomas Grissom, MD, Napoleon Roux, MD, Doug Floccare, MD, Ashton Engdahl, BS, Scott Murray, BS, Rosemary Kozar, MD, Sarah Murthi, MD, Laura Buchanan, MD, Neal Reynolds, MD, Gunjan Parikh, MD, Megan Brenner, MD, without whose help the data collection for this study would not have been possible.
Conflicts of Interest

PH, DS, Colin Mackenzie (part of the VS viewer study group), TS, and SY have US Patent Application 17/676,657 filed on February 21, 2022, titled “Method and Apparatus for Monitoring Collection of Physiological Patient Data.”

Multimedia Appendix 1
Surveyed thresholds for heart rate, systolic and diastolic blood pressure, blood oxygen saturation, and temperature.

References


Abbreviations

HR: heart rate
ICP: intracranial pressure
ICU: intensive care unit
SI: shock index
VS: vital signs

©Shiming Yang, Samuel Galvagno, Neeraj Badjatia, Deborah Stein, William Teeter, Thomas Scalea, Stacy Shackelford, Raymond Fang, Catriona Miller, Peter Hu, VS viewer study group. Originally published in JMIR Human Factors (https://humanfactors.jmir.org), 05.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Human Factors, is properly cited. The complete bibliographic information, a link to the original publication on https://humanfactors.jmir.org, as well as this copyright and license information must be included.
Characterizing and Comparing Adverse Drug Events Documented in 2 Spontaneous Reporting Systems in the Lower Mainland of British Columbia, Canada: Retrospective Observational Study

Erica Y Lau1,2, PhD; Amber Cragg2, MSc; Serena S Small2, MA; Katherine Butcher3, PharmB; Corinne M Hohl1, MSc, MD

1Department of Emergency Medicine, University of British Columbia, Vancouver, BC, Canada
2Centre for Clinical Epidemiology and Evaluation, Vancouver Coastal Health Research Centre, Vancouver, BC, Canada
3Pharmaceutical Science, Vancouver General Hospital, Vancouver Coastal Health, Vancouver, BC, Canada

*these authors contributed equally

Corresponding Author:
Erica Y Lau, PhD
Department of Emergency Medicine
University of British Columbia
7th Floor, 828 West 10th Avenue Research Pavilion
Vancouver, BC, V5Z 1M9
Canada
Phone: 1 6048754111 ext 68926
Email: erica.lau@ubc.ca

Abstract

Background: Robust adverse drug event (ADE) reporting systems are crucial to monitor and identify drug safety signals, but the quantity and type of ADEs captured may vary by system characteristics.

Objective: We compared ADEs reported in 2 different reporting systems in the same jurisdictions, the Patient Safety and Learning System–Adverse Drug Reaction (PSLS-ADR) and ActionADE, to understand report variation.

Methods: This retrospective observational study analyzed reports entered into PSLS-ADR and ActionADE systems between December 1, 2019, and December 31, 2022. We conducted a comprehensive analysis including all events from both reporting systems to examine coverage and usage and understand the types of events captured in both systems. We calculated descriptive statistics for reporting facility type, patient demographics, serious events, and most reported drugs. We conducted a subanalysis focused on adverse drug reactions to enable direct comparisons between systems in terms of the volume and events reported. We stratified results by reporting system.

Results: We performed the comprehensive analysis on 3248 ADE reports, of which 12.4% (375/3035) were reported in PSLS-ADR and 87.6% (2660/3035) were reported in ActionADE. Distribution of all events and serious events varied slightly between the 2 systems. Iohexol, gadobutrol, and empagliflozin were the most common culprit drugs (173/375, 46.2%) in PSLS-ADR, while hydrochlorothiazide, apixaban, and ramipril (308/2660, 11.6%) were common in ActionADE. We included 2728 reports in the subanalysis of adverse drug reactions, of which 12.9% (353/2728) were reported in PSLS-ADR and 86.4% (2357/2728) were reported in ActionADE. ActionADE captured 4- to 6-fold more comparable events than PSLS-ADR over this study’s period.

Conclusions: User-friendly and robust reporting systems are vital for pharmacovigilance and patient safety. This study highlights substantial differences in ADE data that were generated by different reporting systems. Understanding system factors that lead to varying reporting patterns can enhance ADE monitoring and should be taken into account when evaluating drug safety signals.

(JMIR Hum Factors 2024;11:e52495) doi:10.2196/52495

KEYWORDS

adverse drug event reporting systems; side effect; side effects; drug; drugs; pharmacy; pharmacology; pharmacotherapy; pharmaceutic; pharmaceutics; pharmaceuticals; pharmaceutical; medication; medications; patient safety; health information
technology; pharmacovigilance; adverse; safety; HIT; information system; information systems; reporting; descriptive statistics; monitoring

**Introduction**

Over 2 million Canadians visit an emergency department every year because of an adverse drug event (ADE), an unintended and harmful event related to medication use [1,2]. ADEs incur over 700,000 hospital admissions, and cost over CAD $1 billion (USD $7.48 million) in annual health care expenditures across Canada [2,3]. The importance of addressing this issue cannot be overstated: the World Health Organization (WHO) has identified the prevention of ADEs as an urgent global public health priority [4].

In response to this pressing concern, Canada implemented regulations outlined in the Protecting Canadians from Unsafe Drugs Act (Vanessa’s Law) which came into full effect on December 16, 2019. This federal legislation mandates prompt reporting of serious adverse drug reactions (ADRs; a subtype of ADEs) and medical device incidents from hospitals to Health Canada within 30 days of documentation [5]. These regulations serve as a safeguard to protect patients and improve drug surveillance.

Postmarketing pharmacovigilance is crucial in the detection, assessment, and prevention of ADEs under real-world conditions [6,7]. Among the various methods used, spontaneous reporting stands out as one of the most widely adopted approaches in pharmacovigilance [8]. When patients or health professionals spontaneously report ADEs, drug safety monitoring agencies evaluate and integrate these reports into databases, enabling ongoing identification of safety signals [7,8]. This method of surveillance captures data from a broad population and allows us to detect drug safety signals that may not have been identified in the randomized trials used for drug licensing and monitor rare ADEs to medications [9].

It is important to recognize, however, that there is considerable variation in ADE reporting systems worldwide in terms of their design, data fields, terminologies [10], and implementations, which may impact the volume and type of ADEs reported [11]. Variation in design also leads to a lack of standardization of reports, which can in turn prohibit interoperability or effective exchange of ADE reports between systems and may prevent comparisons of ADE events, rates, and risk factors across systems [10].

Despite this variation, the diversity of systems may also be a strength. Each system has the potential to complement others, enhancing the overall quantity and quality of ADE data, if variation in design leads to variation in reporting behaviors or the types of reports that can be entered [12]. To leverage this untapped potential, we need to better understand and compare the events collected through diverse reporting systems [13]. Understanding similarities and differences between systems will enable researchers and drug safety monitoring agencies to more effectively use existing data for accurate signal detection, especially for new or rare ADEs, and prioritize the investigation of drug safety signals. This knowledge will also aid stakeholders in optimizing the design and implementation of new reporting systems to enhance ADE data collection and drug safety surveillance and better align systems with their intended purpose [10].

Health Canada, the regulatory authority for postmarketing pharmacovigilance in Canada, oversees the Canada Vigilance Program, collecting reports of suspected ADEs since 1965. Health professionals and consumers can voluntarily submit reports through various channels, including a web-based platform, phone, fax, or mail. Hospitals are required to submit written reports within 30 days, and Health Canada allows them flexibility by permitting the use of existing systems and processes to meet reporting requirements. With Health Canada’s approval, hospitals may use a third party, such as a regional health authority or other reporting programs, to submit reports [14].

The province of British Columbia (BC) currently uses 2 approved spontaneous reporting systems that enable hospitals to comply with Vanessa’s Law mandates: the BC Patient Safety and Learning System–Adverse Drug Reaction (PSLS-ADR) reporting form and ActionADE. Briefly, PSLS-ADR was developed and implemented as the first province-wide, web-based platform and supports hospitals in meeting the mandatory reporting requirements [14]. ActionADE, implemented later in the timeline, is a research-driven, web-based app that aims to prevent unintentional redispensation of harmful medications by facilitating the sharing of ADE information between providers across health care settings. ADE reporting occurs as a byproduct of enabling safer care provision (Multimedia Appendix 1) [15].

These 2 systems enable a comparison of the quality and quantity of ADE data generated using 2 different designs. Our objective is to describe and compare the ADEs that health care providers documented using PSLS-ADR and ActionADE during the first 3 years following the implementation of Vanessa’s Law.

**Methods**

**ADE Reporting Systems**

**About PSLS-ADR**

BC Patient Safety and Learning System (PSLS) is an initiative of the BC Patient Safety Task Force, developed in collaboration with all 6 provincial health authorities and the Health Care Protection Program, which is part of the Risk Management Branch of the Ministry of Finance that insures BC hospitals [16]. BC PSLS is a web-based safety event reporting and management information system designed to support the identification, investigation, and analysis of safety and risk-related events, including safety hazards, near misses, and adverse events [17]. The system underwent a pilot phase in 2007 and was subsequently implemented province-wide in 2008. BC PSLS has been instrumental in promoting patient safety within the health care system in BC [16].

https://humanfactors.jmir.org/2024/11/e52495
In response to the introduction of Vanessa’s Law and in collaboration with Health Canada, BC PSLS launched PSLS-ADR as a new add-on to the existing system in 2014 and released an updated version in 2019 [18,19]. PSLS-ADR is accessible to health care facilities in all health authorities across BC, including acute care hospitals, long-term care facilities, and outpatient clinics. Authorized health care professionals with access to the secure health authority network, including employees, medical staff, paramedics, contractors, students, and volunteers, can submit reports to PSLS-ADR [15]. Once a report is submitted, the system notifies the medication safety officer in the respective health authority to review and respond to the event [20]. The health authorities send eligible reports to Health Canada for Vanessa’s Law reporting requirements. Reports are not made available to care providers and not integrated into the electronic medical record. They are only generated for the purposes of pharmacovigilance (Multimedia Appendix 2).

The PSLS-ADR data fields are based largely on the Canada Vigilance Adverse Reaction Reporting Form, with additional questions enabling medication safety officers, pharmacy representatives, and others to follow-up with reporters or patients, if necessary [20]. The PSLS-ADR reporting form contains 26 required data fields that collect information about the patient, the adverse reaction (eg, seriousness), the suspected health products (types, name, route used, therapy dates, and treatments), and the reporters.

About ActionADE

Previous studies found that 32.5% of ADE cases observed in emergency departments are repeat events [21], often occurring due to the unintentional re-preservation or redistribution of the same or a same-class medication as one that previously caused harm [22]. This recurrence is attributed to the lack of effective means to communicate and integrate ADE information into clinical workflows. ActionADE, a research-driven initiative, was developed to address this communication gap [23,24].

In collaboration with the Ministry of Health, Vancouver Coastal Health, a technology partner, and health professional organizations and clinicians, our research team developed and piloted ActionADE between 2016 and 2019 using participatory design principles and data standards that were evaluated and subsequently pilot tested to optimize the system’s usability [10,11,15,24-28]. In 2020, we began the implementation of ActionADE in 1 hospital (Vancouver General Hospital) and then expanded its use to 6 hospitals operated by Vancouver Coastal Health and Providence Health Care as part of a research initiative. Although providers were encouraged to use ActionADE, they maintain complete autonomy in choosing between the PSLS-ADR and ActionADE systems to meet their needs.

ActionADE is a web-based app that allows providers to document and communicate ADE information, bidirectionally through its integration (or linkage) with BC’s central drug database (PharmaNet). ActionADE was accessible to a subset of care providers with an eligible prescriber identification number issued by their respective regulatory college (ie, physicians, pharmacists, and nurse practitioners) [29]. Eligible clinicians submit reports to ActionADE from a designated health authority network, and the data are shared with clinicians within the patient’s circle of care via PharmaNet and used to create safety alerts when community pharmacists attempt to redistribute culprit or same-class medications. ActionADE complements the PSLS-ADR system by automating ADE reporting to Health Canada (Multimedia Appendix 3).

The ActionADE data fields were developed based on a systematic review of ADE reporting systems worldwide and participatory action research with clinician end users and are compatible with Health Canada’s Canada Vigilance Adverse Reaction Reporting Form [10,11,15,27,30]. As ActionADE is integrated with PharmaNet, several fields auto-populate based on the patient’s personal health number, including patient’s personal and demographic information (ie, name, date of birth, and sex), reporter’s information (ie, name, role, and site), and patient’s 14-month medication dispensation history. To create a new report, the system auto-populates the patient’s information and medication dispensation history, as well as the reporter’s information. ActionADE contains 5 required data fields that collect information about the suspect drugs, which is automatically generated based on the medication dispensation history or added manually, the ADE type, and details of the event (eg, symptoms or diagnosis, outcome, and certainty; Multimedia Appendix 4).

Study Design

In this retrospective observational study, we analyzed reports documented in PSLS-ADR and ActionADE entered by providers at health care facilities operated by the Vancouver Coastal Health Authority (excluding Providence Health Care, as PSLS-ADR data were unavailable from those facilities) in BC, Canada, between December 1, 2019, to December 31, 2022.

For PSLS-ADR, we included reports documented by authorized health care professionals (eg, employees, medical staff, paramedics, contractors, students, and volunteers) from >120 health care facilities across the province, including hospital, urgent and primary care, long-term care facilities, and community health centers’ clinics. For ActionADE, we included reports documented by eligible clinicians from 4 hospitals where ActionADE is implemented: Lions Gate Hospital, Richmond Hospital, UBC (University of British Columbia) Hospital, and Vancouver General Hospital (Multimedia Appendix 5) [31].

We divided this study’s period into 4 phases: baseline period, when all hospitals across BC only used PSLS-ADR (December 2019 to February 2020); year 1 (March 2020 to November 2020), when 1 hospital (Vancouver General Hospital) had the option to use ActionADE for piloting purposes while all other sites in BC exclusively used PSLS-ADR; and year 2 (December 2020 to November 2021) and year 3 (December 2021 to December 2022), when the 4 hospitals had the option to use PSLS-ADR or ActionADE and all other sites in BC exclusively used PSLS-ADR (Multimedia Appendix 6).

Data Sources

For this study, we requested ADE reports from PSLS-ADR from the BC PSLS central office and retrieved reports documented in ActionADE during the same period from the ActionADE database. We obtained information about hospital
characteristics through the Information Access and Privacy Services at Provincial Health Services Authority, including number of beds, population served, and the number of emergency department visits per year.

### Data Extraction

To allow for direct comparisons between the 2 systems, we combined similar variables wherever possible. A clinical pharmacist classified all free-text drug entries from the PSLS-ADR reports into the equivalent generic drug name that would be present if the same report were entered into ActionADE based on the provincial formulary. We translated continuous age from ActionADE into the age categories in PSLS-ADR. We combined information across platforms to produce combined variables for report date, patient demographics (age group and sex), types of ADE (ADRs and nonadverse drug reactions), ADE outcomes (death, emergency visit, hospitalized or hospital extended, life-threatening, worsened preexisting condition, permanent disability and fetal defect, other, and unknown) [27], and reporter information (role and facility).

### Statistical Analysis

#### Comprehensive Analysis

First, we performed a comprehensive statistical analysis to provide a global view of coverage, usage, and the types of information captured by both reporting systems. We included all events from both reporting systems, excluding reports related to user errors (eg, duplicate reports), refuted allergies, and reports with incomplete data. We calculated descriptive statistics (eg, means and SDs or frequency and percentages) for the following variables: total number and types of the reporting facilities (hospital vs nonhospital), patient’s age group, patient’s sex, roles of reporters, proportion of serious events, and the 10 most reported culprit drugs for all events and serious events. We defined serious events based on the Health Canada’s definition. This definition includes ADEs that require in-patient hospitalization or prolongation of existing hospitalization, cause congenital malformation, result in persistent or significant disability or incapacity, are life-threatening, or result in death [14].

#### ADR Analysis

To allow for direct comparisons between the 2 systems, we then conducted a subsample analysis that only included ADR reports (a subtype of ADE) that met Health Canada’s definition and that could have been reported in both systems. According to Health Canada, ADRs encompass harmful and unintended responses to a health product, including any undesirable patient effects suspected to be associated with health product use. This definition includes unintended effects, health product abuse, overdoses, interactions (including drug-drug and drug-food interactions), and unusual lack of therapeutic efficacy, all of which are considered reportable adverse reactions [14,32,33]. We included eligible ADR reports from both reporting systems from sites where both systems were available, excluding reports related to user errors (eg, duplicate reports), refuted allergies, and reports with incomplete data. We calculated descriptive statistics for the following variables: patient’s age group, patient’s sex, proportion of serious events, the 10 most reported culprit drugs, and mean monthly counts of all events and serious events during each phase of this study period, stratified by reporting system. We conducted all analyses using SAS statistical software (version 9.4; SAS Institute).

### Ethical Considerations

The UBC (University of British Columbia) clinical research ethics board approved of this research (H18-01332 and H22-00312) and provided a waiver for obtaining informed consent as this study meets the Tri-Council Policy Statement minimal risk criteria.

### Results

#### Comprehensive Analysis

We extracted 3248 reports from both reporting systems. After removing 213 reports related to refuted allergies, erroneous reports, and reports with incomplete data, the analytic cohort for the comprehensive analysis comprised 3035 unique ADEs reported in either system (Figure 1). Of these, 12.4% (375/3035) were entered in PSLS-ADR and 87.6% (2660/3035) were reported in ActionADE. Approximately 50% of the events occurred in male patients in both PSLS-ADR (178/375) and ActionADE (1285/3035). The highest proportion of events were from patients aged 45-64 years (32.8%, 123/375) in PSLS-ADR and aged 75-84 years (25.3%, 674/2660) in ActionADE. In total, 12 facilities (5 hospitals and 7 nonhospital facilities) entered reports in PSLS-ADR. The primary reporters in PSLS-ADR were medical imaging staff or technicians (170/375, 45.3%) and pharmacists (174/375, 46.4%). Of the 4 hospitals that entered reports in ActionADE, pharmacists were the reporter for 92.1% (2451/3035) of the events. The proportion of serious events was 36% (135/377) in PSLS-ADR and 28.2% (749/3035) in ActionADE (Table 1).

In PSLS, the most common culprit drugs were iohexol, gadobutrol, and empagliflozin, accounting for 46.2% (173/375) of all events. Empagliflozin, ibuprofen, and iohexol represented 11.8% (16/135) of serious events (Tables 2 and 3). Iohexol and gadobutrol are both contrast agents used for diagnostic imaging, whereas empagliflozin is an oral medication primarily prescribed for managing type 2 diabetes mellitus and ibuprofen is an oral, over-the-counter nonsteroidal anti-inflammatory drug used to relieve pain, reduce inflammation, and alleviate fever.

In ActionADE, the most common culprit drugs were hydrochlorothiazide, ramipril, and apixaban, which accounted for 10.5% (356/3391) of all events; hydrochlorothiazide, empagliflozin, and apixaban represented 11.1% (105/951) of serious events (Tables 2 and 3). Hydrochlorothiazide and ramipril are commonly prescribed for hypertension. Apixaban, an oral anticoagulant, is primarily used for stroke prevention in patients with atrial fibrillation and for treatment and prevention of venous thromboembolism.
**Figure 1.** Flow diagram. ADE: adverse drug event; ADR: adverse drug reaction; PSLS-ADR: Patient Safety and Learning System–Adverse Drug Reaction.

**Table 1.** Descriptive statistics of all events included in the comprehensive analysis by reporting system\(^a\).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>PSLS-ADR(^b) (n=375), n (%)</th>
<th>ActionADE(^c) (n=2660), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type of reporting facilitates</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospitals</td>
<td>5 (41.6)</td>
<td>4 (100)</td>
</tr>
<tr>
<td>Nonhospitals</td>
<td>7 (58.4)</td>
<td>0 (0)</td>
</tr>
<tr>
<td><strong>Patient age group (y)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;1-19</td>
<td>11 (2.9)</td>
<td>22 (0.8)</td>
</tr>
<tr>
<td>20-44</td>
<td>83 (22.1)</td>
<td>299 (11.2)</td>
</tr>
<tr>
<td>45-64</td>
<td>123 (32.8)</td>
<td>523 (19.7)</td>
</tr>
<tr>
<td>65-74</td>
<td>67 (17.9)</td>
<td>544 (20.5)</td>
</tr>
<tr>
<td>75-84</td>
<td>52 (13.9)</td>
<td>674 (25.3)</td>
</tr>
<tr>
<td>&gt;84</td>
<td>39 (10.4)</td>
<td>598 (22.5)</td>
</tr>
<tr>
<td><strong>Patient sex</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>178 (47.5)</td>
<td>1285 (48.3)</td>
</tr>
<tr>
<td><strong>Role of reporter</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physicians</td>
<td>Suppressed(^d)</td>
<td>204 (7.7)</td>
</tr>
<tr>
<td>Nurses</td>
<td>27 (7.2)</td>
<td>__(^e)</td>
</tr>
<tr>
<td>Medical imaging staff or technologists</td>
<td>170 (45.3)</td>
<td>__(^e)</td>
</tr>
<tr>
<td>Nurse practitioners</td>
<td>Suppressed</td>
<td>5 (0.2)</td>
</tr>
<tr>
<td>Pharmacists</td>
<td>174 (46.4)</td>
<td>2451 (92.1)</td>
</tr>
<tr>
<td>Others</td>
<td>Suppressed</td>
<td>__(^e)</td>
</tr>
<tr>
<td><strong>Proportion of serious events(^f)</strong></td>
<td>135 (36.0)</td>
<td>749 (28.2)</td>
</tr>
</tbody>
</table>

\(^a\)The comprehensive analysis included all events from both reporting systems excluding reports related to errors, refuted allergy, and incomplete data on study variables.

\(^b\)PSLS-ADR: Patient Safety and Learning System–Adverse Drug Reaction.

\(^c\)ADE: adverse drug event.

\(^d\)Cell sizes <5 are suppressed.

\(^e\)These personnel are not eligible to report in ActionADE.

\(^f\)Serious events are those with an outcome of fetal defect, permanent disability, hospitalization, extended hospitalization, life threatening, or death.
Table 2. Most frequently reported culprit drugs for all events in the comprehensive analysis by reporting systems.

<table>
<thead>
<tr>
<th>System and drug</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PSLS-ADR</strong> <em>(n=375)</em></td>
<td></td>
</tr>
<tr>
<td>Iohexol</td>
<td>154 (41.1)</td>
</tr>
<tr>
<td>Gadobutrol</td>
<td>12 (3.2)</td>
</tr>
<tr>
<td>Empagliflozin</td>
<td>7 (1.9)</td>
</tr>
<tr>
<td>Rivaroxaban</td>
<td>7 (1.9)</td>
</tr>
<tr>
<td>Furosemide</td>
<td>6 (1.6)</td>
</tr>
<tr>
<td>Nivolumab</td>
<td>6 (1.6)</td>
</tr>
<tr>
<td>Ramipril</td>
<td>6 (1.6)</td>
</tr>
<tr>
<td>Unknown generic drug</td>
<td>6 (1.6)</td>
</tr>
<tr>
<td>Acetylsalicylic acid</td>
<td>5 (1.3)</td>
</tr>
<tr>
<td>Ibuprofen</td>
<td>5 (1.3)</td>
</tr>
<tr>
<td><strong>ActionADE</strong> <em>(n=2660)</em></td>
<td></td>
</tr>
<tr>
<td>Hydrochlorothiazide</td>
<td>113 (4.2)</td>
</tr>
<tr>
<td>Apixaban</td>
<td>103 (3.9)</td>
</tr>
<tr>
<td>Ramipril</td>
<td>92 (3.5)</td>
</tr>
<tr>
<td>Acetylsalicylic acid</td>
<td>88 (3.3)</td>
</tr>
<tr>
<td>Warfarin</td>
<td>88 (3.3)</td>
</tr>
<tr>
<td>Rivaroxaban</td>
<td>79 (3)</td>
</tr>
<tr>
<td>Furosemide</td>
<td>77 (2.9)</td>
</tr>
<tr>
<td>Empagliflozin</td>
<td>63 (2.4)</td>
</tr>
<tr>
<td>Metformin HCL</td>
<td>52 (2)</td>
</tr>
<tr>
<td>Spironolactone</td>
<td>50 (1.9)</td>
</tr>
</tbody>
</table>

*PSLS-ADR: Patient Safety and Learning System–Adverse Drug Reaction.  
ADE: adverse drug event.  
HCL: hydrochloride.*
Table 3. Most frequently reported culprit drugs for serious events in the comprehensive analysis by reporting systems.

<table>
<thead>
<tr>
<th>System and drug</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PSLS-ADR</strong>&lt;sup&gt;b&lt;/sup&gt; (n=135)</td>
<td></td>
</tr>
<tr>
<td>Empagliflozin</td>
<td>6 (4.4)</td>
</tr>
<tr>
<td>Ibuprofen</td>
<td>5 (3.7)</td>
</tr>
<tr>
<td>Iohexol</td>
<td>5 (3.7)</td>
</tr>
<tr>
<td>Nivolumab</td>
<td>5 (3.7)</td>
</tr>
<tr>
<td>Acetylsalicylic acid</td>
<td>Suppressed&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Glyburide</td>
<td>Suppressed</td>
</tr>
<tr>
<td>Rivaroxaban</td>
<td>Suppressed</td>
</tr>
<tr>
<td>Allopurinol</td>
<td>Suppressed</td>
</tr>
<tr>
<td>Amlodipine besylate</td>
<td>Suppressed</td>
</tr>
<tr>
<td>Apixaban</td>
<td>Suppressed</td>
</tr>
<tr>
<td><strong>ActionADE</strong>&lt;sup&gt;d&lt;/sup&gt; (n=749)</td>
<td></td>
</tr>
<tr>
<td>Hydrochlorothiazide</td>
<td>43 (5.7)</td>
</tr>
<tr>
<td>Empagliflozin</td>
<td>26 (3.5)</td>
</tr>
<tr>
<td>Apixaban</td>
<td>24 (3.2)</td>
</tr>
<tr>
<td>Furosemide</td>
<td>20 (2.7)</td>
</tr>
<tr>
<td>Acetylsalicylic acid</td>
<td>18 (2.4)</td>
</tr>
<tr>
<td>Rivaroxaban</td>
<td>18 (2.4)</td>
</tr>
<tr>
<td>Candesartan cilexetil</td>
<td>16 (2.1)</td>
</tr>
<tr>
<td>Ramipril</td>
<td>16 (2.1)</td>
</tr>
<tr>
<td>Chlorthalidone</td>
<td>16 (2.1)</td>
</tr>
<tr>
<td>Spironolactone</td>
<td>15 (2)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Serious events are those with an outcome of fetal defect, permanent disability, hospitalization, extended hospitalization, life threatening, or death.

<sup>b</sup>PSLS-ADR: Patient Safety and Learning System–Adverse Drug Reaction.

<sup>c</sup>Cell sizes <5 are suppressed.

<sup>d</sup>ADE: adverse drug event.

ADR Analysis

We included a total of 2728 reports that met Health Canada’s definition of an ADR from facilities that had the option of using either reporting system during this study’s period (Figure 1) [32,33]. Of the included reports, 12.9% (353/2728) were entered in PSLS-ADR, while the majority (2357/2728, 86.4%) were reported in ActionADE.

The distribution of ADR reports by patient sex, age, primary reporters and proportion of serious events for both systems were similar to the comprehensive analysis (Table 4). However, each reporting system revealed distinct patterns of reporting. In PSLS-ADR, iohexol, gadobutrol, and empagliflozin accounted for 44.8% (168/353) of all events, while empagliflozin, ibuprofen, and nivolumab represented 12.1% (16/133) of serious events. In ActionADE, hydrochlorothiazide, ramipril, and apixaban accounted for 12% (284/2357) of all events. Furthermore, hydrochlorothiazide, empagliflozin, and apixaban represented 13.4% (88/671) of serious events (Tables 5 and 6).

A direct comparison in events reportable through both the PSLS-ADR and ActionADE systems revealed an increase in event reporting, including serious events, following the implementation of ActionADE (Figures 2 and 3). Baseline measurements indicate that the mean monthly counts of all events and serious events across sites were 2.9 (95% CI 2.2 to 3.6) and 1.7 (95% CI 0.8 to 2.5), respectively. In period 3, the mean monthly counts of all events and serious events across sites escalated to 27.2 (95% CI 20.4 to 34.0) and 7.0 (95% CI 4.9 to 9.2), respectively, reflecting a 9- and 4-fold increase over time. Furthermore, the mean monthly counts of all events and serious events during this study’s period within the ActionADE system were 6- and 4-fold greater than that of PSLS-ADR.
### Table 4. Descriptive statistics of events meeting Health Canada’s ADR\(^a\) definition\(^b\) across common reporting sites\(^c\) by reporting system.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>PSLS-ADR(^d) (n=353), n (%)</th>
<th>ActionADE(^e) (n=2357), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient age group (y)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;1-19</td>
<td>11 (3.1)</td>
<td>18 (0.8)</td>
</tr>
<tr>
<td>20-44</td>
<td>77 (21.8)</td>
<td>239 (10.1)</td>
</tr>
<tr>
<td>45-64</td>
<td>114 (32.3)</td>
<td>450 (19.1)</td>
</tr>
<tr>
<td>65-74</td>
<td>64 (18.1)</td>
<td>494 (21)</td>
</tr>
<tr>
<td>75-84</td>
<td>49 (13.9)</td>
<td>606 (25.7)</td>
</tr>
<tr>
<td>&gt;84</td>
<td>38 (10.8)</td>
<td>550 (23.3)</td>
</tr>
<tr>
<td><strong>Patient sex</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>173 (49)</td>
<td>1114 (47.3)</td>
</tr>
<tr>
<td><strong>Role of reporter</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physicians</td>
<td>Suppressed(^f)</td>
<td>161 (6.8)</td>
</tr>
<tr>
<td>Nurses</td>
<td>22 (6.2)</td>
<td>_(^g)</td>
</tr>
<tr>
<td>Medical imaging staff or technologists(^h)</td>
<td>155 (43.9)</td>
<td>_</td>
</tr>
<tr>
<td>Nurse practitioners</td>
<td>Suppressed</td>
<td>5 (0.2)</td>
</tr>
<tr>
<td>Pharmacists</td>
<td>173 (49)</td>
<td>2190 (92.9)</td>
</tr>
<tr>
<td>Others(^b)</td>
<td>Suppressed</td>
<td>_</td>
</tr>
<tr>
<td><strong>Proportion of serious events(^i)</strong></td>
<td>133 (37.7)</td>
<td>671 (28.5)</td>
</tr>
</tbody>
</table>

\(^a\)ADR: adverse drug reactions.  
\(^b\)According to Health Canada adverse drug reaction includes unintended effects, health product abuse, overdoses, interactions (including drug-drug and drug-food interactions), and unusual lack of therapeutic efficacy.  
\(^c\)Common reporting sites included Vancouver General, University of British Columbia, Lions Gate, and Richmond Hospitals.  
\(^d\)PSLS-ADR: Patient Safety and Learning System–Adverse Drug Reaction.  
\(^e\)ADE: adverse drug event.  
\(^f\)Cell sizes <5 are suppressed.  
\(^g\)Not available.  
\(^h\)These personnel are not eligible to report in ActionADE.  
\(^i\)Serious events are those with an outcome of fetal defect, permanent disability, hospitalization, extended hospitalization, life threatening, or death.
Table 5. Most frequently reported culprit drugs for all events meeting Health Canada’s ADR\(^a\) definitions\(^b\) across common reporting sites\(^c\) by reporting system and severity.

<table>
<thead>
<tr>
<th>System and drug</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PSLS-ADR(^d) (n=353)</strong></td>
<td></td>
</tr>
<tr>
<td>Iohexol</td>
<td>139 (39.4)</td>
</tr>
<tr>
<td>Gadobutrol</td>
<td>12 (3.4)</td>
</tr>
<tr>
<td>Empagliflozin</td>
<td>7 (2)</td>
</tr>
<tr>
<td>Rivaroxaban</td>
<td>7 (2)</td>
</tr>
<tr>
<td>Furosemide</td>
<td>6 (1.7)</td>
</tr>
<tr>
<td>Nivolumab</td>
<td>6 (1.7)</td>
</tr>
<tr>
<td>Ramipril</td>
<td>6 (1.7)</td>
</tr>
<tr>
<td>Acetylsalicylic acid</td>
<td>5 (1.4)</td>
</tr>
<tr>
<td>Ibuprofen</td>
<td>5 (1.4)</td>
</tr>
<tr>
<td>Indapamide</td>
<td>5 (1.4)</td>
</tr>
<tr>
<td><strong>ActionADE(^e) (n=2357)</strong></td>
<td></td>
</tr>
<tr>
<td>Hydrochlorothiazide</td>
<td>109 (4.6)</td>
</tr>
<tr>
<td>Ramipril</td>
<td>88 (3.7)</td>
</tr>
<tr>
<td>Apixaban</td>
<td>87 (3.7)</td>
</tr>
<tr>
<td>Acetylsalicylic acid</td>
<td>75 (3.2)</td>
</tr>
<tr>
<td>Warfarin</td>
<td>74 (3.1)</td>
</tr>
<tr>
<td>Rivaroxaban</td>
<td>71 (3)</td>
</tr>
<tr>
<td>Empagliflozin</td>
<td>60 (2.5)</td>
</tr>
<tr>
<td>Furosemide</td>
<td>56 (2.4)</td>
</tr>
<tr>
<td>Metformin HCL(^f)</td>
<td>44 (1.9)</td>
</tr>
<tr>
<td>Ibuprofen</td>
<td>43 (1.8)</td>
</tr>
</tbody>
</table>

\(^a\)ADR: adverse drug reaction.  
\(^b\)According to Health Canada adverse drug reaction includes unintended effects, health product abuse, overdoses, interactions (including drug-drug and drug-food interactions), and unusual lack of therapeutic efficacy.  
\(^c\)Common reporting sites included Vancouver General, University of British Columbia, Lions Gate, and Richmond Hospitals.  
\(^d\)PSLS-ADR: Patient Safety and Learning System–Adverse Drug Reaction.  
\(^e\)ADE: adverse drug event.  
\(^f\)HCL: hydrochloride.
Table 6. Most frequently reported culprit drugs for serious events<sup>a</sup> meeting Health Canada’s ADR<sup>b</sup> definitions<sup>c</sup> across common reporting sites<sup>d</sup> by reporting system and severity.

<table>
<thead>
<tr>
<th>System and drug</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PSLS-ADR&lt;sup&gt;e&lt;/sup&gt; (n=133)</strong></td>
<td></td>
</tr>
<tr>
<td>Empagliflozin</td>
<td>6 (4.5)</td>
</tr>
<tr>
<td>Ibuprofen</td>
<td>5 (3.8)</td>
</tr>
<tr>
<td>Nivolumab</td>
<td>5 (3.8)</td>
</tr>
<tr>
<td>Acetylsalicylic acid</td>
<td>Suppressed&lt;sup&gt;f&lt;/sup&gt;</td>
</tr>
<tr>
<td>Glyburide</td>
<td>Suppressed</td>
</tr>
<tr>
<td>Rivaroxaban</td>
<td>Suppressed</td>
</tr>
<tr>
<td>Allopurinol</td>
<td>Suppressed</td>
</tr>
<tr>
<td>Amlodipine besylate</td>
<td>Suppressed</td>
</tr>
<tr>
<td>Apixaban</td>
<td>Suppressed</td>
</tr>
<tr>
<td>Clopidogrel bisulfate</td>
<td>Suppressed</td>
</tr>
<tr>
<td><strong>ActionADE&lt;sup&gt;g&lt;/sup&gt; (n=671)</strong></td>
<td></td>
</tr>
<tr>
<td>Hydrochlorothiazide</td>
<td>43 (6.7)</td>
</tr>
<tr>
<td>Empagliflozin</td>
<td>25 (3.7)</td>
</tr>
<tr>
<td>Apixaban</td>
<td>20 (3)</td>
</tr>
<tr>
<td>Acetylsalicylic acid</td>
<td>19 (2.8)</td>
</tr>
<tr>
<td>Chlorthalidone</td>
<td>16 (2.4)</td>
</tr>
<tr>
<td>Ramipril</td>
<td>14 (2.1)</td>
</tr>
<tr>
<td>Rivaroxaban</td>
<td>14 (2.1)</td>
</tr>
<tr>
<td>Ibuprofen</td>
<td>13 (1.9)</td>
</tr>
<tr>
<td>Warfarin</td>
<td>13 (1.9)</td>
</tr>
<tr>
<td>Candesartan cilexetil</td>
<td>12 (1.8)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Common reporting sites included Vancouver General, University of British Columbia, Lions Gate, and Richmond Hospitals.

<sup>b</sup>ADR: adverse drug reaction.

<sup>c</sup>According to Health Canada adverse drug reaction includes unintended effects, health product abuse, overdoses, interactions (including drug-drug and drug-food interactions), and unusual lack of therapeutic efficacy.

<sup>d</sup>Serious events are those with an outcome of fetal defect, permanent disability, hospitalization, extended hospitalization, life threatening, or death.

<sup>e</sup>PSLS-ADR: Patient Safety and Learning System–Adverse Drug Reaction.

<sup>f</sup>Cell sizes <5 are suppressed.

<sup>g</sup>ADE: adverse drug event.
Discussion

Principal Findings

Our study aimed to describe and compare ADEs reported using 2 distinct reporting systems that were developed and implemented in different ways. Both PSLS-ADR and ActionADE are currently in use in BC in the first 3 years following the implementation of Vanessa’s Law. We observed differences in reports between the 2 systems regarding their coverage, usage, and the type of ADE data captured.

PSLS-ADR had broader coverage, collecting data from various health care facilities including community health centers, vaccination clinics, and outpatient clinics. Its user base was more diverse including physicians, nurses, medical imaging staff or technologists, nurse practitioners, pharmacists, and other professionals. In contrast, ActionADE coverage was limited to ADEs identified in patients presenting to 4 participating hospitals, with clinical pharmacists as its primary user. The broader coverage of PSLS-ADR can be attributed to its established position as a provincial safety event reporting platform; its accessibility to a broader range of health professions; and a federal mandate for hospitals to stimulate reporting using health authority wide communication efforts including email blasts, information on health authority websites, and presentations to provider groups. Leveraging the insights gained from PSLS-ADR, our research team is actively collaborating with key stakeholders to broaden ActionADE’s
Ibuprofen was the second most commonly reported culprit drug related to serious events in PSLS-ADR, but it barely made the top 10 in ActionADE. This discrepancy may be due to the over-the-counter status of ibuprofen, which means patients can access the medication without a prescription and bypass communication about ADEs from ActionADE that is built into the prescription dispensation process.

While our study primarily focused on comparing these 2 systems, it is crucial to view these findings in the broader context of ADE reporting. Despite these disparities, both systems play vital roles in contributing to patient safety by capturing valuable information on ADEs. PSLS-ADR is an effective means of capturing radiopharmaceutical-related ADEs by using staff and technicians who are not trained in taking medication histories or ADE assessments, while ActionADE is more effective for pharmaceutical-related ADEs by clinical pharmacists that are reported and communicated on a patient-level to improve safety. These systems work in a complementary manner, catering to different areas of the health care system and capturing unique data and thus offering a more comprehensive picture of ADEs. For example, a common signal between the 2 systems might indicate a more serious issue for a specific drug irrespective of context (e.g., empagliflozin). These findings suggest the need for careful attention to the design and implementation of these systems to ensure they effectively serve their intended users and context of use and ensure data resulting from each system are interpreted correctly by end users. The absence of reporting of one type of event may reflect design, implementation, or user characteristics rather than the absence of these events.

**Limitations**

To our knowledge, this is the first study to directly compare 2 ADE reporting systems operating within the same jurisdiction. While the results of our study provided valuable insights into the differences between these systems, it is important to acknowledge several limitations that warrant consideration in interpreting the results. First, our study sample was confined to 2 reporting systems, which may not fully encapsulate the diversity of all systems employed across health care settings globally. As a result, the findings may not be generalizable to other reporting systems. Second, our data set was limited to facilities that used PSLS-ADR or ActionADE for reporting. This reduces the generalizability of our findings to the wider array of health care facilities in BC or nationally. It is plausible that unaccounted-for variations in data and reporting practices among facilities not deploying these 2 systems could exist. Third, our study may be susceptible to unmeasured and uncontrolled confounding variables. For example, the level of organizational emphasis on ADE reporting, differences in implementation, available resources, and providers’ perceptions could have affected the usage and coverage of the 2 systems under study. This variability might have further influenced the nature of ADE information reported. Fourth, the relatively small number of drugs resulting in ADEs prevented us from conducting a robust quantitative comparison of these events. Furthermore, the data we used were a snapshot in time and may not reflect changes in reporting systems or health care facilities that have occurred since then. Lastly, we consciously chose not
to draw comparisons with other studies examining the frequently reported culprit drugs from spontaneous reporting systems in other jurisdictions. This decision stems from the recognition that the diversity in ADE reports—both in terms of numbers and types—is intricately tied to factors such as system design, geography, population characteristics, drug exposures, and the medical system itself. To facilitate meaningful comparisons across studies, a more robust surveillance system is needed.

Conclusions
Understanding the differences between reporting systems can inform future systems design and improvement, including changes to user training and implementation, and inform the use of forthcoming data and procurement decisions for reporting systems. Further research could explore how to integrate the strengths of both systems, potentially leading to more comprehensive safety data to facilitate drug and patient safety and inform pharmacoepidemiologic studies. Continuous evaluation and improvement are essential considering the significant role these systems play to improve our health systems.

Acknowledgments
The ActionADE platform is funded by the Canadian Institutes of Health Research, with support from the Ministry of Health British Columbia, Vancouver Coastal Health, the Lower Mainland Consolidated Pharmacies, and the BC SUPPORT Unit. CMH is supported by the Michael Smith Foundation for Health Research Health Professional-Investigator award. EYL is supported by the Michael Smith Health Research BC Research Trainee Award (GR018656). The funder has no role in the design, conduct, analysis, interpretation of data, or the drafting of scientific papers.

Authors’ Contributions
All authors contributed to this study’s conception and design. EYL requested access and analyzed the data and wrote the first draft of this paper. AC, CMH, and SSS contributed to the refinement of the data analysis. All authors contributed to the interpretation of the findings and commented on previous paper versions. All authors read and approved the submitted paper and have agreed to be personally accountable for their contribution.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Characteristics of (Patient Safety and Learning System–Adverse Drug Reaction) PSLS-ADR and ActionADE.
[DOCX File, 23 KB - humanfactors_v11i1e52495_app1.docx ]

Multimedia Appendix 2
Screenshot of Patient Safety and Learning System–Adverse Drug Reaction (PSLS-ADR).
[DOCX File, 282 KB - humanfactors_v11i1e52495_app2.docx ]

Multimedia Appendix 3
Screenshot of ActionADE.
[DOCX File, 527 KB - humanfactors_v11i1e52495_app3.docx ]

Multimedia Appendix 4
Data fields included in Patient Safety and Learning System–Adverse Drug Reaction (PSLS-ADR) and ActionADE.
[DOCX File, 36 KB - humanfactors_v11i1e52495_app4.docx ]

Multimedia Appendix 5
Characteristics of sites that had options to use Patient Safety and Learning System–Adverse Drug Reaction (PSLS-ADR) or ActionADE systems.
[DOCX File, 30 KB - humanfactors_v11i1e52495_app5.docx ]

Multimedia Appendix 6
Descriptions of the 4-phase study period.
[DOCX File, 159 KB - humanfactors_v11i1e52495_app6.docx ]

References


18. Online ADR reporting is coming!. British Columbia Patient Safety & Learning System Central. 2014. URL: https://bcpslscentral.ca/online-adr-reporting-is-coming/#:~:text=Most%20of%20us%20will%20take%20the%20medication%20as%20intended [accessed 2023-12-22]


20. Online ADR reporting is coming!. British Columbia Patient Safety & Learning System. 2014. URL: https://bcpslscentral.ca/online-adr-reporting-is-coming/ [accessed 2023-05-23]


Abbreviations

ADE: adverse drug event
ADR: adverse drug reaction
BC: British Columbia
PSLS-ADR: Patient Safety and Learning System–Adverse Drug Reaction
PSLS: Patient Safety and Learning System
UBC: University of British Columbia
WHO: World Health Organization

Edited by K Adapa; submitted 05.09.23; peer-reviewed by R Lim, A Austin; comments to author 28.11.23; revised version received 06.12.23; accepted 07.12.23; published 18.01.24.

Please cite as:
Lau EY, Cragg A, Small SS, Butcher K, Hohl CM
Characterizing and Comparing Adverse Drug Events Documented in 2 Spontaneous Reporting Systems in the Lower Mainland of British Columbia, Canada: Retrospective Observational Study
JMIR Hum Factors 2024;11:e52495
URL: https://humanfactors.jmir.org/2024/1/e52495
doi:10.2196/52495
PMID:38236629
Usability and Evaluation of a Health Information System in the Emergency Department: Mixed Methods Study

Christina Østervang¹,², RN, MScN, PhD; Charlotte Myhre Jensen²,³, RN, MEd, PhD; Elisabeth Coyne²,⁴, RN, PhD; Karin B Dieperink²,⁵,⁶, RN, MCN, PhD; Annmarie Lassen¹,², DMSci, PhD

¹Department of Emergency Medicine, Odense University Hospital, Odense, Denmark
²Department of Clinical Research, University of Southern Denmark, Odense, Denmark
³Department of Orthopedic Surgery and Traumatology, Odense University Hospital, Odense, Denmark
⁴School of Nursing and Midwifery, Griffith University, Brisbane, Australia
⁵Department of Oncology, Odense University Hospital, Odense, Denmark
⁶Family Focused Healthcare Research Center (FACE), University of Southern Denmark, Odense, Denmark

Corresponding Author:
Christina Østervang, RN, MScN, PhD
Department of Emergency Medicine, Odense University Hospital
Kløvervænget 25
Odense, 5000
Denmark
Phone: 45 21757009
Email: christina.oestervang@rsyd.dk

Abstract

Background: A lack of information during an emergency visit leads to the experience of powerlessness for patients and their family members, who may also feel unprepared to cope with acute symptoms. The ever-changing nature and fast-paced workflow in the emergency department (ED) often affect how health care professionals can tailor information and communication to the needs of the patient.

Objective: This study aimed to evaluate the usability and experience of a newly developed information system. The system was developed together with patients and their family members to help provide the information needed in the ED.

Methods: We conducted a mixed methods study consisting of quantitative data obtained from the System Usability Scale questionnaire and qualitative interview data obtained from purposively selected participants included in the quantitative part of the study.

Results: A total of 106 patients and 14 family members (N=120) answered the questionnaire. A total of 10 patients and 3 family members participated in the interviews. Based on the System Usability Scale score, the information system was rated close to excellent, with a mean score of 83.6 (SD 12.8). Most of the participants found the information system easy to use and would like to use it again. The participants reported that the system helped them feel in control, and the information was useful. Simplifications were needed to improve the user experience for the older individuals.

Conclusions: This study demonstrates that the usability of the information system is rated close to excellent. It was perceived to be useful as it enabled understanding and predictability of the patient’s trajectory in the ED. Areas for improvement include making the system more usable by older individuals. The study provides an example of how a technological solution can be used to diminish the information gap in an ED context.

(JMIR Hum Factors 2024;11:e48445) doi:10.2196/48445

KEYWORDS
consumer; eHealth; elderly; emergency department; emergency; family members; healthcare professionals; information system; mixed methods research; patients; qualitative interview; questionnaire; technology; usability; usable
**Introduction**

**Background**

Clear communication and information are essential to improving care and patient outcomes in the emergency department (ED) [1-5]. A lack of information during ED visits causes patients and their family members to experience a sense of powerlessness and to feel unprepared to cope with acute symptoms [2,3,6]. Due to the hectic nature of the ED and the constant interruptions, communication from health care professionals is often inadequate or not tailored to patients and their families [4,7]. While this problem has been known for many years, it still persists to this date [1].

Health technologies are implemented in many parts of health care systems to promote quality care and treatment [8]. The design and purpose of health technologies range widely from organizational [9] to person-centered intentions [10]. In the ED, technologies may be used as quality dashboards [9] and more personal information systems on patients’ own devices to support the delivery of health information [11]. However, the successful use of technology in clinical practice is likely to be ineffective if user needs are not carefully addressed and incorporated before attempting a full-scale implementation [9,12]. Thoroughness in integrating and understanding user perspectives will have a direct impact on how well the technology is suited for clinical practice [13,14].

Based on the current findings, patients in the ED and their family members have unmet information needs [1-4]. Hence, guided by the principles of user-driven activities [15], a health information system was developed [16]. The health information system, which is called “Cetrea Clinical Logistic (CCL) for patients,” is available for patients in the hospital’s emergency room and displays real-time information, including (1) person-centered activities, (2) information videos, (3) a notepad, (4) waiting time, and (5) the nurse and physician responsible for care.

Usability is one of the factors affecting the acceptance of health information systems by users, and it is essential for the effective use of the system [17]. A usability evaluation can identify problems and weaknesses in the design and functionalities in the early development phase [18]. Usability tests allow developers to address and adjust concerns and, thus, avoid implementing technologies that will not be useful in the clinical context.

Therefore, a usability evaluation from an end-user perspective was completed to obtain a nuanced understanding of the sustainable use of the system, specifically from the perspective of patients and their family members.

**Objective**

The objective of this study is to gain knowledge about the usability and experiences of the newly developed information system, CCL for patients. This study reports on patients’ and family members’ evaluations of this system.

**Participatory Design and Technology**

This study is the final phase of a 3-phase participatory design study (Figure 1) [19]. Participatory design is a research methodology based on the epistemological position of genuine involvement and understanding of the needs of future end users. A new technology can be designed to improve a real-life problem [20]. The core principles in participatory design methodology have been the theoretical framework of the overall study. In the initial phase, the author group identified the essential needs of patients in the ED, their family members, and ED clinicians [2,3]. The results from phase 1 informed the second phase, in which an information system, CCL for patients, was developed in a cocreation process [16]. The third phase involved testing and evaluation of the system, which is reported in this study. Reporting the evaluation of participatory-designed health technology is a common part of the research methodology [21,22].

The author group has had no financial interest in the system owners of CCL for patients and has no interest in either marketing or promoting the system.

CCL for patients provides information directly to patients and their family members during their stay in the ED. The information provided relates to treatment and time factors and is adjusted toward the individual patient. CCL is an already existing and implemented system for task management for clinicians’ use only [23], whereas CCL for patients is a redesign and further development of the system for patients’ use.

The functionalities of the CCL for patients’ screen are presented in Figure 2.
Methods

Research Design

This is a mixed methods study inspired by a convergent parallel design [24]. This design was chosen to obtain nuanced insights into the usability of the system. Further, we adopted this approach to usability testing because quantitative data can identify usability issues and dissatisfaction with program design, while qualitative data can provide detailed information about the causes of the usability issues and point at potential methods for program optimization. As shown in Figure 3 [24], the study contained the following two parts, ending with a merged result: (1) a questionnaire and descriptive characteristics of the participants, and (2) semistructured interviews with patients and their family members.

Figure 1. Overview of the 3-phase study, highlighting the evaluation phase, which is reported in this study.

Figure 2. Cetrea Clinical Logistics (CCL) for patients and its functionalities, as displayed to patients and their family members, developed in the second phase of the overall study (the figure has been previously published by Østervang et al [16]). (1) Number of the ED room. (2) Name of the hospital department. (3) The name of the patient (no sensitive information is displayed). (4) The nurse who is responsible for care. (5) The physician who is responsible for treatment. (6) Process line with activities. Displaying nurse assessments, blood samples, electrocardiograms, physician assessments, X-rays, etc. (7) Clarification of the different colors in the process line. Gray: not started; blue: activity scheduled; and green: activity finished. (8) Clarification of special activity names. (9) Link to information videos (e.g., information on discharge). (10) Three diverse colors indicate the estimated waiting time: less than 4 hours, equal to 4 hours, or more than 4 hours, respectively. (11) The shared note pad for the patients to write questions to healthcare professionals or messages from family members.
Figure 3. Diagram for a study using convergent design (Creswell and Clark) [24].

Setting
The data were collected in Odense University Hospital’s ED between August 22 and September 29, 2022, on weekdays from 8 AM to 5 PM. The information system was displayed on a laptop personal computer (PC) sitting on the bedside table in the ED room. Four PCs were used during the test phase. They were installed in the specific ED room where the patients participating in the study were admitted.

Inclusion Criteria and Recruitment
All patients admitted to the medical area of the ED without a final plan for treatment and care were eligible for participation. Patients were excluded if they were severely ill or cognitively unable to use the technology. However, patients who were excluded due to a cognitive inability to use the screen but who were still able to give consent for their family members’ participation were enrolled if the family member was interested in participating. Patients were recruited by the first author (CØ) or one of 2 research assistants, all of whom have a Master of Nursing Science degree and research experience. Potential participants were identified and discussed with the responsible care nurse before they were approached to reduce the possibility of any concerns.

Quantitative Phase
A survey was conducted to elicit the opinions and experiences of patients and their family members using the information system.

The Questionnaire
The questionnaire, the System Usability Scale (SUS), contained questions regarding the usability of the system. Answers are rated on a 5-point Likert scale from “strongly disagree” to “strongly agree,” with 5 representing the highest score (strongly agree) [25]. The participants answered 10 questions from the SUS and 2 questions specific to this study (questions 11 and 12) [25]. These 2 extra questions were added to obtain general information about the participants’ experience with CCL for patients (question 11: “I think the system provided a great overview of my stay,” and question 12: “I think the information in the system made sense to me”). As SUS has been translated and validated in a Danish hospital context previously (Cronbach $\alpha=.87$) [26], it was considered suitable for this study.

Sample Size
A total SUS score between 70 and 90 indicates good to excellent usability of the tested system [27]. Based on previous research conducted in Scandinavia using SUS in health care with a reported mean score of 79.81 (SD 14.28), we would gain a 95% CI for a mean score between 77.2 and 82.4 if a total of 120 patients were included [28].

Data Collection
If a patient agreed to participate, the researcher cooperated with the local IT department at the hospital to ensure the patient’s access to the system. Initially, the researcher sent the IT department an SMS text message providing information on the PC number and the ED room number. The IT specialist matched the PC and room numbers. Then, the researcher double-checked that the correct information was displayed before handing it to the patient. All participants were given oral guidance on how to use CCL for patients.

The PC with individual information was placed on the bedside table until either the patient left the ED, the patient had used the system for a minimum of 2 hours, or the patient felt ready to perform the evaluation. All of this had to happen no later than 5 PM, when the IT department closed. When returning the PC, the participants were given an iPad to fill out the questionnaire. The data were stored on the logged server OPEN [29], which is part of Odense University Hospital and the University of Southern Denmark.

Qualitative Phase
Interviews were conducted with individual patients or with the patient together with a family member to get a deeper insight into their experiences using the information system.

Interviews
The qualitative part included a subset of the participants from the quantitative part. Before making CCL for patients available
to the participants, they were asked whether they were interested in participating in an interview.

All interviews were conducted by the first author (CØ). By taking a phenomenological-hermeneutical stance, CØ was allowed to recognize her perceptions as an experienced emergency nurse within hermeneutic interpretation [30]. To bridle her preconceived ideas, CØ wrote down her preunderstandings of why patients lack information in the ED. This reflection provided an initial focus for both the overall research question and the interview questions.

The interviews were conducted in the hospital room after the participants had completed the questionnaire. Notes and quotes were taken during the interview. A summary of the conversation was generated at the end of the interview in the form of member checking [31]. A semistructured interview guide inspired by Kvale was used [32]. An example of a question is: “What was your experience of using CCL for patients?” The interviews lasted up to 30 minutes. The interviews were conducted until no new themes arose [33].

Sample Size
To obtain maximal variation, a purposive sampling strategy was used [33]. The inclusion criteria were the same as for the quantitative part of the study, but they also ensured representation of differences in age and gender.

Analysis
Analysis of the Questionnaires
Only fully completed questionnaires were analyzed (N=120). There were no missing data, as the questionnaire was only considered complete if all the questions were answered. According to the SUS guidelines, we performed an individual analysis of each participant’s SUS score as well as the mean value for the entire population. We separated the 2 self-constructed questions from the original SUS questions in the calculation and interpretation process to ensure that they were accurate and reliable. The final score was between 0 and 100, where a higher score indicates better usability. Odd-numbered questions were positive in tone, and even-numbered questions were negative in tone, so the scale was converted into points ranging from 1 to 5 (1=strongly disagree to 5=strongly agree). The final score was calculated as follows: X = the sum of the points for all odd-numbered questions minus 5. And Y = 25 minus the sum of the points for all even-numbered questions. SUS score = (X + Y) × 2.5 [34]. A system needs a score above 70 to be considered acceptable; better systems will score from the high 70s to the high 80s, and excellent systems will score above 90 [27].

Analysis of the Interviews
The qualitative interviews were analyzed and reported based on Malterud’s [35] systematic text condensation. This process consisted of four steps: (1) transcriptions were read several times to get a total impression of the data and to find preliminary themes; (2) we identified and sorted meaning units based on the preliminary themes and arranged them into code groups; (3) the code groups were reviewed, and the content was reduced into condensates; and (4) the meaning and content of the condensates were synthesized and interpreted [35]. The analysis was completed by CØ using NVivo (version 12; QSR International). The trustworthiness and rigor of the qualitative part of the study were evaluated using Guba’s [36] definition of quality criteria. As part of steps 2, 3, and 4 in the analysis, the emerging themes and codes were discussed in the author group toward strengthening the credibility and reflexivity of our interpretation of the interviews. Using a systematic approach toward the analysis strategy of all interviews ensured confirmability in the data collection and analysis process.

The SQUIRE 2.0 checklist [37] was used to create transparency and ensure that no important information was missed in the reporting of the study.

Integration of Quantitative and Qualitative Results
To achieve an expanded understanding of the results, the qualitative and quantitative results were compared and integrated as the final step of the analysis using joint display tables [24]. In a joint display table, the 2 results are presented in a way that allows comparison, leading to confirmation, disconfirmation, or expansion of each other [24]. The results from the SUS (quantitative results) are presented on a Likert scale, showing the variation of the grades in the different questions. To elaborate on and verify the answers, supportive qualitative quotes were presented for each question. We divided the grades into low (1-3) and high (4-5) to separate the different perceptions of CCL for patients.

Ethical Considerations
All the participants received verbal and written information about the study in accordance with applicable ethical rules [38] and provided their oral and written consent. The study is registered with the Danish Data Protection Agency, Fortegnelsen (19/22672). Approval of the project was granted by the Regional Committee on Health Research Ethics for Southern Denmark (S-20192000–111).

Results
Quantitative Results
In total, 14 family members and 106 patients agreed to participate. A total of 27 patients declined to participate for three main reasons: (1) no interest, (2) no technical skills, and (3) a lack of mental ability due to the acute situation.
Table 1. Demographic descriptions of the participants.

<table>
<thead>
<tr>
<th>Demographic description</th>
<th>Patients (n=106)</th>
<th>Family members (n=14)</th>
<th>Total (N=120)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>55 (51.9)</td>
<td>8 (57.1)</td>
<td>63 (52.5)</td>
</tr>
<tr>
<td>Male</td>
<td>51 (48.1)</td>
<td>6 (42.9)</td>
<td>57 (47.5)</td>
</tr>
<tr>
<td><strong>Age (years), mean (SD)</strong></td>
<td>55.5 (SD 18.7)</td>
<td>66.5 (SD 11.6)</td>
<td>57 (SD 18.3)</td>
</tr>
<tr>
<td><strong>Civil status, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No partner</td>
<td>39 (36.8)</td>
<td>2 (14.3)</td>
<td>41 (34.2)</td>
</tr>
<tr>
<td>In a relationship</td>
<td>67 (63.2)</td>
<td>12 (85.7)</td>
<td>79 (65.8)</td>
</tr>
<tr>
<td><strong>Children, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Having children</td>
<td>81 (76.4)</td>
<td>14 (100.0)</td>
<td>95 (79.2)</td>
</tr>
<tr>
<td>Having children living at home</td>
<td>32 (39.5)</td>
<td>6 (42.9)</td>
<td>38 (30.0)</td>
</tr>
<tr>
<td><strong>Technology, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Having a smartphone</td>
<td>96 (90.6)</td>
<td>14 (100.0)</td>
<td>110 (91.7)</td>
</tr>
<tr>
<td>Using technology on daily basis</td>
<td>102 (96.2)</td>
<td>13 (92.9)</td>
<td>115 (95.8)</td>
</tr>
<tr>
<td><strong>Education, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>21 (19.8)</td>
<td>1 (7.1)</td>
<td>22 (18.3)</td>
</tr>
<tr>
<td>Medium</td>
<td>71 (67.0)</td>
<td>9 (64.3)</td>
<td>80 (66.7)</td>
</tr>
<tr>
<td>High</td>
<td>14 (13.2)</td>
<td>4 (28.6)</td>
<td>18 (15.0)</td>
</tr>
</tbody>
</table>

The respondents were equally represented by gender, with a mean age of 57 years. The mean age of family members was higher than that of the included patients. Most participants had medium education levels, but low and high educational levels were also represented.

Overall, the participants answered the survey positively. As displayed in Tables 2 and 3, each item could have a score contribution between 1 and 5. All the odd-numbered (positive) questions had a score contribution above 4.27-4.53, and all the even-numbered (negative) questions had a score ranging from 1.52 to 1.99. Question 1 had the most positive answers: 94.2% (113/120) strongly agreed or agreed that they would like to use the system if they were hospitalized again. Question 4 had the highest negative score value, indicating that the participants felt they needed help using the system. Of the participants, 50.8% (61/120) indicated that they were confident using the system, answering “strongly agree” to question 9, and 87.5% (105/120) strongly agreed or agreed that most people would be able to learn to use this system.
Table 2. Results of the System Usability Scale for all participants (N=120) and the System Usability Scale score contribution of individual items.

<table>
<thead>
<tr>
<th>System Usability Scale analysis item</th>
<th>Value per 5-point Likert scale response, n (%)</th>
<th>Score contribution (1-5), mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I think I would like to use this system, if I am admitted again.</td>
<td>0 (0) 0 (0) 7 (5.8) 42 (35) 71 (59.2)</td>
<td>4.53 (SD 0.61)</td>
</tr>
<tr>
<td>2. I found the system unnecessarily complex.</td>
<td>61 (50.8) 41 (34.2) 11 (9.2) 5 (4.2) 2 (1.7)</td>
<td>1.72 (SD 0.92)</td>
</tr>
<tr>
<td>3. I thought the system was easy to use.</td>
<td>1 (0.8) 1 (0.8) 8 (6.7) 39 (32.5) 71 (59.2)</td>
<td>4.48 (SD 0.73)</td>
</tr>
<tr>
<td>4. I think that I would need help from the staff to be able to use this system.</td>
<td>49 (40.8) 44 (36.7) 11 (9.2) 11 (9.2) 5 (4.2)</td>
<td>1.99 (SD 1.12)</td>
</tr>
<tr>
<td>5. I found the various functions in the system to be well correlated.</td>
<td>1 (0.8) 1 (0.8) 8 (6.7) 65 (54.2) 45 (37.5)</td>
<td>4.27 (SD 0.69)</td>
</tr>
<tr>
<td>6. I thought there was too much inconsistency in this system.</td>
<td>56 (46.7) 52 (43.3) 8 (6.7) 1 (0.8) 3 (2.5)</td>
<td>1.69 (SD 0.84)</td>
</tr>
<tr>
<td>7. I would imagine that most people would learn to use this system very quickly.</td>
<td>0 (0) 0 (0) 15 (12.5) 53 (44.2) 52 (43.3)</td>
<td>4.31 (SD 0.68)</td>
</tr>
<tr>
<td>8. I found the system very cumbersome to use.</td>
<td>70 (58.3) 42 (35) 5 (4.2) 2 (1.7) 1 (0.8)</td>
<td>1.52 (SD 0.73)</td>
</tr>
<tr>
<td>9. I felt very confident using the system.</td>
<td>2 (1.7) 6 (5) 4 (3.3) 47 (39.2) 61 (50.8)</td>
<td>4.33 (SD 0.89)</td>
</tr>
<tr>
<td>10. I needed to learn a lot things before I could get going with this system.</td>
<td>67 (55.8) 42 (35) 7 (5.8) 4 (3.3) 0 (0)</td>
<td>1.57 (SD 0.75)</td>
</tr>
</tbody>
</table>

Based on the answers to the 2 self-constructed questions, (Table 3), 57.5% (69/120) of the participants strongly agreed that CCL for patients provided a great overview of their stay, and 87.5% (105/120) agreed or strongly agreed that the information in the system made sense to them.

Table 3. Results of general questions calculated by System Usability Scale principles for all participants (N=120) and the System Usability Scale score contribution of individual items.

<table>
<thead>
<tr>
<th>System Usability Scale analysis item</th>
<th>Value per 5-point Likert scale response, n (%)</th>
<th>Score contribution (1–5), mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11. I think the system provided a great overview of my stay.</td>
<td>0 (0) 2 (1.7) 10 (8.3) 39 (32.5) 69 (57.5)</td>
<td>4.46 (SD 0.72)</td>
</tr>
<tr>
<td>12. I think the information in the system made sense to me.</td>
<td>1 (0.8) 5 (4.2) 9 (7.5) 45 (37.5) 60 (50)</td>
<td>4.32 (SD 0.85)</td>
</tr>
</tbody>
</table>

For all participants, the total mean score for the SUS scale was 83.6 (SD 12.8), indicating that the system had close to excellent usability. The median score was 85, and Figure 4 [27] shows the distribution of the individual answers. The scores covered the entire range from 0 to 20 persons per score, and the majority of individuals scored above 70.
Qualitative Results
A total of 10 patients and 3 family members (1 daughter aged 55 years, 1 son aged 65 years, and a husband aged 37 years) were interested in elaborating on their experience of CCL for patients after they had tested the system, and the questionnaire was completed. The patients were aged between 32 and 96 years, with equal representation of men and women, and 3 patients were retired.

The following three main themes emerged from the analysis: (1) future perspectives on usability and design; (2) means toward empowerment; and (3) family implications. These themes will be elaborated on using quotes in the upcoming sections.

Future Perspectives on Usability and Design
The majority of the participants expressed a very positive attitude toward CCL for patients but also offered ideas for the future design of the system. The part of CCL for patients that displayed the estimated waiting time in the ED was found to be intuitive and easy to understand and provided informative insights that prepared the participants for their length of stay. This reduced their frustration with not knowing. However, they expressed concerns about the system’s lack of familiarity and that it could be improved if the design was like other systems they used in everyday life, such as email or smartphone apps.

"The system was not difficult at all but I think it would benefit from more recognizability with others systems, for example, email or iPhone applications." [Male in his 60s]

Most participants valued the line that displayed the boxes with activities the most. They found this part of the system to be essential, as it was the only part that provided direct, personalized information. While they all expressed that they were able to understand the meaning of the changing colors, they also suggested that the text in the boxes could be provided in plain language or a “help” function with text or video could be used to explain the activity in the box.

"The line with the boxes could be much larger, as this is the most important part! It would be great if you could choose whether you would like to see only the line or all actions on the screen." [Joint interview, male patient in his 80s and daughter in her 50s]

The participants all watched more than one video, and there was a consensus that the content in the videos was helpful. A few patients who were placed in the hallway due to crowding found it difficult to listen to the videos not only because of the general noise but also because they were afraid of disturbing others. However, the information provided by using videos instead of text was appreciated.

"The content in the videos was exactly the information I needed. It was nice to be able to revisit the information in the video." [Female patient in her 50s]

A participant found CCL for patients to be too general. More personalized information, such as individual test results, should be incorporated. Moreover, patients who were visually impaired found the system difficult to use.

"I have difficulties with my vision, and I do not think I would have been able to use this without help." [Female patient in her 60s]
Means Toward Empowerment

All of the participants agreed that the system provided an overview that otherwise would not have been accessible for them. Knowing who their treating nurse and doctor were helped them to understand what was going to happen. They described a feeling of not being forgotten in the hectic environment of the ED. Moreover, they valued being able to follow when activities changed from passive to active. Consistency between actions on the screen and in real life provided them with confidence in health care professionals.

When you are here, you can hear people working, but you do not know if anyone is taking care of your situation, or you are forgotten. The system helped us to believe we were not forgotten (...) We loved that when something happens on the screen then it was also reflected in real life. E.g. when the screen said the doctor was on his way- he actually came. [Joint interview, female patient, and husband in their 30s]

Several of the participants stated that having CCL for patients available made them feel calm, as the system provided predictability. Further, having an overview helped them to remain in control of the course of treatment in the ED. Some of the participants said this could save the nurses’ time, as they felt they were more empowered to handle the situation in the ED since they knew what they were waiting for.

The questions that I would have needed a nurse to answer were provided by the system; that was really great. [Female patient in her 30s]

A few of the participants were worried that CCL for patients would need resources from health care professionals that were already scarce.

I am worried that the system takes time away from the patients to support the system. [Joint interview, male patient in his 80s and daughter in her 50s]

Family Implications

Both patients and family members indicated that giving family members direct access was important. CCL for patients gave the family up-to-date information about the care and treatment-related interventions as soon as they attended the hospital room, and they did not need to wait for a nurse or doctor to get an idea of what was planned.

My mother is not able to remember what she is planned for today. I think it was great for me to see she is waiting for X rays. [Female patient in her 90s and son in his 60s]

Furthermore, the family members reported that the system helped them to support the patient, as they could keep track of the interventions provided by CCL for patients. Family members of older patients felt the system was too complicated for the older individuals to use but appreciated that the system was available for them because it allowed them to talk to the patient through the stay in the ED. Moreover, a family member stated that the system made it possible for him to let his wife go to sleep, as they agreed that he would wake her up when he saw that activities were about to happen.

It was nice for me to have a system that told us when things were going to happen. My wife fell asleep, and I knew I did not need to wake her up before I could see the box turned into the blue color. It was easy to understand. [Joint interview, female patient, and her husband in their 30s]

Merged Data

We combined the quantitative and qualitative data in a joint display (Multimedia Appendix 1), providing an assessment of the quantitative and qualitative data together. In this way, the data allow us to expand our understanding of patients and their family members’ experiences with CCL for patients. For example, in question 1, the participants were asked whether they would like to use CCL for patients again. The participants who gave a lower score (1-3) to that question were concerned if the system would replace personal appearance from health care professionals, whereas those who gave it a high score (4-5) valued how the systems helped them to keep control.

Furthermore, question 7 regarding people’s ability to learn to use the system revealed that the participants who gave a low score (1-3) wanted more simplicity, fearing that the older patients would find the system difficult. Meanwhile, the participants who gave high scores (4-5) felt that the system was easy to use. Regarding question 11, the majority of the participants and their family members stated that CCL for patients provided a great overview of the patient’s pathway. They further elaborated on this in the interviews, as they felt that the overview of care in the system helped them to feel less stressed and better understand the treatment pathways.

Discussion

Principal Findings

In this study, we report that the perceived overall usability of the health information system CCL for patients is good to excellent, providing information that is needed during the entire emergency process. The participants rated the system highly (a score of 83.6 points) and reported that the system gave them an opportunity to remain in control, as they knew what they were waiting for and who was responsible for care and treatment.

Technology as a Means to Empower Patients and Family Members in the ED

Looking into previous research on testing systems using SUS [28], a mean score of 83.6, as found in this study, would indicate that the tested system was successful. However, while CCL for patients was evaluated positively overall, we also uncovered technical concerns regarding usability limitations, specifically regarding the older individuals. Our results showed a mean patient age of 57 years, which represents a relatively young ED population. However, the mean age of the family members was almost 10 years (9.5 years) older. The older individuals found the system to be complicated to use and felt that it needed simplified functions, such as a zoom function and recognizability (eg, other well-known systems). Echoing these findings, Verma et al [39] investigated the level of eHealth literacy among older adults and caregivers and found that one main barrier to the adoption of eHealth was a lack of familiarity with the tools.
available. In the development phase of CCL for patients [16], decisions had to be made for the system to work in a clinical setting. One decision was the use of an interface design, which did not allow us to integrate well-known functions, for example, from email or application symbols. Our results highlighted that it might not be possible to design technologies using a one-size-fits-all approach. However, in line with previous research [40], we discovered that the usability testing allowed the developers to adjust and isolate functionalities to provide improved usability outcomes in the future. For example, we found that the participants valued the display with the boxes, which could be promoted in a revised version by the availability of a zoom function.

Furthermore, the participants expressed concerns about whether CCL for patients would influence the health care professionals’ available time to provide actual care. Barriers to the adoption of technology systems in clinical settings include the workflow or demand for more human resources [12]. As the information system is a redesigned patient flow system, it would not require changes in workflow or unduly burden professional health care resources. Another consideration was the need for personal test results. They could not be provided in the current form of the system, as it would require a personal log-on to avoid safety issues related to General Data Protection Regulations.

The participants who rated the usability the highest explained that the system made them feel that they were in control of the situation without the fear of being forgotten. The system provided an overview of the care transition and, therefore, offered predictability. This need to be in control has been identified in another study, which described patients’ and their relatives’ dissatisfaction when visiting the ED [6], as they felt powerless in the ED. Not having knowledge or information available led to such feelings of powerlessness. Nursing rounds were suggested in that study to improve information support [6]. Our results showed that the patients felt more independent because they were able to find the needed information using technology.

Being acutely ill places individuals in a vulnerable situation, and their cognitive capabilities are challenged [2]. Communication from health care professionals and how information is presented have a significant influence on how that information is comprehended [2,41,42]. In this study, we developed information videos related to the journey within the ED, and the participants reported that they were an accessible and usable way to understand information in a stressful situation. Patients and their family members declared that this gave them a feeling of empowerment. Indeed, empowering patients to be in control and involved in their own care is recognized as a core value of high-quality patient-centered care [43]. As Emmamally et al [44-46] noted, improved partnering with family members in the ED is needed. If the family is not included, there is an increased risk of miscommunication and poor understanding of health-related matters [2,44,47-49]. However, creating a closer partnership of care has been described as challenging within the ED due to the high workload, overcrowding, and multitasking [47]. This is echoed in recent findings from studies conducted in a Danish context [2,3], in which family members requested more systematic inclusion in the ED. In this study, the results showed that CCL for patients was perceived as usable and as a useful way to systematically include families during the ED stay.

An update of the Medical Research Council’s guidelines for developing and evaluating complex interventions in health care states that appropriate users should be involved in every part of the development, process, and outcome analysis of a complex intervention to ensure sustainable interventions [50]. In line with best practices, the information system has been developed together with representatives of future users of the system, including health care professionals, managers, patients, family members, and IT specialists [16]. For decades, the ED context has been a hectic environment [4,42,51,52]. This creates challenges at both the information and communication levels, affecting whether patients and their families feel in better control during their stay in the ED [1,4,42,51,52]. In this study, we presented and evaluated a simple but unique system that provides timely information to empower individuals without straining health care professionals’ resources. The usability test was a crucial and important step to inform changes in functionalities and experiences of using IT in the ED.

**Strengths and Limitations**

Questionnaires are a common and recognized method for evaluating the usability of health technologies. However, the contextual factors affecting the results are difficult to determine [53]. The SUS did not provide insights on the effectiveness or efficiency of the system, but it is a validated questionnaire and provided an overall understanding of the system [27]. The mixed methods approach [24] enabled the integration of quantitative and qualitative data. This allowed us to obtain an understanding of how the usability was rated and why the results emerged for the specific questions, which is considered a strength of usability testing [40,54].

Additionally, our findings serve as an inspiration to others about how a participatory design process can develop a technology that is aligned with some of the essential needs described by the users of the ED. The findings provide an example of how a technological solution can be used to reduce the information gap in an ED context, as the provision of adequate information to patients and their families is found to be a major challenge in an ED context [2,4,42].

This study also had some limitations. Using a broader evaluation method, for example, a qualitative evaluation questionnaire or an evaluation instrument with more domains, could potentially have provided the study with more nuances [55]. Patients attending the ED outside of the IT department’s business hours were not able to use the system. Therefore, we do not know if patients attending the ED in the late evening hours or at night would rate the usability differently. Moreover, no cognitive debriefings or adjustments were made specifically for individuals attending an ED, as these tests were conducted before introducing the questionnaire. Multimedia Appendix 2 contains further details about the process as well as final modifications to the questionnaire. In addition, our results are based on a relatively young population (with a mean age of 57 years). Another weakness is that we did not include all users in the evaluation phase, as health care professionals, IT specialists,
and managers were only involved in the development phase and not in the usability testing. For the system to be fully useful, it must run on its own or be serviced directly in the ED. These aspects will be considered in the planning of a future implementation process. Moreover, the transferability of the results is limited to countries with comparable access to and understanding of technologies, as in the Danish population and health care system.

Conclusion

Based on the results of this study, the usability of CCL for patients is rated close to excellent by patients and family members. CCL for patients was perceived to be useful, as it enabled understanding of the ED treatment and pathway. The patients indicated that, from the technology, were able to understand what was going to happen, experienced the feeling of being in control, and found the information to be useful. Areas for improvement include making the system more usable for the older individuals. It is concluded that a technological solution can be used to minimize the information gap in an ED context from the perspective of patients and their family members.

Acknowledgments

The authors would like to thank all those who participated in this study, OPEN [29], and Niels I Agerbek for data management. The authors received financial support for the research, authorship, and publication of this study from the Region of Southern Denmark, the University of Southern Denmark, and Odense University Hospital.

Authors' Contributions

CØ, CMJ, KBD, EC, and AL wrote the protocol for the study. CØ and AL performed the data analysis. CØ wrote the first draft of the manuscript. All the authors reviewed and edited the manuscript and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Joint display.

[PDF File (Adobe PDF File), 55 KB - humanfactors_v11i1e48445_app1.pdf ]

Multimedia Appendix 2

Supplementary file for the Methods section.

[PDF File (Adobe PDF File), 61 KB - humanfactors_v11i1e48445_app2.pdf ]

References


2. Østervang C, Lassen AT, Jensen CM, Coyne E, Dieperink KB. What are the needs and preferences of patients and family members discharged from the emergency department within 24 hours? A qualitative study towards a family-centred approach. BMJ Open 2021;11(11):e050694 [FREE Full text] [doi: 10.1136/bmjopen-2021-050694] [Medline: 34753758]

3. Østervang C, Lassen AT, Øelund K, Coyne E, Dieperink KB, Jensen CM. Health professionals' perspectives of patients' and family members' needs in emergency departments and patient pathway improvement: a qualitative study in Denmark. BMJ Open 2022;12(2):e057937 [FREE Full text] [doi: 10.1136/bmjopen-2021-057937] [Medline: 35228292]


54. Rochat J, Ehrler F, Siebert JN, Ricci A, Ruiz VG, Lovis C. Usability testing of a patient-centered mobile health app for


Intention to Use an Electronic Community Health Information System Among Health Extension Workers in Rural Northwest Ethiopia: Cross-Sectional Study Using the Unified Theory of Acceptance and Use of Technology 2 Model

Tesfahun Hailemariam¹, PhD; Asmamaw Atnafu², PhD; Lemma Gezie³, PhD; Jens Kaasbøll⁴, Prof Dr; Jorn Klein⁵, Prof Dr; Binyam Tilahun¹, PhD

¹Department of Health Informatics, Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia
²Department of Health System and Policy, Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia
³Department of Epidemiology and Biostatistics, Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia
⁴Department of Informatics, University of Oslo, Oslo, Norway
⁵Department of Nursing and Health Sciences Campus Porsgrunn, University of South-Eastern Norway, Porsgrunn, Norway

Abstract

Background: IT has brought remarkable change in bridging the digital gap in resource-constrained regions and advancing the health care system worldwide. Community-based information systems and mobile apps have been extensively developed and deployed to quantify and support health services delivered by community health workers. The success and failure of a digital health information system depends on whether and how it is used. Ethiopia is scaling up its electronic community health information system (eCHIS) to support the work of health extension workers (HEWs). For successful implementation, more evidence was required about the factors that may affect the willingness of HEWs to use the eCHIS.

Objective: This study aimed to assess HEWs’ intentions to use the eCHIS for health data management and service provision.

Methods: A cross-sectional study design was conducted among 456 HEWs in 6 pilot districts of the Central Gondar zone, Northwest Ethiopia. A Unified Theory of Acceptance and Use of Technology model was used to investigate HEWs’ intention to use the eCHIS. Data were cleaned, entered into Epi-data (version 4.02; EpiData Association), and exported to SPSS (version 26; IBM Corp) for analysis using the AMOS 23 Structural Equation Model. The statistical significance of dependent and independent variables in the model was reported using a 95% CI with a corresponding P value of <.05.

Results: A total of 456 HEWs participated in the study, with a response rate of 99%. The mean age of the study participants was 28 (SD 4.8) years. Our study revealed that about 179 (39.3%; 95% CI 34.7%-43.9%) participants intended to use the eCHIS for community health data generation, use, and service provision. Effort expectancy (β=0.256; P=.007), self-expectancy (β=0.096; P=.04), social influence (β=0.203; P=.02), and hedonic motivation (β=0.217; P=.03) were significantly associated with HEWs’ intention to use the eCHIS.

Conclusions: HEWs need to be computer literate and understand their role with the eCHIS. Ensuring that the system is easy and enjoyable for them to use is important for implementation and effective health data management.

(JMIR Hum Factors 2024;11:e47081) doi:10.2196/47081
KEYWORDS

data capturing; data use; eCHIS; electronic community health information system; health extension worker; HEW; intention to use; service provision; Unified Theory of Acceptance and Use of Technology 2; UTAUT2 model

Introduction

Though IT has demonstrated remarkable promise in closing the digital divide in resource-constrained regions and advancing the health care system, there is a global shortage of health workers, which prevents at least half of the world’s population from receiving essential health services [1,2]. Training community health workers (CHWs) in low- and middle-income countries has been a recommended approach to closing the global shortage of health workers [2].

Ethiopia has been implementing the Health Extension Program (HEP) since 2003, comprising female health extension workers (HEWs) to improve the health status of its population [3]. Though various strategies were implemented, and substantial progress was observed in enhancing the community health information system (eCHIS), the performance of HEWs has remained low. The possible reasons for the low performance of HEWs are increased workload; lack of motivation; negligence; and skill gaps in health data production, use, and service provision [4-8].

Due to individual, organizational, and interpersonal level impediments, in most resource-constrained countries, particularly in sub-Saharan regions, health care data generated and used for decision-making are incomplete and inaccurate [9,10]. Likewise, quality health data generation and evidence-based decision-making practices are the remaining challenges for the health care system in Ethiopia [8,11,12].

The growing evidence shows that the penetration of mobile technology improves health service delivery and health outcomes across the world [13-19] and is becoming a solution to strengthen health care industries [8,20-22]. Previous studies in Ethiopia [23], Ghana [24], Uganda [25], South Dakota [26], Indonesia [27], Canada [28], Taiwan [29], South Korea [30], and Jordan [31] indicate that data collection using electronic systems may save time over manual data collection [32,33], and there is the potential to improve health care and the productivity of health staff. For example, digital health solutions may enable CHWs to generate quality health data [34], improve health care delivery [35], and help CHWs be more effective in their job at the community level [32,36].

The eCHIS is one of the evidence-based mobile platforms for CHWs in resource-constrained countries [37], which is an easily customizable mobile health (mHealth) platform for health workers to track and support their interactions with clients. It replaces the conventional practice of a CHW manually tracking their work and carrying large client data and documentation [37].

To tackle the challenges that existed with manual health data generation, use, and service provision, the Ethiopian Federal Ministry of Health has taken the initiative to digitize the existing paper-based Community Health Information System through the eCHIS and started piloting it in 6 districts of the Central Gondar zone, Northwest Ethiopia. The ultimate goal of its implementation is to improve the quality of health data production and service delivery at the community level by transforming the culture of information use by using tablet devices.

The first component of the eCHIS is the HEW component, which supports HEWs in family folder management and the provision of reproductive, maternal, newborn, and child health service delivery and follow-up. The second component is the health center referral component, which enables health center workers to confirm referrals and provide referral feedback to HEWs. The focal person component is the third component, which assists focal persons who are designated at the health center level to provide technical and programmatic support to the HEWs. Therefore, it enables HEWs to manage health post-level data and service provision, as it facilitates referral linkage of clients from health posts to health centers and vice versa.

Although using health system technology has expanded worldwide to leverage quality health data production and use, there is a paucity of evidence on users’ behavioral intention to use health system technology [38]. The intention to use a new system is how much a health care provider intends, plans, and predicts their future behavioral readiness to use health care technology [39]. Studies show that users’ behavioral intention is one of the significant factors of technology acceptance and use.

Hence, it is critical to evaluate the level of users’ intention to use IT before implementing it in the health care system [40-42], as it has a significant role in planning and designing effective implementation strategies for health care programs [43]. Moreover, identifying the level of intention to use the eCHIS for community health data production, use, and service provision and its influencing factors could help to be effective in the implementation and strengthening of the program. To the authors’ understanding, the level of HEWs’ intention to use the eCHIS for community health care data generation, use, and service provision has not been tested using the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) model.

The UTAUT2 model is one of the most mature IT models [44] that has emerged from 8 theoretical models that were primarily developed in psychology and sociology [45]. These include the Technology Acceptance Model, Theory of Planned Behavior, Combined Technology Acceptance Model and Theory of Planned Behavior, Theory of Reasoned Action, Motivational Model, Social Cognitive Theory, Model of PC Utilization, and Innovation Diffusion Theory [45,46].

The UTAUT2 has 3 broad types of integration of concepts. First, the integration was examined in new contexts, new users, and new cultural settings [46]. Second, the addition of new constructs increased the scope of dependent predictors [45]. Third, including independent predictors of the Unified Theory
of Acceptance and Use of Technology (UTAUT) variables made comprehension easier [46]. Its extensive replications, applications, and integration extend the theoretical limits of technology adoption. Therefore, the addition of the 3 predictors (hedonic motivation, price value, and habit) to the previously existing 4 constructs in the original UTAUT model (performance expectancy, effort expectancy, social influence, and facilitating conditions) leveraged the adoption and use of technology (eCHIS in this case). This changes the existing relationships of constructs in the original UTAUT and introduces new relationships among constructs known in the UTAUT2.

We used the UTAUT2 constructs to determine HEWs’ behavioral intention to use the eCHIS [46], as UTAUT2 perspectives are applicable in the health system and the eCHIS is a form of health system technology. Understanding the intention of HEWs using the UTAUT2 model would give insights to health system leaders on how to digitize community health systems in local settings.

Therefore, this study aimed to investigate HEWs’ intention to use the eCHIS and its predicting factors using the UTAUT2 model among HEWs who had received familiarization training on the eCHIS in 6 pilot districts of Northwest Ethiopia.

Since the eCHIS is a form of health system technology, the relationships between UTAUT2 perspectives on accepting and using technology apply to the eCHIS, and the following hypotheses were speculated:

- **Hypothesis 1:** performance expectancy positively influences HEWs’ behavioral intention when using the eCHIS.
- **Hypothesis 2:** effort expectancy positively influences HEWs’ behavioral intention when using the eCHIS.
- **Hypothesis 3:** social influence positively influences HEWs’ behavioral intentions when using the eCHIS.
- **Hypothesis 4:** facilitating conditions positively influence HEWs’ behavioral intentions when using the eCHIS.
- **Hypothesis 5:** hedonic motivation positively influences HEWs’ behavioral intention in using the eCHIS.
- **Hypothesis 6:** self-efficacy positively influences HEWs’ behavioral intention when using the eCHIS.
- **Hypothesis 7:** habit positively influences HEWs’ behavioral intention when using the eCHIS.

In this study, price value was not included in this model because HEWs, the participants in this study, were not directly involved in purchasing the system. Furthermore, the model was not tested on behavioral intention to use the eCHIS among HEWs in Ethiopia.

**Methods**

**Study Design, Period, and Setting**

A cross-sectional study design was conducted from January to February 2021 in the Central Gondar zone, Northwest Ethiopia. The Central Gondar zone has 15 districts, of which 6 districts (Wogera, Mirab Dembia, Misrak Dembia, Enfranz, Takusa, and Belesa) were selected as pilot districts for eCHIS implementation in the zone. The estimated total population of the zone was 2,288,440. The zone has a total of 75 health centers and 404 health posts, and there were 897 HEWs (59 urban and 848 rural) during the study period (Central Gondar Zone Health Bureau report, unpublished data, 2020).

**Population and Participants of the Study**

The source population of the study was HEWs at the primary health care unit level. The study participants were HEWs who were in the pilot districts of the Central Gondar zone and had received initial training for eCHIS implementation. The intervention was skill-oriented training for the implementers of mobile-based community health information system applications based on the training manual prepared by the Ministry of Health, and the training was provided for 1 week by trainers from the regional health bureau and the Ministry of Health. Following the training, each woreda (district) led household registration, tablet usage guideline provision, technical support and mentoring, and periodical communications for 1 year.

**Provision of Mentorship and Technical Assistance**

The University of Gondar assigned three supporting team members who provided technical assistance for implementers with a local mentor every 2 weeks throughout the intervention period. In addition, 1 health information technician (a local mentor) was assigned to provide mentorship and solve eCHIS-related problems during implementation.

**Sample Size and Sampling Procedures**

The initial sample size was calculated using a single population proportion formula, considering the following assumptions: 50% proportion of intention to use the eCHIS, as there was limited evidence in the area; 95% confidence level for estimations; and 5% margin of error. Using these inputs, the initial sample size was estimated at 385. Considering a 10% nonresponse rate, the final sample size was 422. In the pilot districts, however, the total number of HEWs was 460. Therefore, as the initially determined sample size was closer to the population size, it was planned to include all eligible HEWs in the study.

**Study Variables and Measurement**

The dependent variable was the intention to use the eCHIS for health data generation and service provision. Based on the UTAUT2, 8 constructs with a 5-point Likert scale were used to assess the intention to use the eCHIS and were considered potential predictors of the study [46].

1. **Performance expectancy:** the extent to which people believe that using a new technology can improve their job performance [47].
2. **Effort expectancy:** the degree of ease of use associated with the usage of a new technology [46].
3. **Social influence:** the degree of importance others recognize in using a new system [45].
4. **Facilitating conditions:** the degree to which a person perceives that an organization and a technical infrastructure exist to support the intention of people to use technology [45].
5. **Hedonic motivation:** the motivation to do something due to internal satisfaction [48].
6. Habit: the degree to which users perform the usage of
technologies behaviors automatically because of learning [46,49].
7. Self-efficacy: judgment of one’s ability to use technology to
accomplish a particular job or task [45].
8. Behavioral intention: the degree to which a person has
formulated conscious plans to perform or not perform some
specified future behaviors [50].

Data Collection Tools and Procedures
Data collection tools were adapted from the source instrument
used in the UTAUT2 model [46] in the context of the eCHIS
to enhance comprehension by the respondents. The items in the
constructs were performance expectancy (4 items), effort
expectancy (4 items), social influence (3 items), facilitating
condition (4 items), hedonic motivation (3 items), self-expectancy (4 items), habit (4 items), and intention to use
(3 items). The source language of the instrument was translated
forward into the local language of Amharic, and a backward
translation was done to ensure the consistency of the tool.
Experts with health management information system
backgrounds were invited to review the relevance of each
question in the instrument. The experts reviewed the instrument and
checked its content and face validity, and the instrument was
refined according to the comments given. A pretest was
conducted on 5% of the study participants before actual data
collection was started, and the tool was refined based on the
pretest results. A total of 4 data collectors and 2 supervisors
were recruited and trained on the purpose, tools, and procedures
of the study. Self-administered questionnaires were used to
collect data from HEWs with the assistance of data collectors
and supervisors. The data collection period was from January
28 to February 13, 2021, after 2 weeks of eCHIS familiarization
training had been given.

Data Management and Analysis
The data were entered into Epi-data (version 4.02; EpiData
Association) and exported to SPSS (version 26; IBM Corp) for
descriptive statistics such as frequency, cross-tabulations, and
univariate analysis of sociodemographic and model constructs.

Table 1. Structural equation modeling fitness for intention to use electronic community health information system among health extension workers in Northwest Ethiopia, 2021.

<table>
<thead>
<tr>
<th>Fit indices</th>
<th>Threshold value</th>
<th>Authors</th>
<th>Results obtained</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>&lt;3</td>
<td>Bentler [54] (1990)</td>
<td>2.67</td>
<td>Accepted</td>
</tr>
<tr>
<td>Goodness-of-fit index</td>
<td>&gt;0.9</td>
<td>Chau [55] (1997)</td>
<td>0.92</td>
<td>Accepted</td>
</tr>
<tr>
<td>Adjusted goodness-of-fit index</td>
<td>&gt;0.8</td>
<td>Chau [55] (1997)</td>
<td>0.88</td>
<td>Accepted</td>
</tr>
<tr>
<td>Comparative fit index</td>
<td>&gt;0.9</td>
<td>Bentler [54] (1990)</td>
<td>0.97</td>
<td>Accepted</td>
</tr>
</tbody>
</table>
| Root mean square error of
  approximation                   | <0.05           | Browne and Cudeck [56]   (1993) | 0.08 | Accepted   |
| Normed fit index                 | >0.9            | Bentler and Bonett [57]  (1980) | 0.95 | Accepted   |

Reliability and Validity of the Research
Regarding the reliability and validity of the study, Cronbach α
reliability coefficients were computed to determine the internal
consistency of the constructs. Cronbach α of .7 or above
indicates high reliability; between .5 and .7 indicates moderate
reliability; and less than .5 indicates low reliability. We have
used 4-item Likert questions to assess the reliability of the
constructs. Accordingly, the reliability of the constructs assessed
by 3-item questions as follows: performance expectancy
(α=.92), effort expectancy (α=.87), facilitating condition
(α=.75), self-expectancy (α=.88), habit (α=.84), social influence

Simple and multiple structural equation models were carried
out using the AMOS 23 Structural Equation Model in order to
test the relationship between observed and latent variables and
identify the predicting variables of the intention to use the
eCHIS. During analysis, we applied a parceling technique to
increase model efficiency [51]. The subset-item-parcel approach
was used in order to aggregate items into several parcels and
use them as indicators of the target construct [52]. Accordingly,
we created 2 parcels for each factor of target latent constructs
(such as performance expectancy, effort expectancy, facilitator
conditions, self-expectancy, and habit) by aggregating randomly
grouped items within each scale [53]. The remaining 3 latent
target constructs with 3 indicators per construct, such as social
influence, hedonic motivation, and intention to use, remained
as they existed in the original UTAUT2 model [46].

The overall model’s fitness was measured and assessed using
the goodness of fit indices such as chi-square ratio (<3), the
goodness of fit index (>0.9), adjusted goodness of fit index
(>0.8), normal fit index (>0.9), comparative fit index (>0.9),
Tucker-Lewis index (>0.9), and root mean square error of
approximation (<0.08). For the structural equation model,
standardized path coefficients of the regression weight values
were used to estimate the path coefficients of the dependent and
independent variables. Standardized coefficients are not
dependent on the scales as they vary from –1 to 1, where 0
indicates no relationship, 1 indicates a strong positive
relationship, and –1 indicates a strong negative relationship. A
critical ratio (regression weight or standard error) was used to
evaluate whether the constructs had a significant relationship.
The absolute value of a critical ratio greater than 1.96 is an
indication of the significance of the path coefficients. In this
study, the CI and its P value were calculated using
bootstrapping, and the statistical significance of dependent and
independent variables in the model was reported using a 95% CI
with a corresponding P value of <.05 (Table 1). The square
multiple correlation (R²) was used to report the proportion of
variance so that the intention to use the eCHIS could be
explained by the model.
(α=.78), hedonic motivation (α=.90), and intention to use eCHIS (Table 2). In this study, the magnitude of intention to use the eCHIS was assessed by a 3-item Likert question with a reliability test of Cronbach α=.93.

Table 2. Reliability of the constructs on intention to use the electronic community health information system (eCHIS) among health extension workers in Northwest Ethiopia, 2021.

<table>
<thead>
<tr>
<th>Constructs</th>
<th>Sample size</th>
<th>Number of items</th>
<th>Cronbach α</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance expectancy</td>
<td>456</td>
<td>4</td>
<td>.92</td>
</tr>
<tr>
<td>Effort expectancy</td>
<td>456</td>
<td>4</td>
<td>.87</td>
</tr>
<tr>
<td>Social influence</td>
<td>456</td>
<td>3</td>
<td>.78</td>
</tr>
<tr>
<td>Facilitating conditions</td>
<td>456</td>
<td>4</td>
<td>.75</td>
</tr>
<tr>
<td>Hedonic motivation</td>
<td>456</td>
<td>3</td>
<td>.90</td>
</tr>
<tr>
<td>Self-expectancy</td>
<td>456</td>
<td>4</td>
<td>.88</td>
</tr>
<tr>
<td>Habit</td>
<td>456</td>
<td>4</td>
<td>.84</td>
</tr>
<tr>
<td>Intention to use the eCHIS system</td>
<td>456</td>
<td>3</td>
<td>.93</td>
</tr>
</tbody>
</table>

Ethical Considerations

Study approval and ethical clearance were obtained from the University of Gondar’s ethical review board (R.NO. V/P/RCS/05/2020) and a support letter from the ethical review committee of the Amhara Regional Health Bureau Research and Technology transfer office. Study permission was sought at all levels of governmental administration systems including health offices and health facilities. Written consent was obtained, and participants were informed about the objective, importance of the study, procedure and duration, risk and discomfort, benefits of participating in the study, confidentiality, and the right to refuse or withdraw during data collection. To ensure confidentiality, their names and other personal identifiers were not registered. Participants were not compensated for study participation. We confirm that the provided ethics approval documentation covers the study presented in this manuscript.

Results

Sociodemographic and Other Characteristics of the Study Participants

A total of 456 HEWs participated in the study, with a response rate of 99%. The mean age of the study participants was 28 (SD 4.8) years. More than two-thirds (n=314, 68.9%) of the study participants had work experience of more than 5 years. About half of the participants (n=232, 50.9%) were level 4 (10+4) in their educational status, and the majority of the respondents (n=307, 67.3%) were married. The number of HEWs who had difficulties recharging mobile phones was 307 (67.3%). Our study found that 147 (32.2%) HEWs used Microsoft applications daily, 331 (72.6%) had experience using mobile phones for more than 5 years and above, and 421 (92.3%) had informal mobile phone usage practices or were using personal mobile for health post–related activities (Table 3).

According to the findings of this study, 122 (26.8%), 132 (28.9%), and 162 (35.5%) HEWs strongly agreed to intend, predict, and plan to use the eCHIS, respectively (Table 4).
Table 3. Sociodemographic and informal phone use characteristics of the study participants in Northwest Ethiopia, 2021.

<table>
<thead>
<tr>
<th>Variables and categories</th>
<th>Values (N=456), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age group (years)</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;24</td>
<td>104 (22.8)</td>
</tr>
<tr>
<td>25-34</td>
<td>295 (64.7)</td>
</tr>
<tr>
<td>≥35</td>
<td>57 (12.5)</td>
</tr>
<tr>
<td><strong>Marital status</strong></td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>307 (67.3)</td>
</tr>
<tr>
<td>Single</td>
<td>118 (25.9)</td>
</tr>
<tr>
<td>Divorced</td>
<td>26 (5.7)</td>
</tr>
<tr>
<td>Widow</td>
<td>5 (1.1)</td>
</tr>
<tr>
<td><strong>Work experience (years)</strong></td>
<td></td>
</tr>
<tr>
<td>0-2</td>
<td>55 (12.1)</td>
</tr>
<tr>
<td>3-5</td>
<td>87 (19.1)</td>
</tr>
<tr>
<td>&gt;5</td>
<td>314 (68.9)</td>
</tr>
<tr>
<td><strong>Level of education</strong></td>
<td></td>
</tr>
<tr>
<td>Level I</td>
<td>5 (1.1)</td>
</tr>
<tr>
<td>Level II</td>
<td>12 (2.6)</td>
</tr>
<tr>
<td>Level III</td>
<td>196 (43)</td>
</tr>
<tr>
<td>Level IV</td>
<td>232 (50.9)</td>
</tr>
<tr>
<td>Others&lt;sup&gt;a&lt;/sup&gt;</td>
<td>11 (2.4)</td>
</tr>
<tr>
<td><strong>Difficulty with battery recharging</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>307 (67.3)</td>
</tr>
<tr>
<td>No</td>
<td>149 (32.7)</td>
</tr>
<tr>
<td><strong>Using Microsoft applications for work and daily life</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>147 (32.2)</td>
</tr>
<tr>
<td>No</td>
<td>309 (67.8)</td>
</tr>
<tr>
<td><strong>Do you use personal mobile phone for health post-related activities?</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>421 (92.3)</td>
</tr>
<tr>
<td>No</td>
<td>35 (7.7)</td>
</tr>
<tr>
<td><strong>For how long you have used mobile phone (years)?</strong></td>
<td></td>
</tr>
<tr>
<td>0-5</td>
<td>125 (27.4)</td>
</tr>
<tr>
<td>&gt;5</td>
<td>331 (72.6)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Health extension worker with additional diploma, BSc degree, or both.

Table 4. Health extension workers’ intention to use the electronic community health information system (eCHIS) in Northwest Ethiopia, 2021 (N=456).

<table>
<thead>
<tr>
<th>Items</th>
<th>Strongly disagree, n (%)</th>
<th>Disagree, n (%)</th>
<th>Neutral, n (%)</th>
<th>Agree, n (%)</th>
<th>Strongly agree, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I intend to use the eCHIS system in the future</td>
<td>7 (1.5)</td>
<td>12 (2.6)</td>
<td>12 (2.6)</td>
<td>303 (66.4)</td>
<td>122 (26.8)</td>
</tr>
<tr>
<td>I predict I will use the eCHIS system in the future</td>
<td>5 (1.1)</td>
<td>13 (2.9)</td>
<td>9 (2)</td>
<td>297 (65.1)</td>
<td>132 (28.9)</td>
</tr>
<tr>
<td>I plan to use the eCHIS system in the future</td>
<td>4 (0.9)</td>
<td>11 (2.4)</td>
<td>15 (3.3)</td>
<td>264 (57.9)</td>
<td>162 (35.5)</td>
</tr>
</tbody>
</table>
Mean Score of All Predictors and Intention to Use the eCHIS Using the UTAUT2 Model

The mean scores of performance expectancy, effort expectancy, facilitating condition, self-expectancy, and habit with 4-item Likert questions were 17.09 (SD 2.58), 16.22 (SD 2.41), 12.07 (SD 3.46), 13.94 (SD 3.62), and 14.75 (SD 3.14), respectively. On the other hand, social influence, hedonic motivation, and intention to use the eCHIS with 3-item Likert questions had a mean score of 11.63 (SD 2.32), 12.23 (SD 2.01), and 12.57 (SD 2.00), respectively.

Our study revealed that 179 (39.3%; 95% CI 34.7%-43.9%) participants who had the intention to use the eCHIS for community health data generation, use, and service provision had scored above the mean. The mean score of the intention to use the eCHIS was 12.57 (SD 2.00). The maximum score of intention to use the eCHIS was 15, while the minimum score was 3.

Simple Structural Equation Model Analysis

The variance in the dependent variable explained by the independent variables was interpreted using square multiple correlation ($R^2$). The overall $R^2$ of the intention to use the eCHIS is found to be 32%, the variance that was explained by the independent variables in the model. The bootstrap method with a 95% bias-corrected CI was applied to investigate the significance of path coefficients and factors predicting the model. The predictors with $P<.20$ in the simple structural equation model were considered candidate variables for multiple structural equation model analysis. Due to its undependability to scale, we used a standardized beta coefficient to interpret the influence of predictors on the intention to use the eCHIS. A 95% CI with $P<.05$ was considered to declare an association between dependent and independent variables. The study indicated that effort expectancy has the highest direct effect on HEWs’ intention to use the eCHIS, followed by hedonic motivation. The remaining model constructs that have a direct influence on predicting intention to use the eCHIS are social influence and self-expectancy. The structural equation model predicted, with the path coefficients and $R^2$, is represented in Figure 1, and the path coefficients and $P$ value found from the depicted model are presented in the Results section. Moreover, the absolute value of the critical ratio of effort expectancy (3.701), self-expectancy (2.468), social influence (2.782), and hedonic motivation (3.311) indicated that predictors had a significant influence on HEWs’ intention to use the eCHIS. Overall, 32% of the variance with respect to intention to use the eCHIS was reasonably explained by the predictors in the model.

Figure 1. Predictors and intention to use the electronic community health information system among health extension workers at Central Gonda zone, Northwest Ethiopia, 2021. EE: effort expectancy; FC: facilitating condition; HB: Habit; HM: hedonic motivation; ITU: intention to use; PE: performance expectancy; SE: self-expectancy; SI: social influence.
Effort expectancy (the extent to which people believe that using the eCHIS can improve their effort) has a positive influence on HEWs’ behavioral intention ($\beta=.256; P=.007$). Similarly, self-efficacy and social influence had a positive influence on HEWs’ behavioral intention ($\beta=.096; P=.04$), and ($\beta=.203; P=.02$), respectively. Likewise, hedonic motivation to use eCHIS due to internal satisfaction was found to be ($\beta=0.217; P=.03$) and had a significant effect on intention to use the eCHIS. Facilitating conditions ($\beta=0.005; P=.92$), habit ($\beta=0.103; P=.07$), and performance expectancy ($\beta=0.034; P=.61$) had no significant influence on intention to use the eCHIS (Table 5).

Table 5. Multiple structural equations modeling association between predictors and intention to use the electronic community health information system among health extension workers in Northwest Ethiopia, 2021.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Estimate</th>
<th>95% CI</th>
<th>$P$ value</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE$^c$ ⇒ IU$^b$</td>
<td>0.034</td>
<td>-0.91 to 0.190</td>
<td>.61</td>
<td>Not supported</td>
</tr>
<tr>
<td>EE$^c$ ⇒ IU</td>
<td>0.256</td>
<td>0.060 to 0.503</td>
<td>.007</td>
<td>Supported</td>
</tr>
<tr>
<td>FC$^d$ ⇒ IU</td>
<td>0.005</td>
<td>-0.060 to 0.079</td>
<td>.92</td>
<td>Not supported</td>
</tr>
<tr>
<td>SE$^e$ ⇒ IU</td>
<td>0.096</td>
<td>0.004 to 0.191</td>
<td>.04</td>
<td>Supported</td>
</tr>
<tr>
<td>HB$^f$ ⇒ IU</td>
<td>0.103</td>
<td>-0.008 to 0.235</td>
<td>.07</td>
<td>Not supported</td>
</tr>
<tr>
<td>SI$^g$ ⇒ IU</td>
<td>0.203</td>
<td>0.039 to 0.390</td>
<td>.02</td>
<td>Supported</td>
</tr>
<tr>
<td>HM$^h$ ⇒ IU</td>
<td>0.217</td>
<td>0.022 to 0.416</td>
<td>.03</td>
<td>Supported</td>
</tr>
</tbody>
</table>

$^a$PE: performance expectancy.
$^b$IU: intention to use.
$^c$EE: effort expectancy.
$^d$FC: facilitating condition.
$^e$SE: self-expectancy.
$^f$HB: habit.
$^g$SI: social influence.
$^h$HM: hedonic motivation.

**Discussion**

**Principal Findings**

In this study, nearly 2 out of 5 HEWs had an intention to use the eCHIS for community health data generation, use, and service provision. Effort expectancy, self-expectancy, social influence, and hedonic motivation were statistically significant predictors of intention to use the eCHIS. The intention to use the eCHIS by HEWs could be associated with the fact that using the eCHIS is not difficult to understand. It saves time and reduces the amount of effort required to complete health-related tasks [58,59]. Furthermore, it simplifies activities and helps them access data easily. The other could be people around them who have the ability to influence their intention to use the system [58]. For example, HEWs’ activities should be monitored and evaluated by health system leaders. If they give them more attention, they will be encouraged to use the system. The other could be previous exposure to using informal phone for health system activities, such as reminding clients about their health care appointments and facilitating referral linkage between health centers and health posts, as mHealth enhances communication between health workers and clients [58]. Moreover, using the eCHIS creates a conducive environment for HEWs since their usual data handling approach is exhaustive and takes much time to execute activities at the health post level, and using the eCHIS not only helps them to save their time but also creates motivation to do their job at health post level [58,60].

Regarding factors associated with intention to use the eCHIS, effort expectancy had a positive influence on the intention to use the eCHIS among HEWs. This finding was in accordance with a study conducted in Ethiopia [61], Kenya [61], the United States [62], and Portugal [63] and had a positively significant association with the intention of health care providers to use technology. A possible explanation could be the fact that the less effort the user devotes to using the system, the more likely he or she is to continue to use it. A study in this regard showed that individuals often want to face a system that is easy to use [64]. HEWs might perceive that the eCHIS could help them to do their job aids shortly with less strain and increased work efficiency [65], as using the eCHIS would simplify the tasks they are expected to deliver at health post level. A review in this regard showed that using digital tools simplifies work and helps to access data easily [59]. Furthermore, studies indicate that digital health solutions reduce workload and improve work performance [24,25], reduce errors [34], create motivation and learning opportunities [66], promote health care appointment [67], and are easy to use and improve work efficiency [59]. Using the eCHIS could reduce the workload of HEWs since manual data management practice at health post level is exhaustive and takes much time to collect data and conduct routine activities [59]. Moreover, the referral linkage integrated into the eCHIS, including HEWs, midwives, and focal persons, will harmonize HEWs’ activity flow from health posts to health centers and vice versa. Furthermore, using mHealth motivates

https://humanfactors.jmir.org/2024/1/e47081
The intention to use the eCHIS among HEWs who perceived people around them could influence their behavioral intention was positively associated. The current finding corroborates studies conducted among health care providers using the UTAUT2 model in Ethiopia [61], Morocco [68], Taiwan [62], South Korea [30], and the United States [69], showing that social influence significantly predicted health care providers’ intention to use technology. The possible explanation could be that HEWs might perceive peer pressure from health care staff at the woreda and facility levels toward using the eCHIS, which could positively influence their intention to use the eCHIS. The other justification could be the fact that HEWs might get trust from the community for the job aids or activities they are expected to deliver. Hence, health system staff need to understand that peer influence has a positive effect on using a new system. Moreover, making people aware of a new system at the woreda and facility levels in general and at the kebele leaders, women’s development army, and voluntary service providers’ levels, in particular, could influence HEWs’ behavioral intention to use the eCHIS. A study in this regard showed that the more health workers connected to colleagues, the more they improved the use of digital tools and the quality of care [58].

Our study revealed that the magnitude of intention to use the eCHIS among HEWs who had self-expectancy was positively correlated. The findings of past studies in Ethiopia [61], Malaysia [70], Taiwan [71], and Iran [72] showed that digital literacy was correlated with the intention to use technology in health care industries. The possible reason might be that those who had self-expectancy could not face difficulty in adapting the emerging technology to community-level data management and service provision. The current evidence in the feasibility and effectiveness study on digital health indicated that the level of computer literacy had influenced digital health implementation among CHWs [73]. A possible explanation might be the fact that informal mobile phone usage practices of CHWs for health post–related activities could influence behavioral intention to use the eCHIS. A study indicated that in many different settings, CHWs use their personal phone informally for community-based activities so as to fill the gaps in the health care system [74].

Our study revealed that there is a significant association between intention to use the eCHIS and hedonic motivation or perceived enjoyment from using the eCHIS for community health data generation, use, and service provision. A possible explanation could be the fact that using a new system instead of the usual approach to manage community-level data and service provision may create intrinsic motivation for HEWs to obtain fun or pleasure. A study showed that motivation is an important construct for eHealth users, and it could even be a sufficient reason to adopt newly emerging technology in a contextual environment [75]. In addition, using eHealth technology to deal with community health data generation, use, and service provision may be an enjoyable process and will have a positive influence on the behavioral intentions of the users [72]. HEWs were optimistic about using the eCHIS because it could be related to the production of quality health data, ease of data management, reduced errors and false reports, data protection, and increased accessibility. A study also indicated that using digital tools could enhance the productivity of CHWs [76]. Community health digitization using mobile apps support the services delivered by CHWs [77]. Furthermore, studies show that the digitization of health care data has promising results in improving both health care and health outcomes [13-19] and improving health staff productivity and work efficiency [65]. In our study, the level of users’ optimistic perceptions of using the eCHIS could be an advantage in implementing the intervention, as compared to the existing approach [78]. Studies showed that digital health solutions enable CHWs to generate quality health data [34], improve health care delivery [35], and help them to be more effective in their job aids [32,36]. Moreover, digital tools could help them follow the correct order or the required service elements that clients should receive when providing services. It also enables them to communicate with clients in a better way as compared to manual communication since the tool has prespecified data elements that should be asked by CHWs during service delivery [58].

Likewise, it creates enjoyment among users [66] and benefits them by keeping data safe from human and natural factors that could damage the data. In addition, enjoyment could emanate due to the fact that using digital tools can improve data capturing, storing, and reporting of more items that could be more time-consuming during manual data handling and reduce the motivation of health workers to keep data recording [59]. Even though HEWs are optimistic to use the eCHIS, lack of adequate resources for eCHIS implementation at the implementation district could hinder its successful implementation, and therefore resource availability is vital to be effective in community health digitization. Studies show that challenges during digital health solutions implementation, such as the initial and ongoing capacity-building training [73], poor network access and poor access to electricity [58,79], low financial investment [73], and unreliability or absence of infrastructure (eg, electricity and network) [80,81], hinder the implementation. As the skills and knowledge of HEWs vary from one to another, there should be mentoring and supportive supervision during the implementation. Studies showed that the inability to use the system could affect its implementation [26], and intensive training with continuous refreshment could help them realize the digitization of the community health information system [82].

Limitations of the Study

The findings of this study should be interpreted in light of some limitations. Due to the nature of the study, which was cross-sectional, the inability to infer cause-and-effect relationships is present. As the study was focused on HEWs’ intention to use the eCHIS in a pilot district in Northwest Ethiopia, the sample size could affect the findings of this study, and covering larger areas at the regional and national levels is possible, including urban HEWs. Finally, the parcel approach used in this study may introduce parameter estimation bias.
Conclusion
In conclusion, 39.3% (179/456) of HEWs scored above the mean of intention to use the eCHIS for community health data management and service provision. Factors associated with the intention to use the eCHIS were effort expectancy, self-expectancy, social influence, and hedonic motivation. The eCHIS has numerous advantages and a promising future in terms of improving data quality, use, and service delivery. Its adoption in the country, however, should focus on identifying all necessary prerequisites for successful implementation and advancing the community health information system. The implementation of the eCHIS should not skip factors that had no significant effect on intention to use the eCHIS, and further studies at the regional and national levels are recommended to investigate their correlation with intention to use the eCHIS. Model explainability was found in the study using factors that existed in the UTAUT2 model; however, it is recommended to examine the moderating effects of CHWs' related variables to examine how the model constructs could influence HEWs' intention to use the eCHIS.

Acknowledgments
We would like to thank the Central Gondar Zone Health Department, electronic community health information system pilot district health offices, and the University of Gondar Comprehensive Specialized Hospital for their provision of necessary information and support during data collection. Our gratitude also goes to the study participants, data collectors, and supervisors who took part in the study. This work would not have been possible without the financial support of the Doris Duke Charitable Foundation (grant 2017187). The mission of the Doris Duke Charitable Foundation is to improve the quality of people’s lives through grants supporting the performing arts, environmental conservation, medical research, and child well-being, and through the preservation of the cultural and environmental legacy of Doris Duke’s properties.

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

Authors’ Contributions
TH conducted the study. All authors conceptualized the design of the study, provided a review of the methodology and results, contributed to the reviewing the manuscript, and read and approved the final manuscript.

Conflicts of Interest
None declared.

References


Abbreviations

CHW: community health worker

eCHIS: electronic community health information system

HEP: Health Extension Program

HEW: health extension worker

mHealth: mobile health

UTAUT: Unified Theory of Acceptance and Use of Technology

UTAUT2: Unified Theory of Acceptance and Use of Technology 2

©Tesfahun Hailemariam, Asmamaw Atanau, Lemma Gezie, Jens Kaasbøll, Jorn Klein, Binyam Tilahun. Originally published in JMIR Human Factors (https://humanfactors.jmir.org), 04.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Human Factors, is properly cited. The complete bibliographic information, a link to the original publication on https://humanfactors.jmir.org, as well as this copyright and license information must be included.
Interest in mHealth Among Patients With Low Back Pain: Cross-Sectional Study

Jonas Ammundsen Ipsen1,2, MSc; Louise Fleng Sandal3, PhD; Natalie Hong Siu Chang1,4, MSc; Berit Schiøttz-Christensen5, MD, PhD; Karen Søgaard3,6, PhD; Anders Hansen1,4, PhD

1Department of Regional Health Research, University of Southern Denmark, Odense, Denmark
2Department of Physical Therapy and Occupational Therapy, Lillebaelt Hospital, University Hospital of Southern Denmark, Kolding, Denmark
3Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, Odense, Denmark
4Medical Research, Spine Centre of Southern Denmark, University Hospital of Southern Denmark, Middelfart, Denmark
5Research Unit of General Practice, University of Southern Denmark, Odense, Denmark
6Department of Clinical Research, University of Southern Denmark, Odense, Denmark

Corresponding Author:
Jonas Ammundsen Ipsen, MSc
Department of Physical Therapy and Occupational Therapy
Lillebaelt Hospital
University Hospital of Southern Denmark
Sygehusvej 24
Kolding, 6000
Denmark
Phone: 45 76362832
Email: jonas.ammundsen.ipsen@rsyd.dk

Abstract

Background: Digitally supported self-management tailored to an individual's need, in addition to usual care, may reduce pain-related disability compared to usual care alone, and patients with low back pain (LBP) using mobile health (mHealth) solutions express positive experiences. Hence, implementing mHealth solutions designed to support self-management is desirable from a clinical and patient perspective. Easily accessible mHealth solutions that can support the self-management of patients with LBP are available, but interest may be subgroup specific. Understanding the characteristics and preferences of patients with LBP labeled as interested may help to reach relevant LBP patient groups and inform the development and implementation of effective interventions with mHealth for patients with LBP.

Objective: This study aims to explore the proportion of patients with LBP labeled as interested in testing an mHealth solution designed to support self-management in addition to usual care and to assess how these patients differ from those who were labeled as not interested.

Methods: This exploratory cross-sectional study analyzed demographic and patient-reported outcomes from the SpineData registry, a Danish registry of patients with LBP in an outpatient setting. Between February and December 2019, the SpineData registry was used to assess the preliminary eligibility of patients for a clinical trial (selfBACK). Patients were labeled as interested or uninterested depending on if they responded to an invitation to be tested for eligibility for the trial. Outcomes were selected from the International Classification of Functioning core set of LBP using a clinical approach. Associations were assessed in a backward selection process, and the proportion of variance explained was assessed with pseudo-$R^2$ statistic.

Results: This study included 843 patients, with 181 (21%) individuals labeled as interested in participating in the selfBACK trial. Notably, the cohort labeled as interested differed from their uninterested counterparts in two key aspects: age (36-65 years: 116/181, 64.1% vs 347/662, 52.4%; $P=.003$) and smoking status (smokers: 22/181, 12.5% vs 174/662, 26.6%; $P<.001$). Those aged 36-65 years had higher odds of being labeled as interested compared to individuals aged 18-35 years (odds ratio [OR] 0.43, 95% CI 0.26-0.71) and those 65 years or older (OR 0.77, 95% CI 0.53-1.15). Nevertheless, age accounted for only a modest proportion of variance ($R^2=0.014$). Smokers demonstrated lower odds of being labeled as interested (OR 0.39, 95% CI 0.24-0.64), with smoking status explaining a similarly small proportion of variance ($R^2=0.019$). Collectively, age and smoking status accounted for 3.3% of the variance.
Conclusions: Our investigation revealed that 181 (21%) individuals with LBP invited to participate in the mHealth solution trial for self-management expressed interest. Generally, the characteristics of those labeled as interested and uninterested were comparable. Of note, patients aged 36-65 years had a higher frequency of being labeled as interested compared to their younger and older counterparts.

KEYWORDS
low back pain; mHealth solutions; mobile health; characteristics; patient interest; transferability; representativeness

Introduction

Digital health interventions delivered with smartphones (mobile health [mHealth] solutions) are accessible to most patients across age, geography, and socioeconomic status. Thus, clinicians’ expectations of mHealth solutions are significant, and the availability of new solutions on the commercial market every day also indicates a strong general interest in using mHealth solutions [1,2]. Nevertheless, many mHealth solutions have limited download rates, and if downloaded, the use can be scarce [3,4]. This discrepancy may indicate a need for a better understanding of potential users and their characteristics.

For patients with low back pain (LBP; not attributable to a recognizable, known specific pathology such as infections, fractures, or structural deformity), self-management support is recommended as the first line of treatment [5-8]. This may involve empowering patients to know when to consult for diagnostic assessment, symptom relief, or advice [9]. Digitally supported self-management may be delivered through smartphone apps or digital platforms to facilitate and enhance such self-management practices. Research indicates that the integration of such digitally supported self-management strategies, when combined with standard care, can lead to a reduction in pain-related disability [10]. Further, evidence supports that mHealth solutions designed to support self-management are accepted by patients with chronic LBP [9]. Therefore, there is a growing interest in implementing mHealth solutions designed to support self-management from both clinical and patient perspectives. However, despite the potential benefits, the level of acceptance and use of these interventions remains an area that requires further investigation.

However, studies on other patient groups using mHealth solutions report that lower age, higher education, higher income, higher BMI, and higher self-perceived health are associated with increased use [4,11]. In contrast, the cost of using these apps is a significant barrier [11].

Individuals with LBP who use mHealth solutions to self-manage may thus represent a specific subset within the general population. Therefore, this study aimed to investigate the percentage who expressed interest in participating in a trial evaluating an mHealth solution designed for self-management alongside standard care, as well as to evaluate potential distinctions between those who were labeled as expressing interest and those who were not.

Methods

Study Design

This exploratory cross-sectional study used demographic and patient-reported outcomes (PROs) from an internet-based multiuser clinical registry (SpineData) [12]. Reporting follows the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [13].

Setting

Data were collected at the Spine Centre of Southern Denmark, an outpatient hospital that performs clinical spine evaluations [12]. General practitioners or chiropractors typically refer patients to the Spine Centre, which performs a multidisciplinary assessment of its patients, with more than 10,000 new cases yearly.

Before patients are evaluated at the Spine Centre, they provide information in the local SpineData registry [14]. The registry is designed based on the biopsychosocial model of health, and information is collected across the health domains of pain, activity limitation, work participation, psychological factors, physical impairment, and contextual factors [12]. To mitigate nonresponse and missing information, SpineData uses a “waterfall” model (eg, patients in employment are not asked to respond to causes for unemployment). SpineData has an overall completion rate of 80% and approximately 60% of patients agreed to their responses being used for research [14]. The use of this registry allows for the comprehensive assessment of patients consulted at the Spine Centre and provides a rich source of data for research studies, such as the one presented in this paper.

Participants

Between February and December 2019, SpineData was used to identify eligible patients based on the following criteria: consenting to be contacted for research projects, proficiency in Danish, and experiencing LBP in the past 14 days that exceeded their leg pain in severity. Patients with previous back surgery, who were actively filing for a pension, or who were younger than 18 years were not invited. All patients matching the eligibility criteria were sent a letter of invitation to hear more about the selfBACK trial. One reminder was sent. The patients who did not respond to either invitation or reminder were labeled uninterested. The selfBACK trial investigated the effectiveness of the selfBACK digital decision support system that provided patients with LBP individually tailored digital support in an app format using three content domains: (1) physical activity, (2) education, and (3) exercise programs. The trial investigated the
additive effect of the selfBACK system in addition to usual care. Participants in this trial were recruited from primary health care such as chiropractors, physiotherapists, and general practitioners in addition to the Spine Center of Southern Denmark. Recruitment was performed in Denmark and Norway. The population within this study concerns the pool of patients seen at the Spine Center, who would have received an invitation to eligibility screening to the selfBACK trial based on their answers given in the SpineData clinical registry. In this study, all patients who matched the preliminary eligibility criteria for the selfBACK trial were included [15].

Outcomes
The variables of interest were selected from the SpineData registry, based on the International Classification of Function core set for LBP and clinical reasoning [16]. The demographics and clinical characteristics comprised the domains of pain, activity limitation, work participation, and psychological and contextual factors (Textbox 1).
**Textbox 1.** Detailed description of the content and handling of included outcomes.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Description and Handling</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td>Male or female</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td>Patients were categorized into age groups ≤35, 36-65, and &gt;65 years.</td>
</tr>
<tr>
<td><strong>BMI</strong></td>
<td>The anthropometric variables of height and weight were used to calculate BMI (kg/m²).</td>
</tr>
<tr>
<td><strong>Smoker status</strong></td>
<td>Categorical variable that was dichotomized to smoker and nonsmoker strata. If a patient indicated cigarette use of any kind, they were categorized as a smoker.</td>
</tr>
<tr>
<td><strong>Alcohol consumption</strong></td>
<td>Categorical variable that was stratified into two groups based on the consumption of more than 14 alcoholic beverages a week. The threshold was based on the recommendation of the Danish Ministry of Health [17].</td>
</tr>
<tr>
<td><strong>Comorbidities</strong></td>
<td>This variable was based on four dichotomous variables: allergies, including medication; cancers; heart disease; and lung disease. If a patient replied yes to one of these variables, they were categorized as having comorbidities.</td>
</tr>
<tr>
<td><strong>Current work status</strong></td>
<td>The work status variables consisted of different ways of participating in the labor market: working full- or part-time, flex job, in education, job training due to inability to maintain habitual job function, unemployed, early retirement, pensions, stay at home, and other. The variable was dichotomized to working or not by grouping patients indicating working part- or full-time, flex, and students in one group and the remainder in another group.</td>
</tr>
<tr>
<td><strong>Multiple pain sites</strong></td>
<td>SpineData contains a freehanded pain drawing. The pain drawing was post defined into 46 anatomical regions. In this study, the regions were grouped into 9 areas: neck, shoulders, upper back, elbows, lower back, wrists/hands, hips/thighs, knees, and ankles/feet, inspired by Øverås et al [18]. Patients with two or more pain sites were considered as having multiple pain sites.</td>
</tr>
<tr>
<td><strong>Average back pain</strong></td>
<td>The average back pain in the last 14 days was measured on a 0-10 numeric rating scale, with 10 indicating the worst imaginable pain.</td>
</tr>
</tbody>
</table>
| **STarT Back screening tool [19]** | The STarT Back scores categorize patients into three strata based on their risk of developing chronicity: low risk, moderate risk, and high risk of chronicity:  
  - Low risk: <3  
  - Moderate risk: ≥4 and subscore ≤3  
  - High risk: ≥4 and subscore ≥4 |
| **EQ-5D-5L-VAS [20]**           | Numeric rating scale score spanning from 0 to 100, with 100 representing the best possible health state |
| **Oswestry Disability Index (ODI) [21]** | The ODI is a questionnaire containing 10 items that are scored from 0 to 5. The maximum score is 50 points, which indicates that the patient is bedbound. The ODI has been found valid for patients with low back pain [22].  
  - To estimate the patients' functional level, the ODI Stata package was used. The ODI package allows for the imputation of data for one missing value. The missing values in one section were replaced with the average score for all sections. |
| **Anxiety [23]**                 | Numeric score rating from 0 to 10, with 0 indicating no anxiety and 10 a high degree of anxiety |
Social isolation [23]
- Numeric score rating from 0 to 10, with 0 indicating no loneliness and 10 a high degree of loneliness

Catastrophization (terrible pain that will never improve) [23]
- Numeric score rating from 0 to 10, with 0 indicating no catastrophization and 10 a high degree of catastrophization

Catastrophization (the pain is overwhelming) [23]
- Numeric score rating from 0 to 10, with 0 indicating no catastrophization and 10 a high degree of catastrophization

Risk of persisting pain [23]
- Numeric score rating from 0 to 10, with 0 indicating no risk of persisting pain and 10 a high risk of persisting pain

Feelings of sadness, depression, or hopelessness [23]
- Numeric score rating from 0 to 10, with 0 indicating no feelings of depression and 10 a constant presence of depression

Loss of interest or joy [23]
- Numeric score rating from 0 to 10, with 0 indicating no loss of interest or joy and 10 never feeling interest or joy

Fearing activity will damage the back [23]
- Numeric score rating from 0 to 10, with 0 indicating no fear that physical activity will damage the back and 10 completely agreeing that physical activity will damage the back

Fearing activity will increase the pain [23]
- Numeric score rating from 0 to 10, with 0 indicating completely disagreeing to avoid physical activity and 10 completely agreeing to avoid physical activity

**Exposure**

Patients were allocated into two groups based on their response to being invited for eligibility screening for the selfBACK trial. Those who responded positively to the invitation to be screened were labeled as interested in using the mHealth solution, whereas those who did not respond were labeled as uninterested.

**Statistical Methods**

The demographics and baseline characteristics of patients who were or were not labeled as interested in the digital mHealth intervention were assessed using the $\chi^2$ test for categorical variables and 2-tailed Student $t$ test for continuous variables. Baseline characteristics are reported as the proportion and percentage or mean and SD.

To assess the strength of associations between PROs and patients labeled as interested in mHealth or not, we used univariate and multivariate logistic regression analysis with an odds ratio (OR) and 95% CI. The associations were assessed in a backward selection process, and the proportion of variance explained was assessed with McFadden pseudo-$R^2$ statistic. Statistical analyses were performed with Stata statistical software (Release 17; StataCorp LLC). Missing information was handled using pairwise deletion. The ODI Stata package allows for data imputation for one missing value. The missing values in one section were replaced with the average score for all sections. To avoid overparameterizing the model, we aimed for a 1:10 patient-to-variable ratio.

**Ethical Considerations**

The Region of Southern Denmark was the data controller for this project, which is included in its records on personal data processing activities (file 21/13433). Data processing in the project was regulated by the Danish Act on Research Ethics Review of Health Research Projects section 14, subsection 2, which states that health research based solely on questionnaire surveys and registry data is exempt from the obligation to notify the committees. Following the Danish Health Care Act, we obtained approval for using hospital record data for scientific purposes from the council of the Region of Southern Denmark (file 21/25588). After merging, analyses were run on pseudonymized data, and the results presented in this manuscript do not enable the identification of single data participants. Hence, following national laws, no additional informed consent was collected and no remuneration was offered to patients.

**Results**

**Overview**

From February to the end of December 2019, 5796 patients (~80% of those invited) completed the SpineData registry before their diagnostic assessment at the Spine Centre. Of the total sample, 843 (15%) were invited to the selfBACK trial. The mean age of the cohort was 52 (SD 16.2) years, with an even distribution of sexes (male: n=429, 50.1%), and a mean BMI of 27.5 kg/m².
Of the 843 patients invited to the eligibility screen for the trial, 181 (21%) accepted the invitation and were stratified into the group who were labeled as interested in the mHealth solution. Of the 21 included variables, 8 had complete responses, and none of the remaining 13 variables had more than 2.5% missing responses.

Comparison of Patients Who Were Labeled as Interested and Uninterested in an mHealth Solution

Patients labeled as interested in using the mHealth solution were aged 36-65 years ($P<.003$) and had a lower proportion of smokers ($P<.001$) compared to the patients labeled as uninterested. The remaining variables were not different between the patients labeled as interested and uninterested (Table 1).

Table 1. Baseline characteristics of patients labeled as interested in the mobile health solution compared to the uninterested patients.

<table>
<thead>
<tr>
<th>Baseline characteristic</th>
<th>Interested (n=181)</th>
<th>Uninterested (n=662)</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female, n (%)</td>
<td>93 (51.3)</td>
<td>321 (48.5)</td>
<td>.49</td>
</tr>
<tr>
<td>Age (years), n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-35</td>
<td>21 (11.6)</td>
<td>146 (22.1)</td>
<td>.003</td>
</tr>
<tr>
<td>36-65</td>
<td>116 (64.1)</td>
<td>347 (52.4)</td>
<td></td>
</tr>
<tr>
<td>&gt;65</td>
<td>44 (24.3)</td>
<td>169 (25.5)</td>
<td></td>
</tr>
<tr>
<td>BMI (kg/m$^2$), mean (SD)</td>
<td>28.1 (5.8)</td>
<td>27.2 (5.0)</td>
<td>.05</td>
</tr>
<tr>
<td>Smokers, n (%)</td>
<td>22 (12.5)</td>
<td>174 (26.6)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>&lt;14 alcohol consumption per week, n (%)</td>
<td>173 (95.5)</td>
<td>636 (96.0)</td>
<td>.76</td>
</tr>
<tr>
<td>Has comorbidities, n (%)</td>
<td>84 (46.7)</td>
<td>208 (42.3)</td>
<td>.29</td>
</tr>
<tr>
<td>Working, n (%)</td>
<td>95 (52.4)</td>
<td>367 (55.4)</td>
<td>.49</td>
</tr>
<tr>
<td>Has multiple pain sites, n (%)</td>
<td>136 (75.1)</td>
<td>462 (69.7)</td>
<td>.24</td>
</tr>
<tr>
<td>Average back pain (score range: 0-10), mean (SD)</td>
<td>6.3 (2.0)</td>
<td>6.3 (1.9)</td>
<td>.91</td>
</tr>
<tr>
<td>STarT Back tool, n (%)</td>
<td></td>
<td></td>
<td>.34</td>
</tr>
<tr>
<td>Low risk</td>
<td>54 (29.8)</td>
<td>164 (24.7)</td>
<td></td>
</tr>
<tr>
<td>Moderate risk</td>
<td>46 (25.5)</td>
<td>169 (25.5)</td>
<td></td>
</tr>
<tr>
<td>High risk</td>
<td>81 (44.7)</td>
<td>329 (49.7)</td>
<td></td>
</tr>
<tr>
<td>EQ-5D-5L-VAS (score range: 0-100), mean (SD)</td>
<td>59.0 (22.1)</td>
<td>55.2 (23.0)</td>
<td>.05</td>
</tr>
<tr>
<td>Oswestry Disability Index (score range: 0-50), mean (SD)</td>
<td>30.3 (15.6)</td>
<td>31.1 (14.9)</td>
<td>.50</td>
</tr>
<tr>
<td>Anxiety (score range: 0-10), mean (SD)</td>
<td>3.8 (3.0)</td>
<td>3.8 (3.1)</td>
<td>.99</td>
</tr>
<tr>
<td>Loneliness (score range: 0-10), mean (SD)</td>
<td>1.4 (2.4)</td>
<td>1.3 (2.2)</td>
<td>.67</td>
</tr>
<tr>
<td>Catastrophization (terrible pain that will never improve; score range: 0-10), mean (SD)</td>
<td>4.8 (2.9)</td>
<td>5.0 (3.0)</td>
<td>.48</td>
</tr>
<tr>
<td>Catastrophization (the pain is overwhelming; score range: 0-10), mean (SD)</td>
<td>3.7 (3.1)</td>
<td>4.1 (3.1)</td>
<td>.24</td>
</tr>
<tr>
<td>Risk of persisting pain (score range: 0-10), mean (SD)</td>
<td>6.8 (2.6)</td>
<td>6.8 (2.6)</td>
<td>.91</td>
</tr>
<tr>
<td>Sadness (score range: 0-10), mean (SD)</td>
<td>3.5 (3.1)</td>
<td>3.6 (3.1)</td>
<td>.71</td>
</tr>
<tr>
<td>Loss of interest or joy (score range: 0-10), mean (SD)</td>
<td>4.3 (3.3)</td>
<td>4.3 (3.2)</td>
<td>.88</td>
</tr>
<tr>
<td>Fearing activity will damage the back (score range: 0-10), mean (SD)</td>
<td>3.4 (2.9)</td>
<td>3.8 (3.2)</td>
<td>.08</td>
</tr>
<tr>
<td>Fearing activity will increase the pain (score range: 0-10), mean (SD)</td>
<td>4.8 (3.2)</td>
<td>4.4 (3.3)</td>
<td>.16</td>
</tr>
</tbody>
</table>

Our results suggest that patients aged 36-65 years were more likely to be labeled as interested in mHealth solutions compared to patients between 18-35 years (OR 0.43, 95% CI 0.026-0.711) and 65 years or older (OR 0.77, 95% CI 0.525-1.153) and explained a limited proportion of variance ($R^2=0.014$). Smoker (OR 0.39, 95% CI 0.244-0.636) and the association explained a limited proportion of variance ($R^2=0.019$). Combined, the associations of age and smoking explained 3.3% of the proportion of variance.

These findings were supported by univariate regression analysis and a comparison of patients who were labeled as expressing interest in the mHealth solution to those who did not. The proportion of variance explained in the group of patients labeled as interested in mHealth solutions across the 21 selected

---

*Missing: 8 of the 21 variables had complete responses, and none of the remaining 13 variables had more than 2.5% missing responses.*
variables was 0.059, with age and smoking status accounting for 0.033 of the variance (Table 2).

Table 2. Associations to being labeled as interested and proportion of variance explained.

| Age (years)         | Odds ratio (95% CI) | SE  | Z     | People invited, N | P value (>|z|) | R^2 |
|---------------------|---------------------|-----|-------|-------------------|---------------|-----|
| 36-65 (reference)   | —                   | —   | —     | —                 | —             | —   |
| 18-35               | 0.43 (0.02-0.71)    | 0.114| —8.17 | 830               | .002          | .014 |
| >65                 | 0.77 (0.52-1.15)    | 0.151| —9.17 | 830               | .002          | .019 |
| Smoking             |                      |     |       |                   |               |     |
| Nonsmoker (reference)| —                   | —   | —     | —                 | —             | —   |
| Smoker              | 0.39 (0.24-0.64)    | 0.098| —3.72 |                   | <.001         | <.001|

Discussion

Principal Results

This study aimed to explore the proportion of patients with LBP who were labeled as interested in using an mHealth solution designed to support self-management in addition to usual care and assess how these patients differed from those who were labeled as not interested. We found that 21% of the eligible patients were labeled as interested in using the mHealth solution. The groups had no statistically significant differences except that patients labeled as interested were more frequently within the 36-65 years age range and were nonsmokers.

Comparison With Prior Work

Previous evidence of the characteristics and associations of patients with LBP and their interest in mHealth solutions is limited. Contrary to Krebs and Duncan [4], we found a nonsignificant association between BMI and no association between being younger and labeled as interested in mHealth solutions. The key differences between Krebs and Duncan [4] and this study are the target populations (general population) and the type of mHealth solutions included (fitness apps or calorie trackers). Similar to our results, Philip et al [24] identified an association between higher age and increased use of mHealth solutions among patients with chronic pain. We suggest that the differences in results between Krebs and Duncan [4], Philip et al [24], and this study were due to differences between participants from the general population and patients with LBP or chronic pain. Three recent studies have assessed the characteristics and associations of users and nonusers of different mHealth apps, all using participants from the general population, but still lacking consensus. Walrave et al [25] identified no sociodemographic differences between users and nonusers of contact tracking alert apps, including the Belgian Corona alert app. A study of the general US population identified strong associations of age, gender, and education level with the use of fitness apps and calorie counters [26]. Lim et al [27] identified that female patients with higher education were more prevalent users of mHealth apps. Although this lack of consensus regarding patient interest could indicate a call for more research, it could also reflect that the interest in mHealth solutions may be characterized by patients’ preferences and perspectives on the relevance of mHealth solutions.

Strengths and Limitations

This study benefitted from several strengths. First, we had access to comprehensive information on the patients participating through the SpineData registry. Further, we benefited from the fact that SpineData has been in routine use for several years and is frequently updated per clinician and evidence demand [14]. Thus, the PROs were collected using validated questionnaires or questions designed for the LBP population and International Classification of Function core set [14,16,28,29]. The included patients were identified using a computer algorithm, and patients were sent one invitation and one reminder invitation to be screened for eligibility. Thus, the risk of unconscious bias in the recruitment was eliminated. However, using a single data source (SpineData) also limited the variables available to investigate in the univariable model. Low education and economic status have been associated with limited use and adoption of mHealth solutions [26,30], but this information was unavailable in SpineData. Smoking is reportedly more prevalent among patients with a lower socioeconomic or sociodemographic status [31,32]. Further, the use of one registry meant we only had access to PROs, which may be affected by recall bias. The statistically significant difference between being labeled as interested in mHealth solutions by smoking status could reflect a difference in education level. Thus, education level is a parameter that could differentiate the patients labeled as interested and those labeled as uninterested in the mHealth solution, although this hypothesis remains unanswered. Patients referred to the Spine Centre usually have pain for extended periods and at a higher intensity than patients in the primary sector [33]. Thus, these patients potentially have more complex LBP issues than those with LBP who were not referred, which means that our study population may be a subgroup of the general LBP population. The terms “interested” and “uninterested” pose a challenge due to their vague nature. We recognize the distinction between demonstrating a “cursory” interest and moving toward actual participation. After extensive discussions among authors, we chose the terms “interested” and “uninterested.” Despite their less-than-optimal nature, we believe these terms best suit the context where we categorize patients based on their response to an invitation, progressing
from screening to eligibility for participation in a trial evaluating an mHealth solution supporting self-management in patients with LBP. Further, some patients might be interested in testing an mHealth solution but uninterested in participating in a trial or vice versa. Further, those labeled as uninterested in the mHealth solution in this study might see advantages in mHealth solutions that they found more relevant like how to stop smoking or lose weight [34]. This study only addresses patient characteristics; however, investigating clinicians’ perspectives on the use and adoption of mHealth solutions in LBP self-management will similarly inform on barriers to and facilitators of increased mHealth adoption in clinical practice. However, as the SpineData clinical registry only entails patient data, this perspective was not possible in this study. Thus, the results of this study should be interpreted with caution regarding generalizability, and future qualitative or mixed methods studies could explore patients’ preferences and perceptions of the relevance of mHealth solutions. Another important area of research can be clinicians’ acceptability of mHealth solutions and the need for rigorous demonstrations of safety and efficacy to alleviate any reservations or hesitance among clinicians.

Conclusion
This study aimed to explore the characteristics of patients labeled as interested or uninterested in participating in a trial testing an mHealth solution designed to improve self-management. Our study identified that 21% (n=181) of eligible patients with LBP were labeled as interested in participating in the trial testing an mHealth solution to support self-management. Overall, the patients labeled as interested and uninterested, except for age and smoking status were similar. Interestingly, patients aged 36-65 years were more frequently labeled as interested in the mHealth solution. Thus, patients aged 36-65 years may be more interested in adopting mHealth solutions. How to increase interest in mHealth solutions among younger and older patients with LBP is an important consideration for future research and developers, especially as the findings of the selfBACK trial indicate an increased effect for older patients.

Authors’ Contributions
AH, BSC, and KS conceptualized the study. AH, LFS, and NHSC acquired the data and data permissions. JAI performed the data analysis and drafted the manuscript, with support from NHSC, AH, and LFS. All authors helped draft and critically revised the manuscript for important intellectual content and approved the final version.

Conflicts of Interest
None declared.

References


19. STarTBack online. STarTBack. URL: https://startback.hfakenlee.ac.uk/training/resources/startback-online/ [accessed 2024-02-01]


34. Pan J, Dong H, Bryan-Kinns N. Perception and initial adoption of mobile health services of older adults in London: mixed methods investigation. JMRI Aging 2021 Nov 19;4(4):e30420 [FREE Full text] [doi: 10.2196/30420] [Medline: 34807836]

Abbreviations

LBP: low back pain
mHealth: mobile health
OR: odds ratio
PRO: patient-reported outcome
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

©Jonas Ammundsen Ipsen, Louise Fleng Sandal, Natalie Hong Siu Chang, Berit Schiøttz-Christensen, Karen Søgaard, Anders Hansen. Originally published in JMRI Human Factors (https://humanfactors.jmir.org), 12.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMRI Human Factors, is properly cited. The complete bibliographic information, a link to the original publication on https://humanfactors.jmir.org, as well as this copyright and license information must be included.
User-Centered Design and Usability of a Culturally Adapted Virtual Survivorship Care App for Chinese Canadian Prostate Cancer Survivors: Qualitative Descriptive Study

Karen Young\textsuperscript{1,2}, BCom, MSc; Ting Xiong\textsuperscript{1,2}, BSc, MSc; Rachel Lee\textsuperscript{1}, BSc, MHI; Ananya Tina Banerjee\textsuperscript{3}, PhD; Myles Leslie\textsuperscript{4}, PhD; Wellam Yu Ko\textsuperscript{5}, PhD; Quynh Pham\textsuperscript{1,2,6,7,8}, PhD

\textsuperscript{1}Centre for Digital Therapeutics, Techna Institute, University Health Network, Toronto, ON, Canada
\textsuperscript{2}Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
\textsuperscript{3}Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada
\textsuperscript{4}School of Public Policy, University of Calgary, Calgary, AB, Canada
\textsuperscript{5}Men’s Health Research Program, University of British Columbia, Vancouver, BC, Canada
\textsuperscript{6}Toronto General Hospital Research Institute, University Health Network, Toronto, ON, Canada
\textsuperscript{7}Telfer School of Management, University of Ottawa, Ottawa, ON, Canada
\textsuperscript{8}School of Public Health Sciences, University of Waterloo, Waterloo, ON, Canada

Corresponding Author:
Quynh Pham, PhD
Centre for Digital Therapeutics
Techna Institute
University Health Network
Toronto General Hospital, R. Fraser Elliott Building, 4th Floor
190 Elizabeth Street
Toronto, ON, M5G 2C4
Canada
Phone: 1 (416) 340 4800 ext 4765
Email: q.pham@uhn.ca

Abstract

Background: Cultural adaptations of digital health innovations are a growing field. However, digital health innovations can increase health inequities. While completing exploratory work for the cultural adaptation of the Ned Clinic virtual survivorship app, we identified structural considerations that provided a space to design digitally connected and collective care.

Objective: This study used a community-based participatory research and user-centered design process to develop a cultural adaptation of the Ned Clinic app while designing to intervene in structural inequities.

Methods: The design process included primary data collection and qualitative analysis to explore and distill design principles, an iterative design phase with a multidisciplinary team, and a final evaluation phase with participants throughout the design process as a form of member checking and validation.

Results: Participants indicated that they found the final adapted prototype to be acceptable, appropriate, and feasible for their use. The changes made to adapt the prototype were not specifically culturally Chinese. Instead, we identified ways to strengthen connections between the survivor and their providers; improve accessibility to resources; and honor participants’ desires for relationality, accountability, and care.

Conclusions: We grounded the use of user-centered design to develop a prototype design that supports the acts of caring through digital technology by identifying and designing to resist structures that create health inequities in the lives of this community of survivors. By designing for collective justice, we can provide accessible, feasible, and relational care with digital health through the application of Indigenous and Black feminist ways of being and knowing.

(JMIR Hum Factors 2024;11:e49353) doi:10.2196/49353
KEYWORDS
digital health; virtual care; digital therapeutics; prostate cancer; cancer survivorship; user-centred design; usability; supportive care; cultural adaptation; Chinese Canadians

Introduction

Digital health has been posited as a pathway to more equitable and holistic care [1,2]. However, the digital divide, or the capacity for digital technology to exacerbate inequities, has been widely described [3]. Its differential impacts on the social determinants of health are known as the digital determinants of health [4]. Recent years have seen an acceleration of digital health innovations (DHIs) such as digital therapeutics into health care systems, which was supercharged by the COVID-19 pandemic and the resulting widespread implementation of teledmedicine [4]. One such digital therapeutic is the Ned Clinic (“No Evidence of Disease”), which aims to optimize clinical care and patient self-management through virtual asynchronous care delivery for prostate cancer (PCa) survivors [5]. The Ned Clinic platforms, including clinician-led (Specialist Ned) and nurse-led (Ned Nurse) interventions, were developed at the University Health Network in Toronto, Canada, by a consortium of stakeholders [5].

PCa is the most commonly diagnosed nonskin cancer for Canadian male individuals, and most (99%) are estimated to be diagnosed in male individuals aged 50 years and older [6]. Older adults are negatively impacted by the digital divide [7]. Race, a social determinant of health, is also linked to worse survivorship and care outcomes for PCa survivors, most notably for Black male individuals [8]. Asian (generally defined as East Asian and South Asian ethnicity) male individuals have been found to have better survival rates than the median but are more likely to present with advanced PCa, suggesting systemic issues with identifying health issues and obtaining timely appropriate care [9]. These differences carry over into the delivery of follow-up care, as PCa survivors’ care needs and access to care are affected by the complex intersection of ethnicity, culture, and other social and structural factors [10,11].

Cultural adaptation is the process of applying changes to existing health interventions based on “surface” (social and behavioral characteristics) and “deep” (worldview, norms, beliefs, and values) cultural structures [12]. As these structures are known to impact beliefs about illness and well-being, the intent is to provide intervention benefits for communities that have experienced health inequities [13]. Culturally adapted DHIs appear to have been most widely reported in the field of mental health; in contrast, cultural adaptations of cancer survivorship apps have not been published, likely owing to the few interventions in this area [2,14]. The frameworks that appear to be most widely used to adapt health interventions were developed by Bernal et al [15], Resnicow et al [16] (an adaptation of the model by Bernal et al [15]), and Barrera and Castro [17].

However, these guidelines and models often use framings of cultural sensitivity and competency (eg, Resnicow et al [16] and Castro et al [18]), continuing to place the burden of change on individuals rather than addressing the upstream structural determinants of health. These framings can serve to “museumize” and problematize identity categories and culture as causes of ill-health, echoing the long-standing use of culture as a scapegoat to fault specific communities for health inequities. Moreover, defining “culture” for such adaptations can be a complex process in Canada, where culture, race, ethnicity, settler colonialism, and white supremacy (ie, the social and structural determinants of health) all create intersectional and differential lived experiences under a putatively shared identity—Canadian [19-21].

This research reports on the second and final phase of a project to design a cultural adaptation of the patient-facing Ned Clinic virtual follow-up care app for Chinese Canadian PCa survivors. In phase 1, we completed formative work distilling a set of themes relevant to survivors’ user needs for follow-up and virtual care. Following the user-centered design (UCD) framework, we describe the results of the design and formative evaluation of a culturally adapted prototype of the app.

Methods

Study Design

The overall qualitative descriptive study design was structured using the community-based participatory research (CBPR) and UCD frameworks [22-24]. This study was conducted at the University of Toronto between December 2022 and March 2023 during the COVID-19 pandemic. For communities that face barriers to care, it was found that CBPR practices such as our engagement of a key informant and invitations to community members to share their lived experiences through open-ended interviews are appropriate [1,25]. CBPR concepts were applied to meaningfully involve the community (including several authors of this study) and return the results for their benefit. Here, community represents a “symbolic totality as well as a practical multiplicity,” as the Chinese Canadian community is highly heterogeneous [26]. We view our participants as a coalition of self-identified Chinese Canadian individuals impacted by PCa survivorship to attend to their differences.

The Chinese Canadian community is an immigrant community that exists as a result of settler colonialism. In recognizing this, we redefine “immigrants” as “people with ancestral roots outside of Indigenous lands, who are beholden to Indigenous laws and epistemologies” [27]. This definition led us to apply a relational paradigm to this project and an axiology of relational accountability. It also provided a pathway to apply several multilevel Indigenous and Black feminist theorizations, guiding principles, and tools [27-29]. These included decolonial theory, Etuaptmunk (two-eyed seeing), intersectionality, and cultural safety to inform our conceptualization of digital space as intimately related to land [27,30,31]. This approach allowed us to contextualize the place-related experiences of our participants and uncover their desires for relational and culturally safe care [32]. We noted that these desires are not specifically Chinese,
and this presented an opportunity to design for relationally connected digital health.

UCD is a flexible, iterative, and evidence-based 3-step design process framework that consults, involves, and considers the needs of the end user throughout the entire project [23]. Phase 1 of this study encompasses steps 1 and 2; phase 2 encompasses steps 2 and 3. We present this study according to the Consolidated Criteria for Reporting Qualitative Research (COREQ) guidelines [33].

**Step 1: Ideation and Concept Generation**

To contextualize the potential use of this app, we sought to understand the structures that impact Chinese Canadian PCa survivors' experiences with follow-up care and virtual care. The results of this phenomenologically informed exploratory-descriptive qualitative study are described elsewhere [34]. Based on the findings of this formative research, we synthesized a list of design principles (Table 1), which we then categorized into the cultural adaptation taxonomy created by Spanhel et al [14] to systematically adapt the patient-facing prototype.

**Table 1. Summary of design principles for the adaptation of the Ned Nurse patient-facing app.**

<table>
<thead>
<tr>
<th>Research finding</th>
<th>Design principle</th>
<th>Taxonomic classification [14]a</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHI^b freedom: patients felt that they were expected to track and remember overwhelming amounts of information.</td>
<td>The system should automatically update, store, and provide access to PHI on demand.</td>
<td>Content components: ● (9) Goals of treatment ● (10) Methods of treatment</td>
</tr>
<tr>
<td>Access to personalized education and information: patients felt that they were unable to access information about their care options and disease status.</td>
<td>The system should provide access to personalized and evidence-based information regarding staging, self-management, and treatment options.</td>
<td>Content components: ● (9) Goals of treatment ● (10) Methods of treatment</td>
</tr>
<tr>
<td>Continuity of care: patients desired a connection with their provider and the ability to communicate during times of need.</td>
<td>The system should improve accessibility and continuity of care, as strong care relationships create a sense of safety.</td>
<td>Content components: ● (9) Goals of treatment ● (10) Methods of treatment</td>
</tr>
<tr>
<td>Security: patients expressed suspicion about digital health because they had concerns about surveillance and security.</td>
<td>The system should be architected and built with a high level of security and privacy.</td>
<td>Methodological components: ● (12) Functionality</td>
</tr>
<tr>
<td>Accessibility: patients wanted to access care in readable and accessible language formats.</td>
<td>The system should provide readable language and accessible language formats.</td>
<td>Content components: ● (5) Language translation ● (6) Language tailoring</td>
</tr>
<tr>
<td>Digital literacy: patients felt comfortable with their device of choice but desired simplicity, form over function, and accessible help and documentation.</td>
<td>The system should prioritize usability, provide straightforward instruction and support, and maintain simple user interface and user experience design.</td>
<td>Methodological components: ● (11) Structure ● (12) Functionality ● (13) Design and aesthetics</td>
</tr>
<tr>
<td>Care coordination: patients felt like they were expected to coordinate their care, as communication between specialists, primary care, and other services were fragmented.</td>
<td>The system should coordinate and provide a clear follow-up appointment schedule.</td>
<td>Content components: ● (9) Goals of treatment ● (10) Methods of treatment</td>
</tr>
<tr>
<td>Resources: patients felt unable to access, refused, or unaware of needed resources such as mental health support.</td>
<td>The system should provide accessible pathways to resources, such as psychological support, supportive care, and financial support.</td>
<td>Content components: ● (8) Difference in concepts of mental health and its treatment ● (9) Goals of treatment ● (10) Methods of treatment</td>
</tr>
</tbody>
</table>

^aThe design principles used to adapt the Ned Clinic patient app identified here are classified to the corresponding taxonomic components found in Spanhel et al [14].

^bPHI: personal health information.

**Step 2: Design and Development**

We applied these design principles to adapt the Ned Nurse patient app for Chinese Canadian survivors. A composite profile of a sample representative user was created to situate the design team during the development of the wireframes. A list of 5 use scenarios was created to guide the adaptation. These scenarios encompassed the design principles created in step 1 and included actions such as completing follow-up tasks, accessing a follow-up care schedule, and using the app to chat with a clinician. All use scenarios are described in the interview guide (Multimedia Appendix 1). Then, the original Ned Nurse app wireframes were redesigned to reflect the features required to operationalize these scenarios through the app, resulting in a new prototype. The prototype was created in Figma (Figma Inc.) on an iPhone 13 (Apple Inc.) interface. This initial adaptation was iteratively critiqued by a team of researchers and human factors designers to refine the content, user interface, and user experience. Once the adapted prototype was finalized, it was translated from English into written Chinese via the
translation process outlined in Haldane et al [35]. This resulted in 3 versions of the adapted prototype in English, Simplified Chinese, and Traditional Chinese. The app home page in each language version is shown in Figure 1.

Figure 1. Wireframes of the adapted Ned Nurse homepage in all 3 language versions.

Step 3: Evaluation

Overview

We empirically evaluated the acceptability, appropriateness, and feasibility of the adaptation through a cultural safety lens [22]. These dimensions are early-stage implementation outcomes and have also been found to be core to the success of DHIs [36,37]. A moderated cognitive walkthrough approach and the think-aloud protocol were used to construct a semistructured interview guide encompassing the 5 scenarios describing usual tasks that an end user might complete through the app [38].

Usability Testing

Facilitators began each test by outlining the usability testing procedure and think-aloud protocol. Context regarding the intended use and deployment of the Ned Nurse system was provided. Participants were asked to complete a series of actions for each scenario on the prototype to evaluate its design and functionality. We asked participants to think and speak about improvements they desired during their evaluations. In situations where the participant was unable to access the prototype on their device, they were asked to state their intended actions using the think-aloud protocol to the facilitator, who completed the action in the prototype on their behalf.

Interviews were completed through Microsoft Teams or Zoom (Zoom Technologies Inc). Informed consent for this work was previously obtained as part of overall study consent from participants. Participants were provided with the choice of completing their interview in Cantonese, Mandarin, or English and were also able to choose which language they wished to test the prototype in. The results of each usability test were iteratively analyzed via content analysis. Audio recordings of the participant interviews were translated into English as needed, according to the translation process described previously. A deductive and inductive content analysis approach was used, in which analysis of the data was completed by coders (TX and KY) through a process including open coding, creating categories, and abstraction [39]. Recommendations were applied in real time to create a final prototype that incorporated feedback from each user over the course of usability testing.

Positionality

An important marker of excellent qualitative research is “sincerity” or positionality, which indicates that the researcher has thought about and is reflective and aware about their values,
experiences, biases, and inclinations within their research [40]. Here, the lead researcher reports on their social position, personal experiences, and political and professional beliefs to center the active role that the researcher plays in the framing of the research problem, interpretation of data, methods used, and the reporting of the results [41].

KY is a health informatics trainee and second-generation Chinese Canadian settler who was born and raised in the Greater Vancouver Regional District (GVRD) by a working-class, first-generation immigrant family with roots in southeastern China. She does not have any direct experience with PCa and has not previously provided care for PCa survivors. KY works primarily from a relational paradigm, focusing on the structures, contexts, and relationships that shape the design, development, and implementation of digital therapeutics and health technologies. She led and participated in all study activities.

Setting and Place
This study was conducted in the GVRD, located on the current, unceded, and future territories of the Tsleil-Waututh, Squamish, and Musqueam First Nations. The GVRD is home to one of Canada’s oldest and largest living Chinese communities, including persons and families whose stories and identities span multiple geographies and generations [42]. The lead (KY) and senior author (QP) established relations with a supportive care program that provides care for Chinese Canadian PCa survivors and a Chinese PCa support group in this area. A key community informant agreed to guide this study and review and approve study materials.

Ethical Considerations
Research ethics approval for this study was obtained from the University of Toronto research ethics board (Human Protocol #43145). Written and verbal informed consent to participate in both phases of the project was obtained from all participants prior to interviews via the REDCap tool (Research Electronic Data Capture; Vanderbilt University), hosted at the University of Toronto. All data collected and disseminated here have been de-identified. Participants were provided with an honorarium of $50.00 CAD ($37.65 USD) per hour in appreciation of their time.

Results
Demographics
Usability testing was performed by 6 user testers, convenience sampled from the pool of 14 survivors and partner-caregivers who participated in the first phase of work as a form of member checking. This sample was also informed by Nielsen-Norman usability testing guidelines [43]. The reasons for nonparticipation were not collected. To protect the privacy of the participants involved in this phase, a demographic overview of the overall research project is provided here. Of the 14 participants in the first phase of this project, all survivors identified as men (n=12, 86%), and all partner-caregivers identified as women (n=2, 14%). A total of 13 (93%) participants indicated that they spoke English as an additional language. Most made an income between CAD $15,000 (US $11,048) and CAD $100,000 (US $73,653; n=12, 86%), lived in an urban area (n=13, 93%), were married (n=12, 86%), were educated beyond high school (n=13, 93%), and were retired (n=9, 64%). A 50/50 split emerged between preferences for smartphone or desktop or laptop use. Most (n=10, 71%) self-rated as being comfortable with their device. Participants indicated that they had 2 or fewer smartphone health apps (n=13, 93%).

Phase 1: Ideation
Table 1 summarizes the user requirement findings that emerged from previous formative research in phase 1 of this project and their subsequent translation to design principles.

Phase 2: Design and Development
Overview of Ned Nurse
An overview of the Ned Nurse clinical trial protocol is described by Pham et al [5]. The findings from formative work on the perspectives of health care providers, patients from the wider PCa survivor community, and the service design of the platform are forthcoming. Briefly, Ned Nurse digitally operationalizes a nurse-led model of survivorship care. Patients complete a series of tasks or access resources designed to support them in their survivorship. The platform aims to facilitate holistic care for patient quality of life.

Overview of the Adapted Patient-Facing System
The patient-facing adaptation set 2 user-input “care tasks,” a validated questionnaire (Expanded Prostate Cancer Index Composite-Clinical Practice [EPIC-CP]) and a needs assessment survey, to constitute a single Ned Nurse “review” [5,44]. Language within the app avoided wording such as appointment, visit, and so forth to clarify the differences between synchronous and asynchronous care encounters. The user interface and user experience were designed to draw the user’s attention to these tasks on the homepage immediately after login. All features were accessible via an in-app hamburger menu.

User inputs to the questionnaire were triaged via a decision-tree algorithm [45]. The algorithm was designed to return in-app self-management resources within a progress note (“Nurse’s Note”) automatically available to the user after input submission. If the algorithm detected that the patient required further support, they were prompted to specify domains for follow-up and asked to select their preferred contact method. This action would flag this patient to the nurse for follow-up. Resource links would appear on the homepage after the note was read and cleared.

To ensure that patients were aware of their review schedule, a feature was designed to display the last date, frequency, and next date of their expected reviews. The name of the nurse in charge and an explanation of their Ned Nurse role were provided to strengthen the perceived connection between the user and the nurse. This feature also set expectations for manual response times and included a link to users’ previous submissions for on-demand access.

Resources were made available in 3 separate categories: symptom self-management advice, PCa information and education, and support and programmatic resources. Within
each category, resources were further categorized. For example, symptom management included resources for symptoms such as anxiety, urinary incontinence, and hot flashes. Each resource provided an overview; relevant self-management steps; off-app links; and the ability to email, print, or save the resource. The feature home page also sectioned resources saved by the user (“Saved Resources”) and resources picked for the user (“Picked for Me”) by their nurse.

All available and historical prostate-specific antigen and testosterone blood work results were made available in chronological order to the user on-demand in a separate feature. Finally, a chat feature was designed to explore whether users might find it useful. It incorporated both responses in English from an automated support assistant (chatbot) and manually submitted by the nurse. This feature was simulated for evaluation.

**Phase 3: Evaluation**

Of the 6 participants, 2 (33%) tested in Cantonese, 3 (50%) tested in English, and 1 (17%) tested in Mandarin. These ratios correspond to testing of the Traditional Chinese, English, and Simplified Chinese versions. We note that patients who completed their testing in 1 language were functional to fluent in 1 or all of the other languages and provided critique for multiple versions.

Overall, there was strong agreement that the adaptation presented here would be acceptable, appropriate, and feasible for use, with the exception of the chat feature. Participants agreed that this app would make them feel comfortable and safe by allowing them to have more control over their care, access to resources, and stronger connections to their providers. They were encouraged by its perceived ability to meet their needs by protecting their connection with their providers, leveraging the functional flexibility of digital health, and providing resources beyond what they currently accessed. It was particularly valuable that features could be accessed at their convenience, as some felt that their follow-ups were far too short to meet their needs. Overall, 5 (83%) of 6 participants indicated that the level of support provided by this app was beneficial enough that it should be offered to patients prior to beginning treatment, or even at the point of diagnosis.

Participants’ critiques centered on expanding flexibility, access to information, and streamlining responses. They felt that responses for some assessment questions (from 4 to 8 options) were overwhelming and should be reduced (3/6, 50%). English-Chinese translations would increase self-confidence in navigating the health care system. Medication names were spotlighted as particularly difficult. This was noted as an opportunity to expand the app’s personal health information (PHI) storage, as a feature containing self-reported PHI (including medications) would be helpful to reference. Pictures and videos were desired instead of textual explanations. Laboratory results were asked to be displayed graphed or with severity indicators by 1 participant, and a text size adjustment function was requested by another.

Support for sexual dysfunction was not requested explicitly but appeared to be implied (3/6, 50%). A sexual therapy resource section was requested by 1 participant. Another noted that they would be more comfortable with nurses gendered as men as they felt uneasy when discussing sexual dysfunction with women. A final participant was keen to indicate that sexual dysfunction was a major area of concern when completing the EPIC-CP questionnaire.

As resources could be accessed on demand, some indicated that more would be beneficial. However, other participants expressed that the number displayed in the prototype were more than sufficient, reflecting our previous study findings on the bifurcated information-seeking behaviors of Chinese Canadian PCa survivors. Participants were also asked if they might find having their imaging results helpful. Although the majority (4/6, 67%) said no, those who said yes (2/6, 33%) were keen on having this information, especially if they needed to travel outside of Canada.

The questionnaire and assessment were generally deemed to be acceptable by most participants (4/6, 33%), with several notable dissents (2/6, 33%). The EPIC-CP question regarding hormonal function was highlighted as confusing by some because the connection between hormonal function and fatigue was not readily apparent. The spiritual domain in the needs assessment was flagged, as some thought that it would not be appropriately addressed by the nurse. Those who felt uncomfortable with this domain noted that they would prefer speaking about these needs to a spiritual leader. Agreement on appropriate response times also varied.

The chat function was deemed possibly helpful but likely unnecessary (4/6, 67%). As all chat interactions were in English, participants who were not confident in their English communication skills felt that their use of this feature would be limited (3/6, 50%). Others felt reminded of troubleshooting cable services rather than feeling connected to their provider. It was emphasized that any opportunity to improve connections to their providers through the app would be appreciated.

**Discussion**

**Principal Findings and Implications**

This study provides an applied example of a DHI for Chinese Canadian PCa survivors, which is based on broader principles of collectivism and relationality from Indigenous and Black feminist theory. Our initial aim was to co-design a cultural adaptation of the Ned Clinic to provide compassionate care and meet the unmet needs of Chinese Canadian PCa survivors. Participants were also asked if they might find bifurcated information-seeking behaviors of Chinese Canadian PCa survivors. Participants were also asked if they might find difunction useful but likely unnecessary (4/6, 67%). As all chat interactions were in English, participants who were not confident in their English communication skills felt that their use of this feature would be limited (3/6, 50%). Others felt reminded of troubleshooting cable services rather than feeling connected to their provider. It was emphasized that any opportunity to improve connections also varied.

The chat function was deemed possibly helpful but likely unnecessary (4/6, 67%). As all chat interactions were in English, participants who were not confident in their English communication skills felt that their use of this feature would be limited (3/6, 50%). Others felt reminded of troubleshooting cable services rather than feeling connected to their provider. It was emphasized that any opportunity to improve connections to their providers through the app would be appreciated.

**Discussion**

**Principal Findings and Implications**

This study provides an applied example of a DHI for Chinese Canadian PCa survivors, which is based on broader principles of collectivism and relationality from Indigenous and Black feminist theory. Our initial aim was to co-design a cultural adaptation of the Ned Clinic to provide compassionate care and meet the unmet needs of Chinese Canadian PCa survivors via digital health.

However, attending to cultural adaptation theory and the lived realities of settler colonialism identified gaps to interweave Indigenous and Black feminist teachings. We began by synthesizing design principles that surfaced as critical to our participants and their feelings of comfort and safety when receiving follow-up care. This allowed us to leverage digital health to strengthen relations between the survivor and their providers; improve accessibility to resources; and honor desires for relationality, accountability, and care [46,47]. Rather than adapting by defining Chinese Canadian culture, we co-designed
to intervene in structural causes of health inequities created by settler colonial culture instead [21,48].

We applied Etuaptmumk by interweaving strengths from different ways of being and knowing, including those from Indigenous, Western, Chinese, and Black feminist traditions in relation to PCa follow-up and virtual care [27,30,31]. These included prioritizing relational care, accounting for the use of prostate-specific antigen screening as a recurrence monitoring tool, and the benefits of supportive care programs to create adaptation features [30,49]. The EPIC-CP validated questionnaire is a key part of clinical follow-up care, as it allows clinicians to identify possible areas of concern during follow-up [50]. The needs assessment addresses domains beyond clinical care, reflecting the holistic nature of the medicine wheel [51]. Access to resources includes education and guidance for the self-management of concerns across multiple domains. The app presents a “care contract” in the form of a schedule that clearly states the “terms” and dates of the user’s follow-ups [52]. It also respects the user’s privacy by providing access and allowing them to share their PHI on their terms [53]. Only key inputs are communicated for triage and response. Finally, language access is built into the app as a question of communication accessibility, rather than only culture.

This design approach and these features do not deny the fact that culture is a real influence and can be a source of strength in many peoples’ lives. However, we must go beyond implicating culture when designing DHIs for communities made vulnerable and instead address the overarching and underlying structures that create health inequities. Our design approach looked “up” at these structural causes rather than looking “down” and museumizing culture for participants through cultural sensitivity and competency. We demonstrate that a structural approach that applies teachings such as cultural safety and intersectionality can result in DHIs that are found to be acceptable, appropriate, and feasible for use while still leaving room for users to self-define and practice culture on their own terms. We are supporting, not replacing, the labor and acts of caring with digital health. Beginning with a paradigm shift opened a window to design for collective care, a scalable opportunity to benefit communities beyond Chinese Canadians with this Ned Nurse patient-facing app adaptation.

Strengths and Limitations

We have created the first “cultural” adaptation of a PCa follow-up care application for Chinese Canadian survivors. We extended the accessibility of this prototype by offering it in 3 language versions and tested its validity through member checking by returning it to participants who had provided their experiences and expertise as part of the first phase of this project. The findings should be considered with some limitations. Our sample does not fully represent the Chinese Canadian PCa community, as the heterogeneity of the community makes it difficult to recruit a fully representative sample [42]. User testing did not differentiate between results derived from users who interacted with the app themselves and users who directed a facilitator to perform actions on their behalf. However, all participants received the same set of instructions to apply the think-aloud method. A broad description of our theoretical stance, setting and place, methods, and results are provided to enhance understanding. We think of and encourage the transferability of this research as to how it might be made meaningful (ie, valid) for other communities in places where they may be subject to similar constructs and patterns of oppression [32]. Finally, this study does not include the provider perspective, although Ned was developed with clinicians who provide follow-up care for patients from this community. Future studies should examine the clinician’s perspective on the design and development of similar DHIs, including provision of care through these apps, acceptability and feasibility, and implementation readiness.

Conclusions

This study demonstrates the relationality of Indigenous and Black feminist ontologies, epistemologies, and methodologies to digital health design by providing a worked example of its empirical use for an adaptation of a PCa follow-up care app, the Ned Nurse Clinic, for Chinese Canadian PCa survivors. We applied UCD principles to develop a prototype design that supports the relational act of caring through digital technology by identifying structures that create inequities in the experiences of this community of survivors and designing to intervene and provide accessible, connected care instead. We hope that this prototype serves as a tool to help regenerate places of caring, as we have learned from Indigenous and Black feminist scholars’ teachings on power, place, and digital technologies.

Acknowledgments

We are deeply grateful for the time and effort gifted to us by the community members who shared their stories. We acknowledge and appreciate the staff and members of the Prostate Cancer Supportive Care Program, including Dr C Higano and Ms M Sundar, and the Richmond Chinese Prostate Cancer Support Group (列治文華人前列腺癌支援網絡), led by Mr D Leung and advised by Dr W Yu Ko. We thank the Ned team at the Centre for Digital Therapeutics (University Health Network) for their efforts and commitment to delivering a DHI for Canadian prostate cancer survivors. We give special thanks to Laura Parente for her incisive and human-centered design guidance. The study was funded by the Canadian Institutes of Health Research (CIHR) through the CIHR Canadian Cancer Society Survivorship Grant (CCS: 706713; CIHR: 168606) and the Canada Graduate Scholarship program, as well as by AGE-WELL Networks of Centres of Excellence through the Early Professionals & Inspired Careers in AgeTech (EPIC-AT) Fellowship and Graduate Award. These agencies were not involved in the design or analysis of this study.
Data Availability
The data sets generated and analyzed during this study are available from the corresponding author upon reasonable request.

Authors' Contributions
KY, QP, ATB, ML, and WYK contributed to project conceptualization and study design. KY, RL, and TX contributed to data collection and analysis. KY prepared the first manuscript draft, with contributions from TX, WYK, ATB, ML, and QP. All authors contributed, reviewed, and approved the manuscript.

Conflicts of Interest
QP and the University Health Network (Toronto, Ontario) jointly own intellectual property rights to the Ned app. Under the respective agreements with their organizations, QP is entitled to personally benefit from any commercial use of the intellectual property.

Multimedia Appendix 1
Semistructured usability testing interview guide.

References


Abbreviations

- CBPR: community-based participatory research
- COREQ: Consolidated Criteria for Reporting Qualitative Research
- DHI: digital health innovation
- EPIC-CP: Expanded Prostate Cancer Index Composite-Clinical Practice
- GVRD: Greater Vancouver Regional District
- PCA: prostate cancer
- PHI: personal health information
- UCD: user-centered design

©Karen Young, Ting Xiong, Rachel Lee, Ananya Tina Banerjee, Myles Leslie, Wellam Yu Ko, Quynh Pham. Originally published in JMIR Human Factors (https://humanfactors.jmir.org), 01.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Human Factors, is properly
cited. The complete bibliographic information, a link to the original publication on https://humanfactors.jmir.org, as well as this copyright and license information must be included.
Effectiveness and User Perception of an In-Vehicle Voice Warning for Hypoglycemia: Development and Feasibility Trial

Caterina Bérubé¹, PhD; Vera Franziska Lehmann¹,², PhD; Martin Maritsch¹, PhD; Mathias Kraus³, Prof Dr; Stefan Feuerriegel¹, Prof Dr; Felix Wortmann¹,⁵, Prof Dr; Thomas Züger¹,²,⁶, MD; Christoph Stettler¹,², Prof Dr; Elgar Fleisch¹,⁵, Prof Dr; A Baki Kocaballi⁷, PhD; Tobias Kowatsch¹,⁵,⁸,⁹, Prof Dr

¹Centre for Digital Health Interventions, Department of Management, Technology, and Economics, ETH Zurich, Zurich, Switzerland
²Department of Diabetes, Endocrinology, Nutritional Medicine and Metabolism, Bern University Hospital, University of Bern, Bern, Switzerland
³School of Business, Economics and Society, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nürnberg, Germany
⁴School of Management, Ludwig-Maximilians-Universität München, Munich, Germany
⁵Centre for Digital Health Interventions, Institute of Technology Management, University of St Gallen, St Gallen, Switzerland
⁶Department of Endocrinology and Metabolic Diseases, Kantonsspital Olten, Olten, Switzerland
⁷School of Computer Science, University of Technology Sydney, Sydney, Australia
⁸Institute for Implementation Science in Health Care, University of Zurich, Zurich, Switzerland
⁹School of Medicine, University of St Gallen, St Gallen, Switzerland

Corresponding Author:
Tobias Kowatsch, Prof Dr
Centre for Digital Health Interventions
Department of Management, Technology, and Economics
ETH Zurich
WEV G 216
Weinbergstrasse 56/58
Zurich, 8092
Switzerland
Phone: 41 44 632 94 88
Email: tkowatsch@ethz.ch

Abstract

Background: Hypoglycemia is a frequent and acute complication in type 1 diabetes mellitus (T1DM) and is associated with a higher risk of car mishaps. Currently, hypoglycemia can be detected and signaled through flash glucose monitoring or continuous glucose monitoring devices, which require manual and visual interaction, thereby removing the focus of attention from the driving task. Hypoglycemia causes a decrease in attention, thereby challenging the safety of using such devices behind the wheel. Here, we present an investigation of a hands-free technology—a voice warning that can potentially be delivered via an in-vehicle voice assistant.

Objective: This study aims to investigate the feasibility of an in-vehicle voice warning for hypoglycemia, evaluating both its effectiveness and user perception.

Methods: We designed a voice warning and evaluated it in 3 studies. In all studies, participants received a voice warning while driving. Study 0 (n=10) assessed the feasibility of using a voice warning with healthy participants driving in a simulator. Study 1 (n=18) assessed the voice warning in participants with T1DM. Study 2 (n=20) assessed the voice warning in participants with T1DM undergoing hypoglycemia while driving in a real car. We measured participants’ self-reported perception of the voice warning (with a user experience scale in study 0 and with acceptance, alliance, and trust scales in studies 1 and 2) and compliance behavior (whether they stopped the car and reaction time). In addition, we assessed technology affinity and collected the participants’ verbal feedback.

Results: Technology affinity was similar across studies and approximately 70% of the maximal value. Perception measure of the voice warning was approximately 62% to 78% in the simulated driving and 34% to 56% in real-world driving. Perception correlated with technology affinity on specific constructs (eg, Affinity for Technology Interaction score and intention to use, optimism and performance expectancy, behavioral intention, Session Alliance Inventory score, innovativeness and hedonic motivation, and negative correlations between discomfort and behavioral intention and discomfort and competence trust; all
Introduction

Background

Type 1 diabetes mellitus (T1DM) is a chronic condition caused by an inability of the pancreas to produce insulin and requires lifelong insulin therapy [1]. Hypoglycemia, also known as low blood glucose, is a frequent and acute complication in patients with T1DM [2,3]. Symptoms range from autonomic reactions such as trembling, anxiety, and hunger (ie, mild hypoglycemia) to neuroglycopenic reactions such as vision impairment, weakness, or cognitive impairments (ie, severe hypoglycemia) [2,4-6]. Hypoglycemia is a major issue in the context of driving: research has shown that hypoglycemia is associated with a higher risk of car mishaps [7-9]. In fact, drivers experiencing hypoglycemia are recommended by the local authorities [10] to stop the car and treat their condition. However, drivers do not always comply with these recommendations [11,12]. Thus, to help reduce hypoglycemia-related car accidents, there should be an effective warning that informs the driver about an upcoming hypoglycemic episode and supports the driver in coping with the situation. Currently, hypoglycemia can be detected and signaled through flash glucose monitoring (FGM) or continuous glucose monitoring (CGM) devices (ie, wearable receivers connected to a sensor inserted in the subcutaneous tissue of the arm or abdomen) [13]. These allow for glucose monitoring by displaying the values either continuously (ie, CGM) or upon active retrieval (ie, FGM) and deliver alerts in the form of a tone or vibration in case of out-of-range values. However, these devices present limitations in the context of driving. For instance, FGM needs to be held close to the sensor to transfer the data from the subcutaneous sensors to the monitoring device, that is, the driver needs to actively engage in a manual gesture to access the glucose value and to look at a visual display moving the focus of attention from the driving task. In contrast, allowing the drivers to receive an alert in a hands-free mode will facilitate warning reception [14] and lower worry associated with driving with T1DM [15]. However, hypoglycemia is known to cause a decrease in attention [2,4-6], thereby challenging the effectiveness of such devices. As 90% of road accidents are caused by human error, the European Commission has set new safety technologies as mandatory equipment for vehicles as of 2022 (eg, driver drowsiness and distraction warnings and speed assistance) [16]. In-vehicle warning systems for impaired driver states, such as fatigue [17], distraction [18], and breath alcohol concentration [19], are increasingly being developed. However, to the best of our knowledge, there is no existing implementation for hypoglycemia. Such technology would be aligned with the “healing car” concept [20], where vehicles become environments promoting well-being for passengers, including ergonomic seats, ambient lighting, relaxation exercises [21], and detection of health-critical states [22]. This concept is still in its early stages, but it may become a standard in car manufacturing in the future. So far, the only attempts of in-vehicle glucose monitoring are either only proof of concept without user validation [23] or conceptual work [24]. However, the online community clearly expressed a need for in-vehicle glucose monitoring and warning [25].

A growing number of automotive companies are introducing voice assistance technology into their products [26,27]. Voice assistants add value not only for the associated consumer experience but also for their greater safety. Indeed, vocal interactions have been observed to be the least cognitively demanding while driving compared with visual and haptic interactions [28,29]. Moreover, voice assistants are increasingly being implemented to deliver digital health interventions [30-33]. Although research is still in its infancy, efforts have been made to develop voice-based conversational agents to monitor and support individuals with chronic diseases such as cancer, cardiovascular diseases, cognitive disorders, or diabetes [30]. Other recent examples include prevention of excessive alcohol consumption [34], health education and monitoring, physical and mental exercise, and nutrition [35]. Furthermore, a voice assistant delivering a warning is a form of proactive behavior initiated by the computer rather than the user [36,37]. In-vehicle voice assistants can provide personalized and adaptive suggestions, but users may ignore proactive behavior if it is inopportune, violates privacy, or distracts from driving [38-40]. However, emergencies are the most suitable context for proactive behavior that violates privacy [39].

Objectives

Therefore, we investigate the feasibility of an in-vehicle voice warning delivered by a built-in voice assistant to alert and support drivers with T1DM during hypoglycemia. To the best of our knowledge, there have been no investigations on safe and effective in-vehicle hypoglycemia warnings to support drivers with T1DM or on the perception of such technology. Thus, we sought to answer the following research questions (RQs):

**Conclusions:** This is the first study to investigate the feasibility of an in-vehicle voice warning for hypoglycemia. Drivers find such an implementation useful and effective in a simulated environment, but improvements are needed in the real-world driving context. This study is a kickoff for the use of in-vehicle voice assistants for digital health interventions.

**KEYWORDS**

hypoglycemia; type-1 diabetes mellitus; in-vehicle voice assistant; voice interface; voice warning; digital health intervention; mobile phone

(CJMIR Hum Factors 2024;11:e42823) doi:10.2196/42823

P<.05. Compliance was 100% in all studies, whereas reaction time was higher in study 1 (mean 23, SD 5.2 seconds) than in study 0 (mean 12.6, SD 5.7 seconds) and study 2 (mean 14.6, SD 4.3 seconds). Finally, verbal feedback showed that the participants preferred the voice warning to be less verbose and interactive.
• RQ1: How do drivers perceive an in-vehicle voice warning for hypoglycemia while driving?
• RQ2: How effective is an in-vehicle voice warning in prompting drivers to cope with hypoglycemia?

RQ1 refers to the attitude of drivers toward the warning, whereas RQ2 refers to the driver’s compliance behavior once the warning is delivered. Answering these RQs will allow us to conclude on the feasibility of an in-vehicle voice warning for hypoglycemia. To control for individual factors influencing the perception of the warning [41], we also assessed technology affinity.

Methods

Study 0: Preliminary Assessment With Healthy Individuals in Simulated Driving

Driving Setting

Participants performed the task in a driving simulator (Carnetsoft Inc) with 3 monitors displaying the front, left, and right views. The central monitor also showed the cockpit and navigation arrows. The participants used a steering wheel and pedals (Logitech Driving Force G29) to control the simulator, which was set to automatic (ie, no clutch or gear shifter). The simulator’s computer was connected to a stereo speaker with a subwoofer, which was kept at a constant volume. To control for driving difficulty, 3 environments were used: highway, countryside, and town, with the first and last being the least and most difficult, respectively.

In-Vehicle Voice Warning Simulation

Before testing a hypoglycemia voice warning with people with T1DM, we tested the concept of a car voice assistant as an interface between a dedicated monitoring system and the user with healthy participants. As the participants were not affected by hypoglycemia, the first version of the warning was a simulated low fuel warning (“The car needs a refill. Please pull over and turn off the engine”). Although not health related, it signaled an event of reasonable urgency that required safely stopping the car. Note that the participants were informed that this message aimed to ask them to stop the car as soon as possible and that they did not need to look for a gas station.

The warning was simulated using the Wizard-of-Oz method, where the conversational turns produced by the voice assistant were played by the experimenter [42] from a laptop using predefined keyboard keys. The turns were based on the Google Cloud text-to-speech engine, with a de-DE-Wavenet-C voice, a speed of 1.11 times the normal native speed of the specific voice, and a pitch of −1.20 semitones from the original pitch. The experimenter’s computer was connected to the same sound system as the driving simulators so that the voice warning could be heard as part of the driving simulation. No visuals were included.

Voice Warning Evaluation Measures

To assess the RQs, we assessed participants’ perception of the warning (self-reported through the modular evaluation of key Components of User Experience [meCUE]; 10 constructs evaluated on a 7-point Likert scale and a general evaluation evaluated on a 10-point scale [43,44]) and participant compliance with the warning (measured by the experimenter manually assessing if the participant would pull over and stop the car following the warning, and reaction time in seconds from the timestamp of the warning to the timestamp of the car fully stopped). As the perception of technology can be influenced by technology-related personality [45], we also measured technology affinity (measured by the Affinity for Technology Interaction [ATI], a 6-point Likert scale [46]). Finally, qualitative feedback was collected informally.

Evaluation Procedure

The participants were welcomed, informed about the procedure, and invited to sit in the simulator. The voice assistant introduced itself and invited the participants to familiarize themselves with the setting, including the 3 environments. The training also screened for motion sickness.

In the experimental session, participants drove 12 times, with 4 blocks of 3 drives each, for approximately 5 minutes per drive. The driving environment’s order and starting point varied to minimize habituation. The drive began when the voice assistant prompted participants to start the engine. A timer started to deliver the low fuel warning at either 100 or 200 seconds to add variation and minimize habituation effects. At the end of the session, participants completed the meCUE.

Data Analysis

Participants were characterized by sex, age, and driver’s license duration. The ATI was aggregated as a whole, and meCUE items were aggregated per construct. All reports were aggregated across the sample, with mean and SD. Compliance was coded as binary (0=not compliant, 1=compliant) and reported in terms of frequency. Reaction time was aggregated in seconds across participants and phases, with mean and SD.

Study 1: Assessment With Individuals With T1DM in Simulated Driving

Following the iterative approach described earlier, we conducted 3 exploratory iterations. This study was part of a clinical trial registered at ClinicalTrials.gov (NCT04035993).

Driving Setting

The driving setting was the same as in study 0.

In-Vehicle Voice Warning Simulation

On the basis of the results of study 0, we adapted the warning to hypoglycemia instead of low fuel, using the fewest conversational turns possible [47]. To ensure that the drivers were available, the voice assistant started with a receptivity check: “May I disturb you?”

We designed the warning based on the guidelines of the Swiss Diabetes Association [10], which recommends taking carbohydrates and stopping the car as soon as signs of hypoglycemia are noticed. To give the driver a sense of autonomy [48], we designed the warning to suggest eating carbohydrates rather than directly engaging in stopping the car. However, if the driver did not have carbohydrates, they were asked to pull over. On the basis of the feedback, we enhanced...
the voice warning used in the following study to recommend pulling over directly (detailed conversation flow is available in Multimedia Appendix 1).

As in study 0, the warning was simulated with a Wizard-of-Oz method [42], and the turns were generated by recording the same voice. However, to reduce fatigue and cognitive load, we decreased the speed and pitch to 0.93 times the normal speed and −4.8 semitones from the original pitch, respectively. As in study 0, the experimenter would play the turns from a Microsoft Windows laptop using predefined keyboard keys to play prerecorded voice sounds. However, in study 1, the laptop program included a visualization mirrored on a smartphone. The visuals consisted of a blue circle that gradually faded in and out when the voice assistant was speaking. As in study 0, the experimenter’s computer was connected to the same sound system as the driving simulators, so that the voice assistant could be heard as part of the driving simulation.

**Voice Warning Evaluation Measures**

Perception assessment focused on evaluating the voice assistant as a trustworthy driving companion. Specifically, participants completed the Acceptance and Use of Technology (AUT) questionnaire [49,50], the Session Alliance Inventory (SAI) [51], and the Emotional Trust and Competence Trust subscales (henceforth Trust) of the Trust and Adoption questionnaire [52].

To assess technology affinity, participants completed the streamlined scale of the Technology Readiness Index (TRI 2.0) [53]. Items were rated on a 5-point Likert scale (ie, 1= totally disagree, 5=totally agree). We also added a question on whether the participants had previous experience with in-vehicle voice assistants (ie, “Have you already had experience with in-vehicle voice assistants?” with a yes or no answer).

Finally, to obtain qualitative and more in-depth feedback for improvement, we conducted a semistructured interview about their experience with the warning (the interview questions are provided in Multimedia Appendix 2).

**Evaluation Procedure**

The procedure was the same as in study 0, except that participants drove only once for 5 minutes (the evaluation procedure is detailed in Multimedia Appendix 3). Before driving, we ensured that the participants had normal blood glucose levels (5-8 mmol/L).

**Data Analysis**

The sample of participants was characterized by sex, mean age, and mean duration of their driver’s license before the study. TRI and SAI were aggregated as a whole, and the AUT and Trust items were aggregated per construct. Scores from the negatively formulated questionnaire items were inverted. Previous experience with an in-vehicle voice assistant was reported in terms of frequency. All these reports were aggregated across participants of each iteration, with mean and SD.

To further explain results in perception, they were associated with technology affinity measures. The difference in perception between participants with and without experience with an in-vehicle voice assistant was tested using a 2-sided \( t \) test, and it was correlated with the TRI constructs using a Pearson test.

Compliance was defined as whether the participant would comply with the warning and was coded as binary (0=did not comply, 1=complied). Reaction time was aggregated in seconds with mean and SD. Compliance behavior was aggregated across participants of each iteration.

Feedback was summarized in positive and negative topics, with a focus on the most prominent suggestions for improvement. Feedback was aggregated across participants of each iteration.

**Study 2: Assessment With Individuals With T1DM in Real-World Driving Undergoing Hypoglycemia**

Following the iterative approach described earlier, we conducted 2 exploratory iterations. This study was part of a clinical trial registered at ClinicalTrials.gov (NCT04569630).

**Driving Setting**

Participants drove in Volkswagen Touran on a closed circuit accompanied by a driving instructor. Dual pedals allowed the driving instructor to intervene and stop the car if necessary. The driving environments on the test track corresponded to the environments of the driving simulator used in the previous studies. Straight paths, turns, crossroads, stop signs, and a pedestrian crossing with a doll were used to implement the highway, countryside, and town scenarios. Artificial obstacles (eg, boxes and lines of traffic pylons) were used to simulate the traffic.

**In-Vehicle Voice Warning Simulation**

On the basis of the participant feedback from study 1, we revised the voice warning and addressed low trust ratings by explaining the cause of the warning. We simulated driving behavior as a trigger to detect hypoglycemia while driving, as in the study by Lehmann et al [54]. We created 2 variations of the simplified hypoglycemia notification—one with a statement of the cause (driving behavior) and one without. The final recommendation was reformulated as stricter but less directive than that in study 1.

In the second iteration, we simplified the conversational flow by removing the receptivity check (“May I disturb you?”) and the final recommendation (Multimedia Appendix 1 provides the conversation flow).

We used the Wizard-of-Oz method to simulate the warning, as in studies 0 and 1. We implemented the voice assistant in a smartphone with the same voice as in study 1. However, the experimenter had to control it remotely (outside the car), so we implemented the interaction in a smartphone app controlled by a remote desktop application. The experimenter used the smartphone screen to control the voice warning delivery; therefore, no visualization was included. Because of network-related slowdowns in the remote control, we used a combination of remote control and speech-to-text programing.

**Voice Warning Evaluation Measures**

All measures were the same as in study 1. Reaction time was calculated from the warning onset until the car reached a velocity of 0. In addition, at the end of the experiment, we
included a questionnaire item asking which of the 2 types of warning they preferred, that is, the warning including a statement of the cause that triggered the warning or the one without it, or if they would not use either of them.

**Evaluation Procedure**

After welcoming participants and explaining the procedure and simulated voice assistant, the voice assistant introduced itself as an in-vehicle assistant to support drivers with hypoglycemia. The participants then completed a training drive.

The warning was delivered at different stages of hypoglycemia (see the study by Lehmann et al [54]). Drive blocks were defined based on blood glucose levels. In the first phase, the participants drove at normal glucose (5-8 mmol/L). In the second phase, blood glucose level was progressively lowered below the moderate hypoglycemia threshold (3.0 mmol/L) to a target range of 2.0 to 2.5 mmol/L. In the third phase, moderate hypoglycemia was maintained. In the fourth phase, participants drove again with normal blood glucose levels (Multimedia Appendix 3).

To explore the effect of blood glucose level on warning perception and compliance, we delivered a warning at the end of the last drive of each phase. Participants received 2 warnings with an explanation and 2 without, in randomized order.

**Data Analysis**

Data analysis was carried out as in study 1.

**Ethical Considerations**

Study 0 was approved by the Ethics Board of ETH Zürich, Switzerland (2019-N-32), and study 1 and study 2 were approved within the context of the HEADWIND study by the cantonal ethics commission of Bern, Switzerland (2020-00685 and 2021-02381, respectively). Study 1 and study 2 are available at ClinicalTrials.gov (NCT04035993 and NCT04569630, respectively). All participants provided written informed consent.

**Results**

**Study 0: Preliminary Assessment With Healthy Individuals in Simulated Driving**

Results are summarized in Figure 1.

**Recruitment and Participants**

We recruited 11 healthy individuals with a valid driver’s license via a web advertisement (ie, University of Zurich marketplace). One participant was excluded owing to simulator sickness. Thus, we included 10 participants (n=4, 40% female; n=6, 60% male) with an average age of 30.4 (SD 7.8; range 23-47) years and holding a license for 11 (SD 7.5; range 2-26) years, on average.

**Technology Affinity Measure**

Participants showed a mean ATI of 4.2 (SD 1; Cronbach α=.91), which is 70% of the maximal value.

**Perception Measure**

The meCUE (Cronbach α=.7) revealed a mean overall evaluation of 6.4 (SD 1.6), which is 64% of the maximal value. Moreover, the highest mean values were achieved for usability (mean 6.2, SD 0.6, 89%) and usefulness (mean 5.6, SD 0.9, 80%), whereas lower values were observed for commitment (mean 1.5, SD 0.4, 21%), positive emotions (mean 2.7, SD 1.1, 39%), negative emotions (mean 2.7, SD 1, 39%), intention to use (mean 3.1, SD 1.1, 44%), and product loyalty (mean 2.6, SD 0.7, 37%). A low value for negative emotions reflects a more positive evaluation.

To explain the perception results with the technology affinity measure, we correlated each meCUE construct with ATI. We observed a correlation between ATI and intention to use (ρ=0.70; P=.02). All the other correlations were not significant at the .05 level.

**Compliance Measure**

All the participants complied with the warning and stopped the car. Participants took 12.6 (SD 5.7) seconds on average.
Qualitative Feedback

Finally, some participants reported that the voice assistant spoke too fast to deliver information during a driving task without being distracting.

Study 1: Assessment With Individuals With T1DM in Simulated Driving

Results are summarized in Figures 2 and 3.

Figure 2. Violin plots of (A) count of previous experience, (B) score values across the constructs of Technology Readiness Index (TRI; min=1, max=5), (C) score values across the constructs of Acceptance and Use of Technology (AUT; min=1, max=5), (D) Session Alliance Inventory (SAI) scores (min=1, max=5), (E) Trust scores (min=1, max=5), and (F) reaction time across iterations in study 1 (n=18). The dots represent the group means; the dashed line represents the overall mean within an iteration. RT: reaction time; sec: seconds.

Figure 3. Thematic summary of participants' feedback in study 1 (n=18).
Recruitment and Participants

We recruited 20 patients with T1DM from the Department of Diabetes, Endocrinology, Nutritional Medicine, and Metabolism at the Bern University Hospital. Participants needed functional insulin treatment, good insulin self-management knowledge, a driver’s license, and active driving in the past 6 months. We excluded one participant owing to simulator sickness and one participant owing to technical errors in the warning delivery. This resulted in a total of 18 participants (n=6, 33% female and n=12, 67% male; mean age 31.4, SD 7, range 24-44 years; mean driver’s license age 13, SD 7.5, range 4.5-28.6 years). The first iteration had 9 participants, the second iteration had 7, and the third iteration had 2 participants. Although the last iteration’s sample size was small, it provided useful feedback to improve the warning for study 2.

Technology Affinity Measure

Seven participants had previous experience with an in-vehicle voice assistant (n=2, 28% in the first iteration; n=3, 43% in the second iteration, and n=2, 28% in the third iteration). TRI was 3.4 (SD 0.6, Cronbach \( \alpha \approx .85 \)), or 68% of the maximum. Specifically, TRI was 3.5 in the first and second iterations (SD 0.6 and 0.7, respectively) and 2.7 in the third iteration (SD 0.9).

Perception Measure

Perception averaged 3.9 out of 5 (SD 0.8, 78%) and remained stable across iterations. Average AUT (Cronbach \( \alpha \approx .81 \)) values were 3.8 (SD 1) in the first iteration, 4 (SD 0.7) in the second iteration, and 3.8 (SD 0.8) in the third iteration. Effort expectancy and facilitating conditions had the highest values across all iterations, whereas behavioral intention always had the lowest values.

SAI (Cronbach \( \alpha \approx .79 \)) averaged 3.7 out of 5 (SD 0.5, 74%) and increased slightly over the iterations, from 3.4 (SD 0.7) in the first iteration to 3.6 (SD 0.6) in the second iteration to 4.1 (SD 0.1) in the third iteration.

Trust (Cronbach \( \alpha \approx .8 \)) averaged 3.1 out of 5 (SD 0.7, 62%), was stable across constructs, and had the lowest values of the 3 perception measures. Trust averaged 3.1 (SD 0.8) in the first iteration, 2.9 (SD 0.6) in the second iteration, and 3.3 (SD 0.6) in the third iteration.

To explain the perception results with technology affinity, we tested the difference in perception (AUT, SAI, and Trust) between participants with and without previous experience with in-vehicle voice assistants. The means of all constructs, excluding facilitating conditions, were slightly higher for participants with previous experience. However, a 2-sided \( t \) test revealed no significant result (ie, \( P > .05 \)).

We also correlated perception with TRI and observed a correlation between the optimism construct and performance expectancy (\( r = 0.49; P = .04 \)), behavioral intention (\( r = 0.52; P = .03 \)), and SAI (\( r = 0.57; P = .01 \)). All the other correlations were not significant (\( P > .05 \)).

Compliance Measure

All the participants complied with the warning. In the first iteration, all participants answered yes or no to the receptivity check (“May I disturb you?”) and when asked if they had carbohydrates on hand. Five of the 9 participants answered yes to the latter question, although they did not. Two of those 9 participants stopped the car although they were not explicitly advised to do so. In the second iteration, all participants answered the prompts with yes and stopped the car as advised. One participant gave an affirmative mmhm when asked, “May I disturb you?” during the hypoglycemic phase but were otherwise compliant. Because we used the Wizard-of-Oz method, the experimenter interpreted the affirmation. However, a current voice assistant might have interpreted it as an error. In the third iteration, both participants answered the prompts with yes and stopped the car. Across iterations, compliance took approximately 22 seconds. In particular, compliance took approximately 20 (mean 20.7, SD 6.2) seconds in the first iteration, approximately 17 (mean 16.7, SD 1.2) seconds in the second iteration, and approximately 31 (mean 31.7, SD 10.6) seconds in the third iteration.

Qualitative Feedback

Participants judged the voice warning as pleasant, simple, and as clear and efficient (n=15, n=11, and n=13, respectively). The topics for improvement are summarized in Figure 3. Note that these results are best understood when compared with Multimedia Appendix 1.

Given that Trust showed the lowest values in the first iteration, in comparison with the other perception measures, we decided to specifically ask participants, in our second and third iterations, what would help them trust the warning more. Of the 9 participants included in both the second and third iterations, 5 (55%) said they would just need to have a prolonged experience with the warning, whereas 3 (33%) said they would need to know what kind of data is used to infer that the driver is about to experience hypoglycemia. One participant did not know what would improve their trust.

Study 2: Assessment With Individuals With T1DM in Real-World Driving Undergoing Hypoglycemia

Results are summarized in Figures 4 and 5.
Figure 4. Violin plots of (A) count of previous experience, (B) score values across the constructs of Technology Readiness Index (TRI; min=1, max=5), (C) score values across the constructs of Acceptance and Use of Technology (AUT; min=1, max=5), (D) Session Alliance Inventory (SAI) scores (min=1, max=5), (E) Trust scores (min=1, max=5), and (F) reaction time across iterations in study 2 (n=20). The dots represent the group means; the dashed line represents the overall mean within an iteration. RT: reaction time; sec: seconds.

Figure 5. Thematic summary of participants' feedback in study 2 (n=20).
Recruitment and Participants

The recruitment procedure was the same as in study 2. We recruited 21 individuals, and 1 participant was excluded owing to data loss. Thus, we included 20 participants (n=3, 15% female and n=17, 85% male; mean age 40.9, SD 10.6, range 23-57 years; and holding a license on average since 23.7, SD 11.1, range 3.1-42.4 years). The first iteration included a sample of 9 participants and the second iteration included a sample of 11 participants.

Technology Affinity Measure

The pretest measurements revealed that 25% (5/20) of the participants had previous experience with an in-vehicle voice assistant (2 in the first iteration, and 3 in the second iteration), whereas TRI was on average 3.4 (SD 0.7; Cronbach α=.44), which is 68% of the maximal value. In particular, TRI was 3.4 in the first iteration (SD 0.8), and 3.3 in the second iteration (SD 0.7).

Perception Measure

The overall perception score was 1.7 out of 5 (SD 1.3, 34%). The results showed a slight increase in mean AUT (Cronbach α=.95) and Trust (Cronbach α=.85) values between the first and the second iteration, whereas SAI (Cronbach α=.80) showed a slight decrease. AUT also showed a considerable increase in SD. In particular, AUT values were on average 1.4 (SD 1) in the first iteration, and 1.9 (SD 1.6) in the second iteration; SAI was overall 2.8 out of 5 (SD 0.8, 56%). Values were on average 3.1 (SD 0.7) in the first iteration and 2.5 (SD 0.9) in the second iteration; Trust values were on average 1.4 (SD 0.8) in the first iteration, and 2.3 (SD 1.1) in the second iteration.

Similar to study 1, to explain the perception results with the technology affinity measure, we tested the difference in perception (ie, AUT, SAI, and Trust) among participants who had previous experience with in-vehicle voice assistants and those who did not. The means of all perception measures, excluding SAI, were consistently slightly higher in the second iteration. The means of all constructs, excluding SAI, were consistently slightly higher for participants who had previous experience with in-vehicle voice assistants. However, a 2-sided t test revealed no significant result (ie, P>.05). When correlating each perception measure with TRI, we observed a correlation between innovativeness and hedonic motivation (ρ=0.52; P=.02), a negative correlation between discomfort and behavioral intention (ρ=−0.46; P=.04), and a negative correlation between discomfort and competence trust (ρ=−0.45; P=.05). All the other correlations were not significant (P>.05).

Compliance Measure

All the participants complied with the warning. Two drives were excluded: one participant stopped once before the warning was delivered and data from one drive of one participant was lost. The results showed that the reaction time does not seem to vary across glycemic phases and, although minimal, there is a tendency for the reaction time to increase in the second iteration. Participants took 13.6 (SD 4.5) seconds in the first iteration and 15.5 (SD 4.1) seconds in the second iteration.

Preference for the Disclosure of the Triggering Cause

One participant was excluded because of data loss. The results showed that although 10 participants preferred when the warning was delivered with an explanation for the warning being triggered (in this case, driving behavior), 8 participants preferred it without the explanation. One participant stated that they would not use this in-vehicle voice warning either way.

Qualitative Feedback

In general, and similar to study 1, the participants found the communication style pleasant and efficient (n=4 and n=5, respectively). The topics for improvement are summarized in Figure 5. Note that these results are best understood when compared with Multimedia Appendix 1.

Discussion

Principal Findings

Most participants had not previously used an in-vehicle voice assistant, and technology affinity was similar across studies. In general, the voice warning elicited a positive perception, although the perception values were lower in the real-car study. In addition, participants complied with the warning in all studies, and reaction times were shorter in the real-car study than in the simulator study. Finally, the participants preferred the voice warning to be less verbose and prompt fewer interactions with the driver.

Technology Affinity

Although we did not observe a significant effect on the perception of the warning, we suspect that the participants may have experienced a double novelty: using a voice assistant while driving and experiencing a warning from an in-vehicle voice assistant. Thus, future research should include a more balanced sample and compare the perception of a voice assistant–based warning with a standard warning (eg, an acoustic tone). Moreover, although we cannot directly compare ATI (used in study 0) with TRI (used in study 1 and study 2), we can observe that technology affinity was similar across studies. Although ATI showed a mean of 4.2 over 7 (60%), TRI showed a mean of 3.4 over 5 both in study 1 and study 2 (68%). The change in technology affinity measure was the result of an internal discussion between the coauthors, and we recommend the scientific community to use TRI in future research, as it is more widely used and focuses not only on the interaction but also on the general attitudes toward new technologies.

Perception

We observed that AUT, SAI, and Trust values were higher in study 1 (simulated driving) than in study 2 (real-world driving). This evaluation might have been influenced by the driving setting. There can be 2 possible reasons. First, participants may have found the warning to be more distracting in the real car than in the simulator. However, research shows that drivers are more in control in real-world driving than in simulated driving [55]. Second, the technical difficulties in controlling the driver-assistant interaction owing to network slowdowns might have affected the user experience, and thus the perception measures. Future Wizard-of-Oz studies may account for this
methodological weakness with a more accurate text-to-speech technology, avoiding remote control, and reducing interactions. In addition, TRI seemed to have influenced behavioral intention (AUT) but did not consistently influence the other perception measures (ie, other constructs of AUT, SAI, and Trust). Thus, participants may have been excited about the potential of the voice warning, but they may not have been happy with the actual experience of using it.

Compliance
The reaction times were short enough to ensure a timely reaction to the critical event. Blood glucose can change with a maximum rate of 0.22 mmol/L/min [56]. This means that someone driving with a normal glucose of 5.5 mmol/L might reach hypoglycemia (ie, 3.9 mmol/L) within a minimum of 7.5 minutes. Thus, although experiencing hypoglycemia while driving does not require an abrupt stop but rather a careful pullover maneuver and treating the condition, measuring reaction time provided an insight into the time required to take the first measure (ie, pullover). Interestingly, the reaction time was shorter in the real car (study 2) than in the simulator (study 0 and study 1). This difference may be attributed to the lack of traffic in study 2, which allowed the driver to pull over faster.

Feedback
Although we aimed to keep the warning conversational, participants preferred a more direct notification of the problem without specific recommendations (eg, recommending waiting until the blood glucose is at its normal level) or polite formulations (eg, asking for permission to talk). To the best of the author’s knowledge, there was no in-vehicle voice warning at the time of the study, and we mostly relied on the guidelines of the Swiss Diabetes Association [10], while keeping the conversation as simple as possible. The participants’ feedback allowed us to improve the warning in this direction.

Implications and Future Directions

Hypoglycemia Warnings
Reportedly, no research has been conducted for in-vehicle applications providing a hypoglycemia warning. However, smartphone apps for hypoglycemic events tracking have been investigated [57]. Although most of the research on glucose monitoring solutions conducted so far focused on diary apps rather than warning delivery, a pilot study on a smartphone-based hypoglycemia warning showed an improved hypoglycemia awareness and a reduction in daytime hypoglycemia [58] (other research is still in the phase of validation [59]). Future research should investigate such outcomes with an in-vehicle extension of this type of application.

In-Vehicle Warnings
Although there seems to be no related work testing the voice assistant of a private vehicle to deliver hypoglycemia warnings, there is a need for “driver-friendly” in-vehicle glucose monitoring solutions, expressed by the online community [25]. In particular, drivers with T1DM have contributed to the Nightscout Foundation [60], a nonprofit organization founded in 2014 and supporting open source technology for T1DM management, with the development of a data-sharing app, able to connect a car to a CGM, and display the glucose trends while driving on the dashboard of the private vehicle [25]. Moreover, there has been conceptual work manifesting the need for collaboration between automotive and medical industries to improve the safety of drivers with T1DM [24]. However, this work has not been followed by any implementation. Furthermore, no testing with the actual users has been conducted. Our work provides preliminary evidence, both in a simulated and a real-world environment.

Needless to say, recognizing hypoglycemia is only one part of glucose monitoring while driving; general imbalance of blood glucose (including hyperglycemia) can be problematic for the driver, if not dangerous [61]. Our work can be extended to hyperglycemia and, therefore, support further the safety of drivers with T1DM.

Finally, using the in-vehicle voice assistant to deliver a warning is compatible with current technology: not only are cars increasingly equipped with voice assistants [26,27] but also the automotive industry is aware of the relevance of using the upcoming “in-car proactivity” [62].

Warning Escalation
Our results showed 100% compliance in all 3 studies. This can only mean that the warning was clear enough for the participant to understand that it was time to pull over. That is, as all studies were run in a controlled setting, where an experimental team was present, and the participant knew they would be recommended to pull over eventually, we can safely assume that the experiments experienced a participant bias [63]. Thus, we cannot conclude that the warning was compelling enough to motivate the participants to comply (see the Limitations section). Nevertheless, the warning should be designed to allow for escalation, whereas in case the driver does not pull over in due time (eg, 2-3 min [56]) or explicitly rejects the warning, delayed reprompts with an increasingly severe tone would be delivered by the voice assistant (eg, “You are at risk of hypoglycemia. Please stop the car safely and check your blood sugar, then risk of hypoglycemia. Pull over now”).

Hypoglycemia Detection
Finally, in this paper, we focus on the interface between the hypoglycemia detection system and the driver, with the aim of visually distracting them as little as possible. Although the detection side is beyond the scope of this study, the designed warning is intended to be produced by a voice assistant built into the vehicle. Therefore, how a vehicle monitors blood sugar depends on the technology of the car. For instance, the aforementioned open source app displaying the glucose levels on the dashboard of a private vehicle [25] could be enhanced to connect with the in-vehicle voice assistant and use a voice warning instead of a visual one. Furthermore, research has been conducted on how to detect hypoglycemia from the car’s data [54] and from consumer-available wearable devices [64], with the argument that CGM devices can impose a social and financial burden on the individual.
Limitations and Strengths

Despite our best efforts, this research has 3 main limitations. First, the studies included a relatively small sample size. However, this study includes 3 feasibility studies (ie, a preliminary study with healthy individuals and 2 feasibility studies with individuals with T1DM), and the research presented in this paper is intended to be understood as an iterative development of a hypoglycemic warning. As such, this research aimed to pioneer the use of in-vehicle voice assistants for a driver health-related warning, rather than draw conclusions to be generalized to the population with T1DM. Thus, although we included a total sample size of 48 individuals, each feasibility study provides insight into the changes required by the users, and we provide the scientific community with an opening to the design of in-vehicle voice assistant–based health-related warning. Furthermore, previous studies on digital health systems used a similar sample size [65-67]. Thus, we believe that although the sample size does not allow drawing conclusions on the interaction of drivers with T1DM with in-vehicle hypoglycemia warnings, it still reports pioneer research.

Second, the studies were conducted over a short period. The participants had only a short-term experience with the warning. Perception and compliance may therefore be influenced by the novelty of such an experience, whereas perception may stabilize with repeated experience [68]. Future research should investigate the user experience of the warning in a longitudinal study. Third, these studies did not control for all potentially confounding variables related to real-world traffic and driver’s priorities. For instance, both simulator and real-car experiments involved disadvantages: while assessing the warning in a simulator allowed a controlled and safe experiment, such a setting remains artificial and lacks external validity. In contrast, while testing it in a real car increased the ecological validity of the human-machine interaction, it did not allow for as much traffic and speed variation as was possible in the simulator. Future research should investigate the effects of real-world traffic on the perception of the warning and compliance behavior. Moreover, receiving a warning in the presence of a team of experimenters may have influenced the participant’s verbal and behavioral responses; participants knew they would receive the warning sooner or later and had no reason not to follow it (eg, ignoring the warning because of being late for an appointment). In a real situation, drivers may not respond as expected or may even ignore the warning. Future research should test such a warning in a more ecological context, for instance, in a field study where the driver may not fall for a participant bias [69].

Finally, as we aimed to test a voice warning, our studies used a Wizard-of-Oz methodology to avoid problems related to natural language processing. Note that our studies were conducted in German-speaking Switzerland, where the German accents easily vary from region to region. As this aspect was beyond the scope of our research, we did not implement a working voice assistant or account for potential fallback intents triggered by the voice assistant’s failure to understand the user. Future research should push this research further and examine the potential danger of delayed treatment of hypoglycemia owing to the voice assistant’s natural language processing errors.

Conclusions

Although hypoglycemia increases the risk of care mishaps [7,8], current solutions (eg, CGM and FGM) require visual human-machine interaction, which is inappropriate for an in-vehicle context. As voice assistants are increasingly present in private vehicles [26,27] and the European Commission fosters safety technologies inside the car [16], we propose to warn the driver of their critical health state through a voice assistant–based health warning. This paper reports on an iterative development and assessment of a hypoglycemia warning. In particular, we conducted in 3 studies: a preliminary study using a simulator with healthy participants, a test with individuals with T1DM in a simulator, and a test with individuals with T1DM in a real car. This gradual increase in authenticity in the experimental design allowed us to increase the ecological validity of our results while keeping experimental control. To the best of our knowledge, this is the first attempt of such a comprehensive feasibility assessment of an in-vehicle voice warning for hypoglycemia. Our results suggest that a voice warning can be useful, but that proactive behavior in voice assistants is still emerging and unfamiliar. We hope that these preliminary findings will foster future research to further develop in-vehicle hypoglycemia warnings.

Acknowledgments

The authors would like to thank the Swiss Federal Department of Defense, Civil Protection and Sports for providing the closed circuit used in study 2.

This work was partly funded by the Swiss National Science Foundation, Sinergia Project (CRSIiS 183569).

Data Availability

The data supporting the findings of this study are available from the last author, TK, upon reasonable request.

Authors' Contributions

CS and EF were responsible for the oversight, leadership, and management of the research activity and for funding acquisition. CB, MM, VFL, MK, SF, TZ, FW, CS, EF, and TK were responsible for the methodology. CB and TK were responsible for the formulation of the research goals and aims, conceptualizing the voice warning. CB and MM were responsible for developing the voice warnings. CB, MM, FW, and TK were responsible for developing the driving scenarios. FW, TK, and TZ were responsible
for providing the driving simulator, the real car, and the closed circuit. CB and MM were responsible for recruiting participants for study 0. VFL, TZ, and CS were responsible for recruiting participants for studies 2 and 3. CB, MM, and VFL were responsible for data collection. CB was responsible for the data analysis and presentation and the first draft of this manuscript. VFL, MK, SF, TZ, CS, EF, ABK, and TK were responsible for critical feedback and final revisions of the manuscript.

Conflicts of Interest

CB, VFL, SF, FW, TZ, CS, EF, and TK are affiliated with the Centre for Digital Health Interventions, a joint initiative of the Department of Management, Technology, and Economics at ETH Zürich, and the Institute of Technology Management at the University of St Gallen, which is funded in part by the Swiss health insurer CSS. EF and TK are also the cofounders of Pathmate Technologies, a university spin-off company that creates and delivers digital clinical pathways. However, neither CSS nor Pathmate Technologies were involved in any way in the design, interpretation, analysis, or writing. All other authors declare no other conflicts of interest.

Multimedia Appendix 1

Original (German) and translated version of the conversation flow of the hypoglycemia voice warning in study 1 and study 2.

Multimedia Appendix 2

Questions used in the semistructured interview about participants' experience with the warning conducted in study 1 and study 2.

Multimedia Appendix 3

Illustration of the procedure across studies.

References


37. Dey AK, Abowd GD, Salber D. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. Hum Comput Interact 2001;16(2-4) [FREE Full text] [doi: 10.1207/S15327051HCI16234_02]


60. The Nightscout Foundation. URL: https://www.nightscoutfoundation.org/ [accessed 2023-12-22]


**Abbreviations**

**ATI:** Affinity for Technology Interaction  
**AUT:** Acceptance and Use of Technology  
**CGM:** continuous glucose monitoring  
**FGM:** flash glucose monitoring  
**meCUE:** modular evaluation of key Components of User Experience  
**RQ:** research question  
**SAI:** Session Alliance Inventory  
**TIDM:** type 1 diabetes mellitus  
**TRI:** Technology Readiness Index

---

Edited by A Kushniruk; submitted 15.12.22; peer-reviewed by M Chatzimina; comments to author 31.01.23; revised version received 06.02.23; accepted 04.12.23; published 09.01.24.

Please cite as:

Effectiveness and User Perception of an In-Vehicle Voice Warning for Hypoglycemia: Development and Feasibility Trial  
JMIR Hum Factors 2024;11:e42823  
URL: https://humanfactors.jmir.org/2024/1/e42823  
doi:10.2196/42823  
PMID:38194257
A Closed-Loop Falls Monitoring and Prevention App for Multiple Sclerosis Clinical Practice: Human-Centered Design of the Multiple Sclerosis Falls InsightTrack

Valerie J Block1,2, PT, DPTSc; Kanishka Koshal1, BDS, MSc; Jeeleene Wijangco1, BS; Nicolette Miller1, BS; Narender Sara1, MBA, MS; Kyra Henderson1, BA; Jennifer Reihm1, MPA; Arpita Gopal2, DPT, PhD; Sonam D Mohan1, MD, PhD; Jeffrey M Gelfand1, MAS, MD; Chu-Yueh Guo1, MD; Lauren Oommen1, DO; Alyssa Nylander1, MD, PhD; James A Rowson1, PhD; Ethan Brown1, MD; Stephen Sanders1, PhD; Katherine Rankin1, PhD; Courtney R Lyles3,4, PhD; Ida Sim5, MD, PhD; Riley Bove1, MMSc, MD

1Department of Neurology, University of California San Francisco Weill Institute, University of California San Francisco, San Francisco, CA, United States
2Department of Physical Therapy and Rehabilitation Science, University of California San Francisco, San Francisco, CA, United States
3University of California San Francisco Division of General Internal Medicine, Zuckerberg San Francisco General Hospital, San Francisco, CA, United States
4Center for Vulnerable Populations, University of California San Francisco, San Francisco, CA, United States
5Department of Medicine, University of California San Francisco, San Francisco, CA, United States

Corresponding Author:
Riley Bove, MMSc, MD
Department of Neurology
University of California San Francisco Weill Institute
University of California San Francisco
Box 3126 1651 4th St
Room 612A
San Francisco, CA, 94143
United States
Phone: 1 (415) 353 2069
Email: riley.bove@ucsf.edu

Abstract

Background: Falls are common in people with multiple sclerosis (MS), causing injuries, fear of falling, and loss of independence. Although targeted interventions (physical therapy) can help, patients underreport and clinicians undertreat this issue. Patient-generated data, combined with clinical data, can support the prediction of falls and lead to timely intervention (including referral to specialized physical therapy). To be actionable, such data must be efficiently delivered to clinicians, with care customized to the patient’s specific context.

Objective: This study aims to describe the iterative process of the design and development of Multiple Sclerosis Falls InsightTrack (MS-FIT), identifying the clinical and technological features of this closed-loop app designed to support streamlined falls reporting, timely falls evaluation, and comprehensive and sustained falls prevention efforts.

Methods: Stakeholders were engaged in a double diamond process of human-centered design to ensure that technological features aligned with users’ needs. Patient and clinician interviews were designed to elicit insight around ability blockers and boosters using the capability, opportunity, motivation, and behavior (COM-B) framework to facilitate subsequent mapping to the Behavior Change Wheel. To support generalizability, patients and experts from other clinical conditions associated with falls (geriatrics, orthopedics, and Parkinson disease) were also engaged. Designs were iterated based on each round of feedback, and final mock-ups were tested during routine clinical visits.

Results: A sample of 30 patients and 14 clinicians provided at least 1 round of feedback. To support falls reporting, patients favored a simple biweekly survey built using REDCap (Research Electronic Data Capture; Vanderbilt University) to support bring-your-own-device accessibility—with optional additional context (the severity and location of falls). To support the evaluation and prevention of falls, clinicians favored a clinical dashboard featuring several key visualization widgets: a longitudinal falls...
display coded by the time of data capture, severity, and context; a comprehensive, multidisciplinary, and evidence-based checklist of actions intended to evaluate and prevent falls; and MS resources local to a patient’s community. In-basket messaging alerts clinicians of severe falls. The tool scored highly for usability, likability, usefulness, and perceived effectiveness (based on the Health IT Usability Evaluation Model scoring).

**Conclusions:** To our knowledge, this is the first falls app designed using human-centered design to prioritize behavior change and, while being accessible at home for patients, to deliver actionable data to clinicians at the point of care. MS-FIT streamlines data delivery to clinicians via an electronic health record–embedded window, aligning with the 5 rights approach. Leveraging MS-FIT for data processing and algorithms minimizes clinician load while boosting care quality. Our innovation seamlessly integrates real-world patient-generated data as well as clinical and community-level factors, empowering self-care and addressing the impact of falls in people with MS. Preliminary findings indicate wider relevance, extending to other neurological conditions associated with falls and their consequences.

**KEYWORDS**
digital health; mobile tools; falls; prevention; behavioral medicine; implementation science; closed-loop monitoring; multiple sclerosis; mobile phone

**Introduction**

**Background**

Falls are common in patients with multiple sclerosis (MS), occurring in 50% to 70% of published cohorts, a rate similar to that of older adults [1]. Falls often lead to injury, result in significant health care costs [2-5], and increase the fear of falling [6,7]; furthermore, they lead to a decline in physical activity and participation in daily life as well as cause loss of independence [8,9]. Targeted interventions such as physical therapy (PT) can reduce falls and the fear of falling [10-12], but patients often underreport and clinicians undertreat this issue. Indeed, fewer than half of the people with MS who report falls receive falls prevention information from their clinician [13], and there is a lack of self-management apps to engage and empower people with MS about falls prevention [14-16].

To address this gap, multimodal closed-loop tools hold promise. Closed-loop tools can use real-time feedback and patient-generated data (PGD; such as those already validated in MS [17-22]) to continuously monitor and adjust interventions to improve outcomes. Such an approach has been used in biological functions and symptoms, such as insulin delivery or depression [23-25]. Unfortunately, in MS, apps on the commercial market exist outside of the health system, that is, away from the point of care. To close these gaps in care, a tool should close the loop of information flow from the patient to the appropriate clinician (depending on the diagnosis and symptoms being treated, ie, neurologist) at the point of care and back to the patient to support patient-centered care. Furthermore, the tool must address the behavioral barriers to change to promote the behaviors (eg, reporting, screening, treatment recommendations, and follow-up with timely refills or referral scheduling) likely to lead to falls prevention. From previous work, real-time PGD such as prospective near-falls reports, patient-reported outcomes, and changes in step count captured by wearable sensors all provide useful input for the closed-loop models [26,27]. The integration of these in a multimodal tool would enhance falls prediction accuracy and could act as an early warning system for timely PT referrals, reducing falls risk and related injuries [28,29]. However, challenges lie in delivering PGD to the point of care, granting access for prompt intervention, and active self-management. To be actionable, these PGD, generated from remote devices or patient-reported outcomes, must be delivered according to the 5 rights [30]: the right information, to the right person, in the right format, through the right channel, at the right time in the workflow. This is a hurdle that health systems have for the most part not yet overcome, and PGD are not typically integrated into care systems.

To address these challenges, we developed Multiple Sclerosis Falls InsightTrack (MS-FIT), a closed-loop falls monitoring and prevention app. MS-FIT enables seamless information exchange between patients and clinicians, driven by stakeholder input and human-centered design (HCD) principles [31,32]. It empowers individuals with MS to track falls, enhances clinician decision-making by providing real-world insights, and fills a crucial gap in self-management for falls monitoring and prevention.

**Objectives**

This paper describes the iterative process of the design and development of MS-FIT. MS-FIT is designed to integrate various data types to personalize falls risk assessments and interventions for individuals with MS. To achieve this, a planned process of engagement of patients and clinicians (ie, neurologists) was performed to ensure that MS-FIT aligns with user needs, whereas usability evaluations validated its potential impact on falls prevention. Subsequently, we will test the feasibility of implementation and effectiveness of MS-FIT in a larger clinical trial.

**Methods**

**Study Setting**

The primary clinical setting is the University of California San Francisco (UCSF) Multiple Sclerosis and Neuroinflammation Center, which provides specialized care to >6000 adults with MS annually. Clinician stakeholders were approached via email or in person and invited to participate in the study. Patients who had given permission to be contacted for research participation
or who had sustained falls in the past year were invited via secure email to participate as stakeholders.

**Ethical Considerations**

The University of California San Francisco Institutional Review Board approved all study activities (22-36680). Informed consent forms and Health Insurance Portability and Accountability Act documents were signed by each study participant (patients, clinicians, and other interviewees). Patients received US $50 (1-time compensation) for their participation in the study.

**Study Design**

The overarching approach was grounded in the principles and phases of HCD [31]. This process focuses on the usability and needs of those whom the tool is meant to serve, in this case, patients and clinicians. The development protocols included (1) thorough engagement from a comprehensive range of stakeholders, (2) models based on HCD approaches to ensure alignment with the needs of the intended users (patients and clinicians), (3) an evaluation of the tool’s usability using an established framework: the Health IT Usability Evaluation Model (Health-ITUEM) [33], and (4) plans to support the generalizability and scalability of the tool to other clinical settings associated with falls.

HCD involves a series of steps, articulated initially in the context of design [34] and expanded to health care [35]: inspire (empathize with all stakeholders), ideate (define the problem and conceptualize in an open-minded manner), implement (prototype solutions and test), and iterate. Figure 1 illustrates these phases in a modified double diamond approach as they were undertaken in the current project, depicting the iterative broadening and narrowing of content and layout throughout the phases [36]. Figure 2 shows the trajectory of MS-FIT and the assimilation of insights obtained from user interviews (involving patients and clinicians) throughout the phases of discover, define, develop (iterative), and deliver.

The initial prototype (prediscover) was developed based on feedback from extensive HCD of the BRIDGE point-of-care clinical dashboard (refer to the Technological Building Blocks subsection) summarized elsewhere [37,38], where both patients and clinicians expressed a desire for the integrations of features and episodes of falls to be incorporated into the design. The study team initially identified key elements for MS-FIT through a combination of clinical expertise and literature review [39,40] (Figure 3). These elements were then amalgamated into mock app screens using PowerPoint (Microsoft Corp) for the first round of patient interviews. Figure 3 illustrates the inaugural prototype, which was informed by valuable insights from observational [41] and interventional [39] studies that used PGD to monitor walking and falls in individuals with MS. In addition, the prototype draws inspiration from clinician-facing [42] and patient-facing [43] apps designed using HCD principles to promote shared decision-making and evidence-based practice in MS.

**Figure 1.** Modified double diamond approach: phases of development and stakeholder engagement. The double diamond depicts the human-centered design principles and framework, with iterations through the discover, define, develop, and delivery phases. The timeline and workflow of the human-centered design phases depict corresponding interviews and products. The curved arrows between “Define” and “Develop” indicate an iterative process between these 2 phases. MS: multiple sclerosis; MS-FIT: Multiple Sclerosis Falls InsightTrack; MVP: minimum viable product; PD: Parkinson disease; REDCap: Research Electronic Data Capture.
Figure 2. The trajectory of Multiple Sclerosis Falls InsightTrack (MS-FIT) through the phases of development and stakeholder engagement. The final tool components include a patient survey (MS-FIT patient survey) and a clinical dashboard (MS-FIT BRIDGE). The trajectory integrates feedback from user (patient and clinician) interviews through the phases of discover, define, develop (iterative), and deliver. The version numbers indicate a revised version of the patient- or clinician-facing prototype. “Other patients” refers to patients with Parkinson disease as well as orthopedics, neurorecovery, and geriatrics populations. “Full test” refers to the prototype testing in the contextual environment. MS: multiple sclerosis.

Figure 3. Initial proposal for Multiple Sclerosis Falls InsightTrack (MS-FIT), which involved designing a closed-loop integrated MS-FIT personal health library. MS-FIT is designed to enable patients to track their falls in the context of their lived experience, report them to their care team, and gain insight into multimodal contributors to falls, falls’ impact on daily life participation, and interventions likely to prevent falls. Clinicians, by using BRIDGE, can gain insight into which patients are falling between clinical encounters and how best to personalize risk reduction interventions for the individual patient. This prototype was generated from a number of insights from observational and interventional studies that used patient-generated data to monitor walking and falls in people with multiple sclerosis (MS) and from clinician-facing [42] and patient-facing [43] apps designed using human-centered design to facilitate shared decision-making and evidence-based practice in MS. (A) Patient facing app; (B) Live communication with clinician inbox; (C) Clinical dashboard: BRIDGE launches from the EHR using SMART or FHIR. API: application programming interface; EHR: electronic health record; EMR: electronic medical record; FHIR: Fast Healthcare Interoperability Resources; MS-FIT: Multiple Sclerosis Falls InsightTrack; SMART: Substitutable Medical Apps and Reusable Technologies.

Framework for Tool Evaluation

The Health-ITUEM framework appraises both subjective and objective outcomes that inform a tool’s usability [44]. In the design phases described herein, the subjective outcomes (satisfaction measured by the perceived ease of use and perceived usefulness) were primarily evaluated. Furthermore, the 4 key variables proposed by Mathews et al [45] to determine both (1) whether the tool (MS-FIT) reflects HCD principles and (2) whether it is likely to engage patients were applied. These four domains encompass (1) usefulness, (2) ease of use or learnability, (3) likability, and (4) effectiveness. These frameworks were used to categorize critical data and visualization elements, as well as the technological and clinical workflow aspects of MS-FIT [46].
Technological Building Blocks
The architecture of the tool was built leveraging existing tools, primarily BRIDGE and REDCap (Research Electronic Data Capture; Vanderbilt University).

BRIDGE
The BRIDGE precision medicine platform at UCSF is an application programming interface (API) that assembles clinical and research data from a variety of sources into a dashboard customized for a given clinical context, displaying a series of digestible, actionable visualizations [38]. BRIDGE is integrated with the Epic electronic health record (EHR; Epic Systems Corporation), launches from Epic using Substitutable Medical Apps and Reusable Technologies (SMART) on Fast Healthcare Interoperability Resources (FHIR; a standard approach for building reusable and extendable EHR-integrated apps), and is integrated with Epic FHIR APIs and other data integrations. The back-end of BRIDGE is built using Python, the flask framework, and PostgreSQL to store configuration data. Although individual-level data will populate the tool, cohort-level data can become the reference cohort against which an individual’s data can be contextualized. BRIDGE pulls data not only from the EHR but also from a range of custom research databases as well as other APIs, such as REDCap [38]. BRIDGE was developed based on extensive HCD processes both within the field of MS [42,43] and beyond [37]. The data visualizations can be developed using HTML, cascading style sheets (CSS), JavaScript, Data-Driven Documents–JavaScript, and other front-end libraries. Each front-end visualization is modular, allowing for asynchronous loading, and is a parameterized JavaScript component, allowing us to extend the code to additional platforms and data sources. Data formatting standards are also applied to make all visualizations and data inputs modular. All API calls are made in real time; BRIDGE does not store patient data, but there is an option to write back to the EHR by pasting the visuals into a clinical note. Furthermore, the development team follows universal design principles, influenced by the Agency for Healthcare Research and Quality Toolkit for Designing Consumer Health IT [47].

REDCap Tool
REDCap [38] includes editable or annotatable functions to enable patients to keep track of, and annotate, their PGD. Design choices reflect digital health literacy principles and feedback provided from diverse patients. Together, these enhancements make the data understandable and actionable.

Investigator Team
The core team included an MS neurologist with HCD expertise (RB), software engineers (NM and NS), a health literacy and user interface or user experience experts.

Phases of Design
Phase 1: Discover

Stakeholder Advisory Team
An initial stakeholder meeting took place, during which the goals and phases of the project were outlined. Next, the core team met biweekly as a group or as subgroups to discuss an agenda that included the development of patient and clinician interview guides, interview coding schemes and thematic analysis, the practical aspects of the technological lift, the workflow integration of MS-FIT, and the visualization types and customizations. The iterations of mock-ups were revised based on patient and clinician interview feedback.

Interviews
One-on-one interviews were conducted by the health literacy and patient engagement expert (hereinafter referred to as the interviewer) with patients (round 1) and clinicians. Because of ongoing COVID-19 restrictions on in-person engagements, interviews were conducted via the UCSF Zoom video platform (Zoom Video Communications, Inc) using interview guides developed for each audience to elucidate how a tool might be designed to promote behavior change around falls ascertainment, reporting, and prevention. All questions were administered verbally; and interviews lasted between 45 and 60 minutes.

Interview guides included qualitative and quantitative components. Open-ended questions probed around the domains of the capability, opportunity, motivation, and behavior (COM-B) framework to facilitate subsequent mapping to the Behavioral Change Wheel (BCW) proposed by Michie et al [48]. Quantitative questions with Likert-style responses (ranging from 1=lowest to 5=highest) were administered verbally throughout each interview to assess specific aspects of patient and clinician experience related to capability, opportunity, and motivation, as well as the perceived usefulness of mock screen views and workflows. Participants were asked to comment on their Likert-style responses.

Patient interviews were semistructured around 2 key thematic topics: patient experience with (1) falls and activity, including ability to be active, knowledge, communication with care team, experience, feelings, and expectations; and (2) use of technology, including smartphone, tracking devices, apps, and communication with care team. To complement qualitative insights, patients were asked to use a Likert scale to rate the perceived usefulness of each of 3 app screen views featuring different design elements.

Semistructured interviews with clinicians started with a review of the activity blockers and boosters identified during the discovery phase interviews with patients. With this insight,
clinicians were asked a series of open-ended qualitative questions to elicit their perspectives on whether a falls reporting tool might promote sustainable falls prevention, as well as gather feedback on the initial closed-loop design (Figure 3) intended to support falls treatment and clinical decision support. To assess each design feature, clinicians were asked to rate perceived usefulness on a Likert scale.

Analysis
After all interviews were concluded for each audience, the interviewer reviewed each transcript and used inductive coding to develop a coding scheme on the basis of responses to the open-ended questions [46]. Frequently occurring topics were assigned a unique thematic category, and less frequent topics were coded other. Categories were defined by the interviewer, and quotations from the transcript were used to illustrate the type of text coded into the category. Although the interviewer was the sole coder, the stakeholder advisory team provided ongoing consultation on the coding scheme and how to code less frequently occurring responses.

The interviewer transferred Likert-style response data to a spreadsheet to calculate means and SDs for each question. To analyze questions designed to map to the COM-B framework, the interviewer created a data grid where the rows were COM-B categories with subthemes of ability blocker and booster types, and the columns were evidence (quotes) of specific blockers or boosters [49]. Evidence of blockers or boosters that spanned >1 category were placed in all relevant categories to ensure that they would be represented when considering BCW-guided interventions.

After developing the initial COM-B data grid, the interviewer, in consultation with the stakeholder advisory group, expanded the grid to include (1) BCW intervention functions to help users overcome barriers to performing target behaviors and (2) potential intervention solution features designed to be effective for each corresponding blocker category. Intervention solution features were subsequently added to the design road map for immediate or future implementation.

Phases 2 and 3: Define and Develop (Iterative)
Stakeholder Advisory Team
In these phases, the team reviewed qualitative and quantitative findings from additional patient (2 rounds) and clinician (1 round) interviews and used this feedback to further refine MS-FIT tool functionality, including design and technological features. Changes were prioritized according to the strength of feedback (occurrence of themes and usability scores) and technical feasibility.

Interviews
The define and develop phases encompassed a second round of patient interviews, followed by 2 rounds of interviews with clinicians and patients designed to assess MS-FIT generalizability to other high-risk clinical contexts. The same process was followed as that described in phase 1 (discover). One-on-one interviews were conducted by the interviewer via the UCSF Zoom video platform using interview guides. All questions were administered verbally, and interviews lasted between 45 and 60 minutes. With participant consent, interviews were simultaneously recorded and transcribed using Zoom’s video transcription feature.

Patient Interviews (Round 2)
Interview guides included qualitative and quantitative components. In an effort to validate the patient experience findings from round 1 interviews, patients interviewed during round 2 were similarly asked to share qualitative feedback around personal experiences with falls, falls and near-falls reporting, perceived benefits and concerns around using a falls tracking app, and thoughts on what supports would be helpful between appointments. Quantitative questions with Likert-style responses (ranging from 1=lowest to 5=highest) were used to rate 9 mock screens for usefulness, understandability, and importance for each view. Mock screens had been iterated after the discover phase; therefore, patient feedback during this second round further validated and helped refine the designs.

Generalizability to Other High-Risk Clinical Contexts
To ensure that the technological build was not overdesigned for MS and to support the scalability of the tool to other clinical settings, interviews were expanded to intended users in other clinical specialties associated with falls, including geriatrics, orthopedics, neurorecovery (after stroke or traumatic brain injury), and Parkinson disease (PD). Clinicians from each discipline and patients with PD were interviewed. Interview protocols used during the discover phase were adapted to reference specific disciplines and diseases, whereas the questions (qualitative and quantitative) remained the same to yield a parallel assessment of each audience’s experiences, preferences, capabilities, opportunities, and motivations.

Analysis
Qualitative and quantitative interview analysis used the same inductive coding and calculation techniques, respectively, used during the discover phase. The results were analyzed by the interviewer, with ongoing thematic consultation with the stakeholder advisory team, and used to inform and prioritize design and content iterations.

Phase 4: Deliver
Stakeholder Advisory Team
The core team met with stakeholders on an ad hoc small-group basis during this phase to plan observation and tool-scoring protocols, specifically to identify a subset of questions from the Health IT Usability Evaluation Scale (Health-ITUES) derived from the Health-ITUEM to assess the 2 subjective components of usability—usefulness and ease of use [33]—as well as the Patient Education Materials Assessment Tool for Audiovisual Materials to assess understandability and actionability [50]. As recommended for digital tool validation [45], a single survey question—Net Promoter Score (NPS)—was asked regarding the likelihood that users (patients and clinicians) would recommend the MS-FIT to colleagues or friends. Additional conversations focused on the scalability of the tool, as well as the qualitative and quantitative feedback received.

https://humanfactors.jmir.org/2024/1/e49331
Observations and Scoring
Observations and scoring for the patient-facing falls assessment survey took place with 2 audiences: people with MS and people with PD. Patients scheduled for a routine upcoming in-person clinical visit with their neurologist were contacted and invited to participate in testing and evaluating the tool. After providing informed written consent, and while being observed by the interviewer, participants were asked to engage with the MS-FIT minimum viable product consisting of the falls assessment survey and accompanying patient instructions while being observed by the interviewer. Patients were specifically asked to complete the falls assessment survey by entering up to 5 falls (real or hypothetical) that had occurred in the prior 2 weeks and responding to on-screen prompts to provide context about each reported fall. Patients could ask questions of the interviewer, if needed. After survey submission, each patient was asked to complete an 18-item survey about their experience to assess usability, usefulness and ease of use, likability, understandability, actionability, and NPS. Patients were subsequently asked if they had any feedback about their experience. Feedback was documented in field notes captured contemporaneously.

Clinicians seeing people with MS and those with PD who had just been observed entering data in the falls assessment survey were asked to launch the MS-FIT BRIDGE app in real-time clinical encounters with these patients to review the falls and contextual data the patient had entered and to engage with the various widgets designed to help evaluate and address reported falls. The interviewer met with the clinicians immediately after the encounters to conduct in-person exit interviews and administer a 9-item survey to assess usability, usefulness and ease of use, likability, understandability, actionability, and NPS. Clinicians were subsequently asked whether they had any feedback about their experience, including any barriers to use and functionality challenges. Feedback was documented in contemporaneous field notes.

Analysis
Qualitative feedback, although limited, was analyzed by the interviewer using the same inductive coding technique used during the previous 2 interview phases. Quantitative questions with Likert-style responses (ranging from 1=lowest to 5=highest) were used to score likability, usability, usefulness, and ease of use. Understandability and actionability were assessed using a binary agree or disagree scale. Another member of the research team entered quantitative responses into REDCap, which was used to calculate means and SDs for all Likert-style responses and total binary responses. NPS responses (0-10 scale) were calculated by the interviewer by subtracting the percentage who were detractors (those who scored 0 to 6) from the percentage who were promoters (those who scored 9 or 10). An NPS >0 was considered good, >20 was considered favorable, and >50 was considered excellent.

Development Action Items
Once the tool was live, the developer was able to debug MS-FIT; iterate based on patient, clinician, and stakeholder feedback; and redebug as needed.

Results
Overview
Demographic information about each interview panel is shown in Multimedia Appendix 1 [42]. Altogether, 30 patients of diverse ages, disability levels, and technological literacy as well as 14 clinicians provided at least 1 round of feedback. The level of involvement from the users ranged from testers to informants [32]. Feedback from both rounds of interviews with people with MS, MS clinician comments, and feedback from other high-risk clinical context patient and clinician interviews were integrated into the final MS-FIT design. Iterative interview feedback was categorized into activity blockers (what keeps people from performing a behavior) and boosters (what is already working well that we can build on) in the COM-B model. Examples of how interview feedback findings fit into the COM-B and BCW, along with intervention function solution features integrated into MS-FIT, are shown in Figure 4. Details are provided in Table 1. Further discussions with clinicians and patients in other high-risk clinical contexts confirmed the findings from the MS context. Across these specialties, the main barriers to falls prevention efforts included access to specialized PT (availability and physical ability to access it), insurance coverage, ability to adapt the home to improve safety, the adequate use of assistive devices, and COVID-19–related restrictions to community exercise areas.

The overview of findings from interviews with clinicians and participants with MS, highlighting areas that block or boost patient and clinician behavior change with regard to falls and falls prevention, are shown in Multimedia Appendix 2.
Figure 4. Example of mapping blockers and boosters relating to falls prevention (findings from interviews with patients with multiple sclerosis [MS]) to the Behavioral Change Wheel [49] and associated behaviorally informed intervention solution features. Examples for each of the sections of the capability, opportunity, motivation, and behavior (COM-B) model are highlighted, showing how these integrate into the Behavioral Change Wheel. The examples provided relate to patients’ reported goals, blockers (features that block falls prevention behavior), and boosters (features that boost behaviors related to falls prevention).

Table 1. Scoring of the final University of California San Francisco Multiple Sclerosis Falls InsightTrack app (REDCap [Research Electronic Data Capture]) by patients with multiple sclerosis: usability, ease of use, and likability (n=10).

<table>
<thead>
<tr>
<th>Health-ITUES&lt;sup&gt;a&lt;/sup&gt;–based questions for usability, ease of use, and likability</th>
<th>Score, mean (SD)</th>
<th>Score &lt;4 out of 5, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“It is useful to report if I’ve had any falls or near falls every 2 weeks”</td>
<td>4.80 (0.42)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>“It is useful to have my survey answers sent to my care team”</td>
<td>4.90 (0.32)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>“The survey asks about important topics”</td>
<td>4.70 (0.48)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>“I am comfortable with my ability to complete the survey”</td>
<td>4.80 (0.42)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>“I find the survey easy to use”</td>
<td>4.80 (0.42)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>“I can easily remember how to access the survey through my email”</td>
<td>4.60 (0.70)</td>
<td>1 (10)</td>
</tr>
<tr>
<td>“I like the survey”</td>
<td>4.80 (0.42)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Health-ITUES: Health IT Usability Evaluation Scale (scores range from 1=strongly disagree to 5=strongly agree).

**Tool Components**

Thematic saturation was reached after 5 patient interviews (round 1), and we incorporated these insights into prototypes for an additional 5 patient interviews (round 2), which were then iteratively reviewed.

**UCSF Support Self-Monitoring: A Patient-Facing Tool to Track Falls and Self-Monitor**

**Tool Architecture**

One key and consistent theme emerging from patient interviews was a preference for a simpler design for the patient-facing tool than had been initially conceived. Combined with a goal of maintaining confidentiality and keeping personal information within our university firewall, the study team opted for a tailored REDCap app rather than a custom new app.

**Tool Components**

Key features informed by patient and clinician feedback are detailed in Figure 5. Key features mapping to the COM-B framework (Multimedia Appendix 2 and Figure 4) are denoted by a red number and described in Textbox 1.
Figure 5. University of California San Francisco (UCSF) Multiple Sclerosis (MS) Falls Tracker: a patient-facing tool to track falls and support self-monitoring. This is the “MS-FIT [Multiple Sclerosis Falls InsightTrack] patient survey V2.0,” sent via email to patients with a REDCap (Research Electronic Data Capture) survey link. Key features mapping to user-generated perspectives and feedback and to the capability, opportunity, motivation, and behavior (COM-B) framework are denoted by a red number and described inTextbox 1.
**Textbox 1.** University of California San Francisco Multiple Sclerosis Falls InsightTrack: key features. The numbers correspond to the red numbers in Figure 5, which denote key features mapping to the capability, opportunity, motivation, and behavior (COM-B) framework.

### Concise and precise falls screening

1. Clear definitions were preferred to distinguish between a fall and a near fall to support the reporting of meaningful data.
2. An easy-to-use and simple 1-question tool that could be completed frequently (every 2 wk) was preferred to relying on “flawed memory” to report falls during sporadic clinic visits: if “No,” then the survey ends at this point; if “Yes,” then branching logic continues.
3. The ability to easily report each fall or near fall separately was preferred. The ability to edit (return later) was important for reducing burden.
4. Simple reporting for near falls (yes or no and overall number) was preferred, given the large volume of near falls experienced by some patients and the potential burden and time commitment of providing details.
5. The 2-wk epoch between reporting was determined feasible (balance between memory and overburdening).
6. The ability to report activity limitations was preferred because these pertain to primary goals with regard to the “ability to continue independence for activities of daily living” and to “stay active.”
7. Because of the heterogeneity in answers, a free-text option would allow patients to add further details regarding activity limitations.
8. Indicating whether the patient has seen a neurorehabilitation specialist could help clinicians triage the continued plan of care.

### Detailed context of falls (optional)

1. Recording the date of the fall using a simple button allows the tool to display each fall into the longitudinal representation (refer to Textbox 2; Figure 6).
2. The time of falls can also inform falls context (eg, in the dark or when fatigued). The 24-h day was divided into time blocks for clarity and to reduce recall error of exact time.
3. Information regarding the medical consequences of a fall can inform both its severity and the clinical follow-up needed.
4. Injury after a fall is considered distinct from seeking and receiving medical attention.
5. Fall location can inform prevention efforts, including home safety; “some falls inside the home can be avoided through modifications such as removing a rug, better lighting etc.”
6. Other details of the fall location can also inform home safety and prevention (eg, curb, stairs, and poor lighting).
7. Specifying whether falls occur because of factors related to multiple sclerosis or other factors (obstacle, etc) is important owing to the heterogeneity of fall triggers and of clinical responses.
8. The question “If you have a mobility aid, were you using it when you fell?” can remind patients to use the assistive device and can cue clinicians of the need to modify or change the current assistive device.
9. A falls log is provided to patients and shows the reported falls over time.

### Closing the loop: real-time in-basket messaging

1. Enabling the reporting (patient) and ascertainment (clinician) of falls at regular intervals optimizes timeliness (vs periodic visits) while maintaining low burden (vs daily or “at time of fall”). If a severe fall is reported on the biweekly survey, an in-basket message to the electronic medical record alerts the care team in a manner integrated into the clinical workflow (Figure 6, #15).
Textbox 2. Multiple Sclerosis Falls InsightTrack clinical management dashboard integrated into the Epic electronic health record: key features. The numbers correspond to the red numbers in Figure 6, which denote key features mapping to the capability, opportunity, motivation, and behavior (COM-B) framework.

<table>
<thead>
<tr>
<th>Longitudinal multiple sclerosis trajectory widget (visualizes patients’ disease and medication trajectory over time with integrated normative ranges)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Ability to toggle through disability measures (eg, Expanded Disability Status Scale [EDSS] and Timed 25-Foot Walk)</td>
</tr>
<tr>
<td>2. Succinct overview of patient’s longitudinal MS trajectory, including relapses, disability, medications, and normative data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Longitudinal falls widget (visualizes falls reported every 2 wk by the patient using their patient-facing app [Figure 5] data regarding date, time, and severity of each fall on 1 display)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Fall severity visualized by color shade (grading falls by severity considered important to trigger an alert to the care team and to inform type of clinical response)</td>
</tr>
<tr>
<td>2. Ability to include a way to visualize the falls log with falls over time</td>
</tr>
<tr>
<td>3. Estimated time of day of the fall can inform further interventions needed, including vision check, home safety evaluation, and medication review (especially for Parkinson disease)</td>
</tr>
<tr>
<td>4. Time of day visualized with colors for daytime (lighter: yellow) and nighttime (darker: blue) preferred by all stakeholders</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Community resources widget (map automatically displays the patient’s home community and allows for web-based identification of MS health care professionals in their community)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Automated display of MS professionals (physical therapist, occupational therapist, and talk therapist) in the patient’s community, which reduces barriers for patient to identify local resources once physical therapy or other referrals have been placed</td>
</tr>
<tr>
<td>2. Contact information and driving navigations between the patient’s home and the resources automated, can be pasted into the patient’s after-visit summary or the clinician’s note</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cross-sectional widget (summary display with 2 tabs displaying clinical disability outcomes and patient-reported outcomes [PROs]; clinician can toggle between time points)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Clinic-based performance measures (walking speed, hand function, and cognition) and disability outcomes (EDSS with separate functional system scores) as well as PROs can inform a more global assessment of the patient at given time point</td>
</tr>
<tr>
<td>2. Color-coded normative ranges can provide rapid assessment of whether patient’s given function is within “normal” range</td>
</tr>
</tbody>
</table>

| Falls treatment– and action-prompt widget (tabulates core data needed for a comprehensive assessment of falls risk and prevention; for each category, patient’s score is colored according to severity, and possible action prompts are displayed) |
Phase 4: Deliver

Altogether, 15 patients (10 with MS and 5 with PD) with an age range of 34 to 79 years and 6 MS clinicians with a clinical experience range of 2 to 22 years (Multimedia Appendix 1) launched the tool components live and provided feedback.

**Patient-Facing UCSF MS-FIT**

**People With MS**

Of the 15 patients, 10 (67%) had been diagnosed with MS; they had a mean age of 48.8 (SD 8.8; range 34-60) years, with disability level (EDSS score) ranging from 1.5 to 6.0 and a median disease duration of 14.5 (IQR 6.3-24; range 2-27) years. The feedback from people with MS was overwhelmingly positive (Table 1). Likability scores were all NPS ≥ 100 (all promoters). The survey was found to be brief and clear. Patients appreciated the benefit of the closed-loop system and the overall impact on clinical encounters.

**Patients With PD**

Of the 15 patients, 5 (33%) had been diagnosed with PD; they had a mean age of 60.6 (SD 13.2; range 46-79) years, with a median disability level (Unified Parkinson’s Disease Rating Scale score) of 31 (IQR 30.3-8.5; range 17-42) and a median disease duration of 4 (IQR 1.5-8.5; range 1-10) years. Overall, the NPS was found to be 0 (20%-20%, with 1/5, 20% detractor, 1/5, 20% promoter, and 3/5, 60% passive scores that trended toward promoters), indicating that patients with PD could be easily swayed to use MS-FIT. The mean scores on the Health-ITUES questions were all >4 (ie, agree or strongly agree), and only 1 score was <3 out of 5 (Table 2).

Qualitative insights from the interviews revealed that falling, fear of falling, and thinking about falling were “not at the top of their list,” in contrast to people with MS. Nevertheless, patients with PD found the tracker “easy to fill out,” and they “liked the idea of reporting falls and reporting if [they] experienced fear of falling.” Patients with PD felt that it was important to have the ability to increase the font size (incorporated into MS-FIT patient survey v 2.0; Figure 5).

For future use in PD, patients reported that it would be important for ease of use and usability to have the ability to report motor vehicle accidents and specific PD symptoms as they relate to falls risk. Patients with PD also reported greater issues with using an iPad (motor or tremor issues).
Table 2. Scoring of the final University of California San Francisco Multiple Sclerosis Falls Tracker (REDCap [Research Electronic Data Capture]) by patients with Parkinson disease: usability, ease of use, and likability (n=5).

<table>
<thead>
<tr>
<th>Health-ITUESa–based questions for usability, ease of use, and likability</th>
<th>Score, mean (SD)</th>
<th>Score &lt;4 out of 5, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“It is useful to report if I’ve had any falls or near falls every 2 weeks”</td>
<td>4.40 (0.89)</td>
<td>1 (20)</td>
</tr>
<tr>
<td>“It is useful to have my survey answers sent to my care team”</td>
<td>4.20 (1.22)</td>
<td>1 (20)</td>
</tr>
<tr>
<td>“The survey asks about important topics”</td>
<td>4.00 (1.00)</td>
<td>2 (40)</td>
</tr>
<tr>
<td>“I am comfortable with my ability to complete the survey”</td>
<td>4.60 (0.89)</td>
<td>1 (20)</td>
</tr>
<tr>
<td>“I find the survey easy to use”</td>
<td>4.60 (0.89)</td>
<td>1 (20)</td>
</tr>
<tr>
<td>“I can easily remember how to access the survey through my email”</td>
<td>4.20 (1.10)</td>
<td>2 (40)</td>
</tr>
<tr>
<td>“I like the survey”</td>
<td>4.20 (1.10)</td>
<td>2 (40)</td>
</tr>
</tbody>
</table>

aHealth-ITUES: Health IT Usability Evaluation Scale (scores range from 1=strongly disagree to 5=strongly agree).

MS-FIT Clinical Management Dashboard

Overall, the MS clinicians (n=6) rated the dashboard highly (NPS=16.67; Table 3):

* I like that [the app] summarizes important clinical information in an easily digestible format, and the new widget that includes an MS [multiple sclerosis]-specific review of systems and actionable items seems like it will help ensure well-rounded MS care! [Clinician 1]

With regard to reporting falls and near falls, the MS clinicians noted multiple benefits to aiding with patient care:

* You can infer a lot from [fall data] in terms of disease activity, disease course, changes in a patient’s life, their living setting, their support. If you see a jump in falls or the onset of falls in a patient who wasn’t falling—it is worthy of clinical attention and needs to be addressed. It would give us an objective way to know if interventions are helping to reduce falls. [Clinician 2]

Near falls are particularly underscreened, so any granularity on near falls would be helpful. [Clinician 3]

* For some patients, near falls may not be worth reporting—may just be part of life. But other patients it could make sense for. Any change from baseline has potential to be significant. Near falls can be [a] canary in a coal mine. [Clinician 5]

Table 3. Scoring of the final University of California San Francisco Multiple Sclerosis Falls BRIDGE dashboard by multiple sclerosis clinicians: usability, ease of use, and likability.

<table>
<thead>
<tr>
<th>Health-ITUESa–based questions for usability, ease of use, and likability</th>
<th>Score, mean (SD)</th>
<th>Score &lt;4 out of 5, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“The information that appears in BRIDGE is useful to me.”</td>
<td>4.80 (0.41)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>“It is useful to be updated on my patient’s significant fall activity between appointments.”</td>
<td>4.50 (0.84)</td>
<td>1 (17)</td>
</tr>
<tr>
<td>“I find BRIDGE easy to use.”</td>
<td>4.20 (0.75)</td>
<td>1 (20)</td>
</tr>
<tr>
<td>“I can always remember how to access BRIDGE.”</td>
<td>4.00 (1.10)</td>
<td>1 (20)</td>
</tr>
<tr>
<td>“I like BRIDGE.”</td>
<td>4.50 (0.55)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

aHealth-ITUES: Health IT Usability Evaluation Scale (scores range from 1=strongly disagree to 5=strongly agree).

Discussion

Principal Findings

To our knowledge, this is the first tool designed using the HCD framework, anchored in the COM-B approach to behavior change, and capable of delivering relevant information at the point of care in line with the 5 rights with the aim of preventing falls in people with MS. Other apps have been developed, although the focus has mainly been on 1 component of falls (eg, evaluating falls risk [51]) at a time. In addition, many large-scale clinical research projects, such as those conducted at the Stanford Center for Digital Health and the Remote Assessment of Disease and Relapse–Central Nervous System program, are exploring applications of wearable data. However, most of the collected wearable data remain inaccessible for visualization or integration within a clinic’s EHR. This limitation can impede the effective use of PGD by clinicians and compromise patient-physician collaboration related to PGD [52]. MS-FIT fills a critical gap in multimodal closed-loop self-management apps for falls monitoring and prevention.

Through extensive stakeholder engagement, MS-FIT offers novel aspects of customization, generalizability, and scalability, integrating multiple data streams relevant to reducing falls. It provides rapid personalized in-basket notifications, limited to severe falls, and digitally displays PGD through the EHR, increasing the likelihood of adoption by patients and clinicians.

Designed in collaboration with patients and clinicians, MS-FIT has emerged as a well-received closed-loop tool for tracking falls and reducing falls risk in individuals with MS. Patients liked its brevity, simplicity, and overall utility, recognizing its
potential to enhance clinical discussions. The utility of between-visit reporting and contextualized information for identifying modifiable falls risks was acknowledged by both patients and clinicians. The trial phase aims to validate its low-burden design in practice. Clinicians welcomed the closed-loop system, foreseeing proactive interventions and streamlined implementation. Longitudinal falls visualization, incorporating time and severity, along with clinician prompts targeting MS symptoms and medication effects, was favored for its ability to capture often overlooked components during regular visits.

Another noteworthy finding was the minimal number of interviews required to attain thematic saturation in our initial discover phase, indicating that some clear guidance for potentially high-value initial design features was achieved with a minimal sample size. This could be due to the fact that MS-FIT was based on an initial prototype developed during a prediscover phase using patient and clinician feedback. It could also be attributed to homogeneous samples of study participants consulted throughout the discover and design phases. Overall design efficiency was likely aided by the experience and regular input of interprofessional teams.

The ongoing process involves testing MS-FIT in a prospective longitudinal study in a cohort of 100 adults with MS over 12 months. The primary objectives of this larger study include assessing the adoption rate of the tool, evaluating the level of sustained use of the tool, monitoring adherence to falls reporting, and assessing study retention over the 12-month period. Secondary and exploratory analyses will center around the prediction of adoption, sustained use, adherence to action prompts, and study retention. To determine effectiveness, the study will compare in-study falls with a prior falls data set (Fitbit remote monitoring in MS) [41], and patient satisfaction will be assessed during an exit interview.

Scalability
Our approach, characterized by the selection of key technological and clinical features, allows for the scalability and generalizability of the tool’s modular infrastructure to various symptoms, conditions, and clinical settings for other high fall-risk diseases as well as other symptoms within MS (eg, bladder dysfunction). Technological factors for scalability include (1) high-quality, widely shareable static visualizations; and (2) optimized industry standards for code sharing with clinicians in other health care settings, such as other MS centers using Epic EHR. However, successful integration into other health systems depends on the internal governance and motivation within each system.

Limitations
All interviews were conducted remotely, using the UCSF Zoom video platform, which may have biased the patient stakeholders to people who are technologically literate and have access to the internet. However, 92% of people in the United States have access to the internet [53], and given that MS-FIT is an app, users (patients or caregivers) are expected to possess a certain level of technical proficiency. Only clinicians at UCSF and patients seen by this (broad) group of clinicians were interviewed; therefore, we may have missed important feedback from a wider cohort of users. Although HCD is favored for user-driven eHealth innovations, certain limitations exist [32], including a narrow focus; thus, exploring alternatives such as value-sensitive design, citizen science, and more-than-human design could enhance inclusivity and impact within eHealth innovation [54]. Finally, having the interviewer serve as the primary coder could have introduced bias into the qualitative analysis process. Stakeholder advisory group engagement in the coding process was an effort to reduce any potential bias.

Conclusions
MS-FIT delivers relevant data to clinicians through an embedded window within the EHR, following the 5 rights approach. By using MS-FIT for data processing and algorithms, we reduce clinician burden while enhancing care. Our innovation extends to enabling and integrating real-world PGD as well as clinical and community-level factors, providing actionable information to empower self-care and addressing the impact of falls in people with MS. Our preliminary data indicate that this tool and design extend beyond MS and can be applied to other conditions associated with falls as well as the fear of falls and their associated consequences. To test the feasibility and effectiveness of the app, a clinical trial is ongoing (University of California San Francisco Clinical Trials identifier: NCT05837949).

Acknowledgments
This study was funded by the National Library of Medicine award (R01LM013396; principal investigator: RB). VJB is supported by the National Multiple Sclerosis Society Career Transition Award. RB is funded by the National Multiple Sclerosis Society Harry Weaver Scholar Award.

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

Conflicts of Interest
VJB is funded by the National Multiple Sclerosis Society Career Transition Award. CYG provides medical consulting for EMD Serono, Genentech, and Horizon Therapeutics. JMG receives research support to University of California San Francisco from Genentech, Hoffmann-La Roche, Vigil Neuroscience, and consulting for Aerialys. EB reports research funding from the Michael J Fox Foundation, the Gateway Foundation for Brain Research, the National Institutes of Health, and Biogen Inc. EB has also
received honoraria for his work as neurology section editor of NEJM Knowledge+ and consulting fees from Rune Labs, Inc. IS is the cofounder of Open mHealth; general assembly member for The Commons Project; serves on the medical advisory board for 98point6 Technologies; is cofounder, board director, and consultant for Vivli; is on the scientific advisory board for Myovant; and holds stocks in Myia. RB is a recipient of funding from the National Multiple Sclerosis Society Harry Weaver Award, the National Institutes of Health, DOD, and NSF, as well as Biogen, Novartis, and Roche Genentech. She has received personal fees for consulting from Alexion, EMD Serono, Horizon, Jansen, Genzyme Sanofi, and TG Therapeutics. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Demographic information for each round of interviews that led to the development of Multiple Sclerosis Falls InsightTrack.

Multimedia Appendix 2
Overview of findings from interviews with clinicians and participants with multiple sclerosis (MS), highlighting areas that block or boost patient and clinician behavior change with regard to falls and falls prevention. The table indicates whether intervention solution features were incorporated into the Multiple Sclerosis Falls InsightTrack (MS-FIT) patient survey, the clinician dashboard, or both.

References


Abbreviations

API: application programming interface
BCW: Behavioral Change Wheel
COM-B: capability, opportunity, motivation, and behavior
CSS: cascading style sheets
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
HCD: human-centered design
Health-ITU: Health IT Usability Evaluation Model
Health-ITUES: Health IT Usability Evaluation Scale
MS: multiple sclerosis
MS-FIT: Multiple Sclerosis Falls InsightTrack
PD: Parkinson disease
PGD: patient-generated data
PT: physical therapy
REDCap: Research Electronic Data Capture
SMART: Substitutable Medical Apps and Reusable Technologies
UCSF: University of California San Francisco
Please cite as:
A Closed-Loop Falls Monitoring and Prevention App for Multiple Sclerosis Clinical Practice: Human-Centered Design of the Multiple Sclerosis Falls InsightTrack
JMIR Hum Factors 2024;11:e49331
URL: https://humanfactors.jmir.org/2024/1/e49331
doi:10.2196/49331
PMID:35206662

©Valerie J Block, Kanishka Koshal, Jaelene Wijangco, Nicolette Miller, Narender Sara, Kyra Henderson, Jennifer Reihm, Arpita Gopal, Sonam D Mohan, Jeffrey M Gelfand, Chu-Yueh Guo, Lauren Oomen, Alyssa Nylander, James A Rowson, Ethan Brown, Stephen Sanders, Katherine Rankin, Courtney R Lyles, Ida Sim, Riley Bove. Originally published in JMIR Human Factors (https://humanfactors.jmir.org), 11.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Human Factors, is properly cited. The complete bibliographic information, a link to the original work on https://humanfactors.jmir.org, as well as this copyright and license information must be included.
Digital Triage Tools for Sexually Transmitted Infection Testing Compared With General Practitioners’ Advice: Vignette-Based Qualitative Study With Interviews Among General Practitioners

Kyma Schnoor\textsuperscript{1,2}, MSc; Anke Versluis\textsuperscript{1,2}, PhD; Niels H Chavannes\textsuperscript{1,2}, MD, PhD; Esther P W A Talboom-Kamp\textsuperscript{1,2,3}, MBA, MD, PhD

\textsuperscript{1}Public Health and Primary Care, Leiden University Medical Center, Leiden, Netherlands
\textsuperscript{2}National eHealth Living Lab, Leiden University Medical Center, Leiden, Netherlands
\textsuperscript{3}Zuyderland, Sittard-Geleen, Netherlands

Corresponding Author:
Kyma Schnoor, MSc
Public Health and Primary Care
Leiden University Medical Center
Hippocratespad 21
Leiden, 2333 ZA
Netherlands
Phone: 31 71526 8433
Email: k.schnoor@lumc.nl

Abstract

Background: Digital triage tools for sexually transmitted infection (STI) testing can potentially be used as a substitute for the triage that general practitioners (GPs) perform to lower their work pressure. The studied tool is based on medical guidelines. The same guidelines support GPs’ decision-making process. However, research has shown that GPs make decisions from a holistic perspective and, therefore, do not always adhere to those guidelines. To have a high-quality digital triage tool that results in an efficient care process, it is important to learn more about GPs’ decision-making process.

Objective: The first objective was to identify whether the advice of the studied digital triage tool aligned with GPs’ daily medical practice. The second objective was to learn which factors influence GPs’ decisions regarding referral for diagnostic testing. In addition, this study provides insights into GPs’ decision-making process.

Methods: A qualitative vignette-based study using semistructured interviews was conducted. In total, 6 vignettes representing patient cases were discussed with the participants (GPs). The participants needed to think aloud whether they would advise an STI test for the patient and why. A thematic analysis was conducted on the transcripts of the interviews. The vignette patient cases were also passed through the digital triage tool, resulting in advice to test or not for an STI. A comparison was made between the advice of the tool and that of the participants.

Results: In total, 10 interviews were conducted. Participants (GPs) had a mean age of 48.30 (SD 11.88) years. For 3 vignettes, the advice of the digital triage tool and of all participants was the same. In those vignettes, the patients’ risk factors were sufficiently clear for the participants to advise the same as the digital tool. For 3 vignettes, the advice of the digital tool differed from that of the participants. Patient-related factors that influenced the participants’ decision-making process were the patient’s anxiety, young age, and willingness to be tested. Participants would test at a lower threshold than the triage tool because of those factors. Sometimes, participants wanted more information than was provided in the vignette or would like to conduct a physical examination. These elements were not part of the digital triage tool.

Conclusions: The advice to conduct a diagnostic STI test differed between a digital triage tool and GPs. The digital triage tool considered only medical guidelines, whereas GPs were open to discussion reasoning from a holistic perspective. The GPs’ decision-making process was influenced by patients’ anxiety, willingness to be tested, and age. On the basis of these results, we believe that the digital triage tool for STI testing could support GPs and even replace consultations in the future. Further research must substantiate how this can be done safely.

(JMIR Hum Factors 2024;11:e49221) doi:10.2196/49221
Introduction

Background

The use of eHealth, health services delivered through the internet or related technologies, is increasing, especially since the COVID-19 pandemic [1,2]. The COVID-19 pandemic has shed light on the crucial role of digitization in health care [2]. An important and promising element of digitization in health care are digital triage tools consisting of a questionnaire for patients to identify the risk of a medical problem. These tools use a digital questionnaire typically administered by a health care professional, and an algorithm based on a medical decision tree generates automatic advice for follow-up, for example, a web-based symptom checker. In this paper, we discuss a digital triage tool that advises whether a specific diagnostic test for a specific combination of symptoms is necessary. This specific digital triage tool is based on Dutch medical guidelines.

Such a digital triage tool for different problems and symptoms could be an efficient and accessible method for citizens with medical questions. In addition, this digital triage tool could possibly lower the workload of general practitioners (GPs) as it can replace the triage that health care professionals would do themselves [3]. However, it is important that triage leads to responsible and appropriate care given the situation. Digital triage tools should not result in "over-triage" or "under-triage" [4]. Overtriage is when a patient is advised to undergo a medical treatment or diagnostic test when they do not have an (urgent) medical problem [4]. Undertriage is when a patient is told that they do not have an (urgent) medical problem when they do, with the advice that a diagnostic test or medical treatment is not necessary [4]. It is important to know whether the digital triage tool for diagnostic tests is in line with daily medical practice to maximize its validity.

In daily practice at GPs’ offices, medical guidelines are used to support their decision-making. GPs following guidelines has been an important research subject into the decision-making process of GPs with that of a web-based tool. At the same time, this research provides insights into the GP’s decision-making process for a diagnostic test. As a starting point, we investigated these research questions for sexually transmitted infection (STI) triage as the medical guidelines are straightforward (eg, clear risk factors and answer categories). Much research has been conducted on digital applications for STI testing, such as websites in which tests can be ordered, with positive feedback from patients about their usability [10]. Moreover, research has shown that a digital triage tool can potentially lower the threshold for STI testing [10] as this problem can be associated with feelings of shame [11]. To answer the research questions, a vignette-based qualitative study was conducted based on different STI-related patient cases [12].

Methods

Study Design and Participants

A qualitative vignette study was conducted using semistructured interviews with GPs as participants. Data saturation was expected after 10 interviews [13]. There were no specific exclusion criteria. GPs in training, practicing, or retired (for ≤5 y) could participate. In the interviews, the participants were presented with different patient vignettes (see the Materials section for details). After each vignette, the participants were asked about their clinical decision regarding STI diagnostic testing and to describe their thinking and decision-making process. This approach is called the “Think Aloud” method.
which allows for a description of how information is structured during a problem-solving task [14]. In addition, it provides rich data for analysis [15].

**Ethical Considerations**

This study was declared not to fall within the scope of the Dutch Medical Research Involving Human Subjects Act by the departmental ethics committee of the Leiden University Medical Center (reference 22-3002).

**Materials**

A vignette is a short hypothetical description of a patient representing a standardized combination of specific characteristics [16]. Vignettes made it possible to present patients with the same characteristics to every participant (eg, complaints, relationship status, and age) and, in this way, minimize variations between patients, which is not possible in real life. In this study, the vignettes were based on different aspects of the Dutch medical guidelines for STI testing [17]. In the medical guidelines, different aspects are taken into account to calculate the risk of an STI, such as endemic areas, unsafe sex, and different complaints. The following factors were incorporated into the vignettes: age, gender, sexuality, relationship status, employment (eg, full-time job or student), history of unsafe sex and how long ago it took place, number of sexual partners, frequency of unsafe sex, frequent GP visits, symptoms, and ethnicity. Some of these factors are not in the guidelines but were included to research whether they influenced the decision-making process of the GP (eg, situation and if the GP was visited often by that patient). In addition, the vignettes were designed in such a way that they would lead to advice from participants to undergo a diagnostic test for STIs or not. In total, 6 different vignettes were created and used (Multimedia Appendix 1). In Textbox 1, a short description of the vignettes is provided. The Dutch vignettes were designed with a GP and checked by another GP. An example of a translated vignette can be found in Textbox 2.

**Textbox 1. Short description of the vignettes.**

**Vignette 1**
- Woman, aged 20 years, from Spain, student, had unsafe sex multiple times >3 weeks ago, itching of the vagina, does not visit her general practitioner (GP) often

**Vignette 2**
- Man, aged 26 years, plumber, steady relationship, has irritation at the urethra and sensitivity when urinating, visits GP often

**Vignette 3**
- Woman, aged 17 years, high school student, had unsafe sex <3 weeks ago with no complaints, the first time she comes to the practice

**Vignette 4**
- Man, aged 24 years, has a relationship with a man, his partner has sexual contact with other men, has difficulty urinating

**Vignette 5**
- Woman, aged 45 years, has a steady relationship but thinks her partner cheated 6 months ago, has contact bleeding, visits the GP often

**Vignette 6**
- Woman, aged 35 years, has a steady relationship, comes from Surinam, has a burning sensation when urinating, visits her GP often

**Textbox 2. Vignette 1 translated from Dutch to English.**

- Mrs A is aged 20 years and studies in the Netherlands but comes from Spain originally. She has not visited you at the practice often. She is not in a committed relationship and has had unprotected sex several times in the past 6 months for more than 3 weeks. She experiences vaginal discharge and itching and irritation in her vagina. She wonders whether she might have a sexually transmitted infection.

**Procedure**

Participants were recruited via a LinkedIn post that included the email address of the researcher. Interested participants were instructed to send an email if they wanted to take part. In addition, participants were emailed from the network of the researchers, and the GPs could reply to the email if they wanted to participate. Interested participants were sent information and the informed consent form. In addition, different data and time points were included in the interviews, which could be face-to-face or digital (based on the preference of the participant). Participants had the right to withdraw at any time.

An interview protocol guided the semistructured interviews (Multimedia Appendix 2). All interviews were audio recorded. Each interview started with a short explanation of the study. The first vignette was then read out loud to the participant. They were asked whether they would advise undergoing diagnostic tests for STIs. Next, they were asked to share their reasoning process. These 2 steps were repeated for each vignette (ie, 6 in total). The first interviews were conducted with both interviewers present (KS and Fleur Rekveld), and KS was the

https://humanfactors.jmir.org/2024/11/e49221
Service: Digital Triage Tool

The digital triage tool was developed by a Dutch diagnostic center [18] based on a decision tree with Dutch medical guidelines [17]. The digital triage tool was developed in cocreation with GPs and clinical chemists. A Dutch academic knowledge center assessed the digital triage [19]. During triage, users first go through a series of questions. Their answers determine what question they have to answer next and, in the end, what advice is given. For example, the first question is “Did you have unsafe sex?” If the answer is “no,” the advice is not to be tested. If the answer is “yes,” a follow-up question appears: what is your gender? Gender is asked about as differences in gender result in different advice (eg, for women users who are advised to undergo a chlamydia test, it means that the service could advise doing a vaginal swab). Ultimately, the digital triage tool advises whether a diagnostic test for STIs is necessary and, if yes, which one (eg, chlamydia, gonorrhea, or HIV). The digital triage tool is now used in 2 digital services of the diagnostic company where patients can order diagnostic tests themselves with or without a health care professional. These diagnostic services are Directlab, where users can order web-based diagnostic test packages independent of a health care professional, and Homelab, where patients in the digital environment of their GP can order diagnostic test packages. In regular daily practice in the Netherlands, the patient needs to ask for a consultation with the GP (on the phone or in person) and ask for a diagnostic test for STIs. In this situation, the GP performs triage to identify whether it is necessary to conduct an STI test.

Data Analysis

To determine the diagnostic test advice of the digital triage tool, the characteristics of each vignette were entered into it. The ensuing advice was compared with the test advice of the GPs per vignette. To learn which factors influenced the GPs’ decision-making process, the combination of the think-aloud process, vignettes, and semistructured interviews was used as a triangulation method to obtain a complete range of data to result in a strong conclusion [12,20]. All interviews were transcribed (intelligent) verbatim. When the transcripts were completed and uploaded to ATLAS.ti (version 22; ATLAS.ti Scientific Software Development GmbH), the audio recordings were deleted. In total, 2 authors (Fleur Rekveld and KS) conducted the qualitative data analysis according to the principles of thematic analysis. Fleur Rekveld and KS developed a preliminary coding scheme based on the coded data from the first 8 participants. The final coding scheme emerged after all the coding was performed by the 2 authors independently. The codes were grouped into themes and subthemes.

Results

Characteristics of the Study Population

Data saturation was reached after 10 interviews. The characteristics of the participants are presented in Table 1. Their ages ranged from 32 to 70 years, with a mean of 48.30 (SD 11.88) years. The number of men and women was almost equal (6/10, 60% and 4/10, 40%, respectively). Of the 10 GPs, 1 (10%) was retired, 3 (30%) were working part time as GPs, and 6 (60%) were working full time.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Age (y)</th>
<th>Gender</th>
<th>Employment status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32</td>
<td>Woman</td>
<td>Full time</td>
</tr>
<tr>
<td>2</td>
<td>55</td>
<td>Man</td>
<td>Part time</td>
</tr>
<tr>
<td>3</td>
<td>38</td>
<td>Man</td>
<td>Part time</td>
</tr>
<tr>
<td>4</td>
<td>59</td>
<td>Man</td>
<td>Full time</td>
</tr>
<tr>
<td>5</td>
<td>70</td>
<td>Man</td>
<td>Retired</td>
</tr>
<tr>
<td>6</td>
<td>53</td>
<td>Man</td>
<td>Full time</td>
</tr>
<tr>
<td>7</td>
<td>55</td>
<td>Woman</td>
<td>Full time</td>
</tr>
<tr>
<td>8</td>
<td>43</td>
<td>Man</td>
<td>Full time</td>
</tr>
<tr>
<td>9</td>
<td>38</td>
<td>Woman</td>
<td>Part time</td>
</tr>
<tr>
<td>10</td>
<td>40</td>
<td>Woman</td>
<td>Full time</td>
</tr>
</tbody>
</table>

Testing Advice of Digital Triage Tool Versus GPs

Table 2 shows, for each vignette, whether the digital tool would advise conducting an STI test and what each participant would advise to do. For 50% (3/6) of the vignettes (ie, numbers 1, 4, and 5), the digital triage tool’s advice aligned with all participants’ advice. For all 3 vignettes, the advice was to conduct a diagnostic test for STIs. For those 3 vignettes, the patients’ risk factors were sufficiently clear for the participants to advise to conduct a test.

In vignette 1, the most important decision-making factor was the patient’s age; young age combined with women was an important factor influencing the participants’ test advice as having an STI could make this woman infertile. Participant 7 answered the following:

I would test her, always with women of her age who are sexually active.

In addition, unsafe sex was an important factor in the decision to test.
For vignette 4, the main factor in advising to test was the “men having sex with men” risk factor. Participant 5 answered the following:

*It is male-male contact, and in addition, there are changes in sexual contacts so that he can do an STI test.*

For vignette 5, all participants would advise conducting an STI test as well. Furthermore, 80% (8/10) mentioned that they would also conduct cervical cancer diagnostic tests because of the symptom of contact bleeding. Participant 9 mentioned the following:

*In the case of contact bleeding, more research than only an STI is needed. It could be Chlamydia, but a smear test is needed to exclude cervical cancer.*

For the other 50% (3/6) of the vignettes, not all participants gave the same advice as each other or as the digital triage tool. For vignette 2, a total of 60% (6/10) of the participants agreed with the advice of the digital tool, and for vignettes 3 and 6, the proportions were 70% (7/10) and 80% (8/10), respectively. It is important to mention that the initial answer of the participants is presented in Table 2. It could be the case that participants answered “no” to advising an STI test for the patient initially. However, the participants mentioned that they would advise conducting an STI test after excluding other diseases. In addition, sometimes, the participants wanted more information about the patient’s situation before advising to conduct an STI test.

For vignette 2, most participants wanted to know more about the patient’s case before giving the advice to test for an STI. In addition, they wanted to conduct a physical examination or other tests, such as a test to exclude urinary infection, as the patient’s symptoms seemed not totally compliant with those of an STI. Participant 2 said the following:

*I would like to know a little more; why does he think he has an STI? Does he have other contacts next to his current relationship or an open relationship? Has he heard anything from his wife?*

Participant 4 answered the following:

*I would check his urine.*

Participants answered that the symptoms and risk factors were too unclear to advise an STI test. A minority of the participants would test for an STI to exclude it or to satisfy the patient’s request. Participant 2 answered the following:

*He asked for an STI test so I would do one.*

The participants mentioned that, sometimes, a patient does not have an apparent reason for wanting to take an STI test or the patient has no symptoms that fit with those of an STI. However, sometimes patients do not want to discuss this in detail, and participants found it important to allow for testing at a low threshold if patients asked for it themselves. Participant 9 mentioned the following:

*Maybe he (or his wife) is cheating, and they do not want to tell you that directly. It is always the question if the patient is honest with you, so I would test at a low threshold after I did a urine infection test, and then I think he would accept that.*

**Table 2.** Advice of the digital tool and the participants to test for a sexually transmitted infection.

<table>
<thead>
<tr>
<th>Vignette</th>
<th>Digital triage tool</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>P9</th>
<th>P10</th>
<th>Agreement, n (%)^b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vignette 1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>10 (100)</td>
</tr>
<tr>
<td>Vignette 2</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>6 (60)</td>
</tr>
<tr>
<td>Vignette 3</td>
<td>Later</td>
<td>Later</td>
<td>Later</td>
<td>Later</td>
<td>Later</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Later</td>
<td>Later</td>
<td>Later</td>
<td>7 (70)</td>
</tr>
<tr>
<td>Vignette 4</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>10 (100)</td>
</tr>
<tr>
<td>Vignette 5</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>10 (100)</td>
</tr>
<tr>
<td>Vignette 6</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>8 (80)</td>
</tr>
</tbody>
</table>

^aP: participant.  
^bPercentage of participants who agreed with the advice of the digital triage tool.

For vignette 3, most participants (7/10, 70%) answered that the patient could take an STI diagnostic test but at a later time. At this time, it was too early to detect an STI. A total of 20% (2/10) of the participants also mentioned that they would talk to the patient about her contraception and provide education about safe sex. Participant 2 said the following:

*She had unsafe sex, so I would do two things. Maybe check if she uses birth control, and I would tell her that she can do an STI test after two weeks.*

Vignette 6 involved a patient from an endemic area. In total, 25% (2/8) of the participants who agreed with the advice of the digital tool mentioned the endemic area as a reason for testing. Participant 10 mentioned the following:

*I would ask her some more questions; however, she is from Surinam, a risk area. So I would test her at a low threshold, especially for a serological test.*

The other 62% (6/8) of the participants mentioned low-threshold testing because of the patient’s symptoms. Most participants (6/10, 60%) mentioned that they would check for a urinary infection, some before conducting an STI test and others in addition to it. Participant 1 mentioned the following:

*I would check her urine first to ensure she has no urinary infection.*
It is important to note that almost all participants mentioned that, if a patient requested an STI test, they would meet the request. They also mentioned that, in some cases, they would also give patients more information about safe sex or conduct a physical examination. The decision to do so often depended on age or other risk factors such as contact bleeding. Especially in the case of younger patients, GPs educated them about safe sex and birth control. However, this information provision was not part of their decision-making process but rather of their consultation.

**Extra Factors That Influenced the Decision of the GPs**

There were several factors that the participants considered in their decision that were not included in the digital triage tool. The most important additional patient-related factors were anxiety about infection, the wishes of the patient, and age. Among all participants (10/10, 100%), the patient’s anxiety was an additional reason for referring them to an STI test. The participants reasoned that a request for an STI test is not made easily and that there may be an unknown reason behind it. In their opinion, when patients experience fear-related stress, it might hurt their health. Participant 10 mentioned the following:

> Sometimes you feel that there is more than they want to say, and then you decide to test at a low threshold.

Age played a role in the decision-making process of the GPs. This was especially the case in vignettes 1 and 3. The GPs mentioned that checking for STIs was important at a fertile age, especially for women. In the Dutch medical guidelines, it is noted that, below the age of 25 years, there needs to be a low threshold for STI testing even if patients report no complaints. Participant 6 answered the following in the interview about vignette 3:

> Especially in younger patients, you want to know what they know about sex and the transmission of STIs.

In 2 vignettes, the GPs felt the need to ask additional questions or conduct a physical examination. The digital triage tool only provides advice on an STI test. However, the symptoms may also indicate a urinary tract infection or a stage of cervical cancer. These tests are not advised via the digital tool but were advised by the participants in this study for those 2 vignettes.

One GP also considered who had to pay for the test and whether it was affordable. Participant 3 mentioned the role of the payer or possible reimbursement in the decision. He answered the following about vignette 6:

> If she wants to pay for a test and she wants to do a test...Then, she can do a test.

In summary, it can be generally said that GPs in this study paid extra attention to patient-related factors such as fear of infection, desire to undergo the test, and young age when deciding whether to request an STI test.

**Discussion**

**Principal Findings**

In this study, we tried to identify whether the advice of a digital triage tool based on medical guidelines aligned with GPs’ medical practice. The results showed that other factors, which are not part of the guidelines, played a role in the GPs’ decision-making process when determining whether to advise an STI test for a patient. The most important additional patient-related factors were the patient’s anxiety, wishes, and age. The GPs also considered who had to pay for the test and whether it was affordable. Finally, the GPs were willing in some vignettes to ask additional questions or conduct a physical examination. The most notable factors are discussed in this section and compared with the literature.

In line with other research, the GPs’ decision to test depends sometimes on the anxiety and wishes of the patient [7]; these factors were not included in the studied digital triage tool. This additional aspect aligns with the research by Hajjaj et al [5,7]. In addition, our results align with those of a study that researched the barriers to following guidelines among GPs [6] that showed that the patient’s preferences were considered more important than following guidelines.

The interviews showed that the age of the patients was an important factor that influenced the GPs’ advice. Specifically, younger age was an important reason to advise an STI test because of the risk of infertility and the sexual activity in this group. Age was not included as a factor in the digital triage tool. As STIs mainly occur under the age of 30 years, it is not surprising that GPs tend to advise testing more for patients in this age group [21].

From the literature, it was found that the factor “knowing the patient” influences the decision-making process of GPs [22]. Accumulated knowledge about the patient influences the context and interpretation of the conversation between the patient and the health care professional, especially in the case of psychosocial or unspecific problems such as fatigue. However, in this study, knowing the patient was not a factor that was considered in the vignettes. For this reason, the decisions that the GPs made in this study could be different in real life as they might know the patients.

In addition to patient-related factors (eg, the wishes of the patient), GP-related factors also influenced the decision-making process. The extent to which GPs were open to discussion with patients about why they wanted an STI test or to which GPs patients about why they wanted an STI test or to which GPs were willing to address patients’ concerns influenced the decision. In addition, based on the findings of this study, it seems that the GPs expressed a preference for obtaining a complete set of information before deciding. For example, some GPs wanted to have more information about the situation of the patients and their partners. In some cases, GPs wanted to conduct a physical examination or other diagnostic tests (eg, urinary infection) to exclude other diseases. The digital triage tool is strictly bound to the guidelines set up without paying attention to, for example, the anxiety of the patient or the need for additional information. Other guidelines have been developed for possible symptoms of urinary tract infection or cervical problems, which have not yet been combined on the internet.

The advice of the digital triage tool is straightforward and always in line with a strict algorithm. In this study, GPs were found to recommend a diagnostic test for STIs more often than...
the digital tool. In the Netherlands, a study showed that unnecessary diagnostics (overdiagnostics) are a common problem among Dutch GPs; slightly more than half of the participating GPs indicated that patients could submit a complaint for not requesting an examination that was indicated and that this played a role to some or a significant extent in the request for diagnostic testing [23].

Our study did not investigate whether the digital tool can prevent overdiagnostics, but we assume that it can be a powerful decision support tool for daily general practice, just as tools for pharmacotherapy are already in use. More research is needed to confirm this.

Another possible reason why GPs are more inclined to test seems to be that it could save them time [24]. For example, if a patient has vague symptoms, it would be easy to request some tests first without having a thorough conversation. Another possible reason specifically for low-threshold STI testing could be feelings of embarrassment to ask about sexual behavior [25]. Recently, a Dutch center for sexual health found that talking about sexual behavior is not done as often as it should by health care professionals [26]. This could be seen as an additional justification for supporting GPs with digital tools for STI testing.

This study does not suggest that digital triage is the holy grail to prevent overdiagnostics or that it is the solution to lower the work pressure of GPs. However, this vignette study confirms that GPs have a more holistic approach to their patients compared with a digital triage tool. A digital triage tool primarily relies on specific responses to predefined questions, whereas a GP can consider more factors such as social factors, lifestyle, and personal context. On the one hand, the comprehensive perspective of GPs might result in a higher frequency of diagnostics when compared with a digital triage tool. This is due to the GPs considering additional factors. Given the high workload and time constraints of GPs, the investigated digital tool can play a helpful role in daily decision-making. In contrast, this holistic approach by GPs could potentially lead to fewer diagnostics. Given their deep understanding of the patients’ condition, GPs are better positioned to assess the necessity of tests.

This study has several limitations. It could be that social desirability influenced the GPs’ answers on the vignettes and interviews. Potentially, the advice of the GPs was more in line with the guidelines compared with that in their daily practice as they were aware of the fact that they were part of research on this topic [12]. It is also worth mentioning that there could be a disparity between what people think they would do in a particular situation and their actual behavior [27]. In addition, this study is not generalizable to the entire field of diagnostics at general practices because of its focus on STI testing. As a starting point, this study identified factors that influenced the decision-making process of GPs for STI testing. In future research, we recommend investigating digital tools and the decision-making process of GPs for other common diagnostic tests.

A strength of this study is the combination of the vignette method, the think-aloud process, and the semistructured interviews, which aimed to obtain a complete range of data on the topic (triangulation). Although no actual patients were included in this study, we aimed to make the vignettes as valid as possible by developing and testing them with GPs. In addition, providing the same vignettes to different GPs made it easier to compare patients within different general practices instead of comparing real-life patients with different complaints and characteristics. Currently, we are working on a real-life study in which patients in the waiting room of a GP’s office complete digital triage for STI testing (the result of the digital triage tool is not shown to the patient), after which they go on to have their planned consultation with the GP. At this consultation, the GP will also advise whether to test for an STI; the advice of the digital tool and of the GP will be compared. We expect more detailed and practical information to further refine this working method using a digital tool.

A qualitative study in which GPs were interviewed about their general attitude toward the use of digital tools by patients in their practice showed that GPs’ attitudes toward digital STI diagnostic services were positive, and they acknowledged that the use of eHealth in their practice could result in a more efficient workflow [28].

It will be interesting to further investigate whether GPs are also willing to use digital triage tools as a standard gateway for their practice for some diagnostic tests. When a digital triage tool is implemented and integrated into the care pathway, it is important to investigate what users think of this integration and whether they are satisfied with this change in their way of working. For future research, it could be beneficial to make a comparison of the experiences of patients with a digital triage tool, triage at the GP’s office, and a mix. Notably, recent studies on digital chatbots for medical questions have shown that patients perceived the chatbot’s responses to be superior to those provided by GPs [29]. For future applications, it is essential to consider patients’ eHealth literacy before using a digital triage tool as the primary tool in daily general practice [30,31]; hybrid care might be a solution to address all types of patients. Finally, it is important to realize that the tool in the care pathway needs to stay up-to-date and needs to be changed when the medical guidelines are updated [32]. This study showed that (holistic) factors that are not part of the digital triage tool affect GPs’ decision-making. This is an interesting topic for future research as digital tools and artificial intelligence are increasingly being used in health care. Nowadays, GPs use digital medication prescription tools to support their decision-making, which could help with handwriting errors but also with poor treatment decisions [33]. Another example is an artificial intelligence system that could help GPs decide on the early detection of skin cancer [34,35]. Digital technologies such as these should be researched carefully to see what the impact and consequences are for both GPs and patients.

Conclusions

This study shows that, in some cases, patients receive different advice to undergo an STI test from a digital tool and from a GP. Other factors that are not part of medical guidelines play a role in the GPs’ decision-making process when deciding whether to request an STI test. The most important additional patient-related factors were the patient’s anxiety, wishes, and age. One GP also
considered who had to pay for the test and whether it was affordable. Finally, some GPs expressed a desire to ask additional questions or conduct a physical examination in certain vignettes. In comparison, the digital triage tool adhered more closely to the medical guidelines, with GPs being more inclined than the digital tool to recommend an STI test for the same patient case. Alignment between the digital tool and GP advice only occurred when the risk factors for STI testing were unequivocally evident. This confirms that GPs decide from a holistic perspective. On the basis of these initial findings, we cautiously posit that a digital triage tool for STI testing can potentially support GPs and may even serve as a substitute for in-person consultations in the future. However, it is imperative to conduct further research to establish safe and effective methods for implementing such a transition.

These conclusions should be approached carefully, recognizing that this study represents an initial exploration and that additional research is required to substantiate and refine these findings.

Acknowledgments
The authors would like to thank all general practitioners who participated in this study. They appreciate the time they invested in this research. In addition, the authors want to thank Fleur Rekveld for her contribution to this paper. Funding was provided by Unilabs. Unilabs played no role in the research design or data collection, analysis, or interpretation.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Translated vignettes from Dutch to English.
[DOCX File, 13 KB - humanfactors_v11i1e49221_app1.docx ]

Multimedia Appendix 2
Semistructured interview protocol.
[DOCX File, 16 KB - humanfactors_v11i1e49221_app2.docx ]

References


Original Paper

Evaluating Users’ Experiences of a Child Multimodal Wearable Device: Mixed Methods Approach

Nancy L McElwain1,2, DPhil; Meghan C Fisher1, DPhil; Camille Nebeker3, MS, EdD; Jordan M Bodway1, BSc; Bashima Islam4, DPhil; Mark Hasegawa-Johnson2,5, DPhil

1Department of Human Development and Family Studies, University of Illinois Urbana-Champaign, Urbana, IL, United States
2Beckman Institute for Advanced Science & Technology, University of Illinois Urbana-Champaign, Urbana, IL, United States
3Herbert Wertheim School of Public Health and Human Longevity Science, University of California, San Diego, La Jolla, CA, United States
4Department of Electrical and Computer Engineering, Worcester Polytechnic Institute, Worcester, MA, United States
5Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, Urbana, IL, United States

Corresponding Author:
Nancy L McElwain, DPhil
Department of Human Development and Family Studies
University of Illinois Urbana-Champaign
2004 Doris Christopher Hall
904 West Nevada Street
Urbana, IL, 61801
United States
Phone: 1 217 244 7168
Email: mcelwn@illinois.edu

Abstract

Background: Wearable devices permit the continuous, unobtrusive collection of data from children in their natural environments and can transform our understanding of child development. Although the use of wearable devices has begun to emerge in research involving children, few studies have considered families’ experiences and perspectives of participating in research of this kind.

Objective: Through a mixed methods approach, we assessed parents’ and children’s experiences of using a new wearable device in the home environment. The wearable device was designed specifically for use with infants and young children, and it integrates audio, electrocardiogram, and motion sensors.

Methods: In study 1, semistructured phone interviews were conducted with 42 parents of children aged 1 month to 9.5 years who completed 2 day-long recordings using the device, which the children wore on a specially designed shirt. In study 2, a total of 110 parents of children aged 2 months to 5.5 years responded to a questionnaire assessing their experience of completing 3 day-long device recordings in the home. Guided by the Digital Health Checklist, we assessed parental responses from both studies in relation to the following three key domains: (1) access and usability, (2) privacy, and (3) risks and benefits.

Results: In study 1, most parents viewed the device as easy to use and safe and remote visits as convenient. Parents’ views on privacy related to the audio recordings were more varied. The use of machine learning algorithms (vs human annotators) in the analysis of the audio data, the ability to stop recordings at any time, and the view that the recordings reflected ordinary family life were some reasons cited by parents who expressed minimal, if any, privacy concerns. Varied risks and benefits were also reported, including perceived child comfort or discomfort, the need to adjust routines to accommodate the study, the understanding gained from the study procedures, and the parent’s and child’s enjoyment of study participation. In study 2, parents’ ratings on 5 close-ended items yielded a similar pattern of findings. Compared with a “neutral” rating, parents were significantly more likely to agree that (1) device instructions were helpful and clear ($t_{109}=-45.98; P<.001$), (2) they felt comfortable putting the device on their child ($t_{109}=-22.22; P<.001$), and (3) they felt their child was safe while wearing the device ($t_{109}=-34.48; P<.001$). They were also less likely to worry about the audio recordings gathered by the device ($t_{108}=6.14; P<.001$), whereas parents’ rating of the burden of the study procedures did not differ significantly from a “neutral” rating ($t_{109}=-0.16; P=.87$).

Conclusions: On the basis of parents’ feedback, several concrete changes can be implemented to improve this new wearable platform and, ultimately, parents’ and children’s experiences of using child wearable devices in the home setting.

(JMIR Hum Factors 2024;11:e49316) doi:10.2196/49316

https://humanfactors.jmir.org/2024/1/e49316

JMIR Hum Factors 2024 | vol. 11 | e49316 | p.257
(page number not for citation purposes)
KEYWORDS
wearable devices; multimodal sensing; user experience; usability; privacy; children; mobile phone

Introduction

Background
Advances in pervasive sensing, internet of medical things, and digital health strategies more broadly [1-6] have rapidly accelerated over the past decade. Although digital health research among adults and adolescents has predominately used smartphones [7-9], parallel work with infants and children tends to use wearable devices [10], including motion sensors to detect body posture and physical activity [11], audio recorders to assess language environment and development [12,13], heart rate sensors to assess psychophysiology [14], and head-mounted cameras to capture infants’ visual perspective of the physical and social environment [15]. Such wearable technology, especially when paired with machine learning algorithms, permits the automated detection of children’s behavioral and physiological states, as well as caregivers’ responses, and has the potential to transform the field of child development through the collection of big data in real-world environments [16].

At the same time, the use of wearable devices among infants and young children in home environments raises unique ethical, legal, and social implications and logistical challenges. As such, careful attention to the perspectives and experiences of end users of such technology, in this case, parents and their children, is required. In this study, we assessed parents’ perceptions of and experiences with a novel wearable device, LittleBeats, developed specifically for use with infants and young children. LittleBeats, which is not Food and Drug Administration approved and used only for research purposes, integrates a microphone, a 3-lead electrocardiogram (ECG) sensor, and an inertial motion sensor to synchronously collect information about infant vocalizations, cardiac physiology (heart rate and respiratory sinus arrhythmia), and motion (eg, physical activity level, position, and discrete movements). The electronics are housed in a 3D-printed case (55×57×13 mm), which is placed on a specially designed shirt that the child wears. Data can be collected throughout the day at home, without the researchers present. In prior papers, we reported on machine learning algorithms used to detect and classify child and parent vocalizations using audio data from the LittleBeats device [17] and child sleep states using all 3 sensor modalities [18]. We also conducted technical validation studies to assess the signal quality of each sensor modality in relation to established laboratory protocols and gold-standard equipment [19]. Complementing these prior reports, we focus here on the critical issue of “user experience” among families and their children aged 1 month to 9.5 years. Using semi-structured interviews and parent questionnaires to assess parents’ experiences and perceptions, our mixed methods investigation examined usability, privacy, and perceived risks and benefits.

The “Digital Health Checklist” for Use in Child Development Research

The proliferation of digital health technologies has spurred a parallel examination of ethical practices and related decision-making processes around the use of such technologies with human participants. To evaluate the LittleBeats platform, we used the Digital Health Checklist developed by Nebeker et al [20,21]: it is grounded in the ethical principles of the Belmont Report [22], which speaks to beneficence, respect for persons (or autonomy), and justice, and the Menlo Report [23], which added the principle of respect for law and public trust. These principles form the foundation of a 4-domain framework that includes privacy, access and usability, data management (eg, collection, storage, interoperability, and sharing), and assessment of risks and benefits (Figure 1).

To date, the research and development of the Digital Health Checklist has been applied to digital health protocols in adult samples, including for use in cardiovascular disease prevention [24]; studies of human emotion [25]; and improvement of informed consent communications [26]. The current investigation extended the use of the Digital Health Checklist to research involving parents of infants and children. In doing so, we integrated ethical considerations specific to research with children [27]. Specifically, children are a heterogeneous group, and the potential benefits and risks to child participants need to be understood within the context of the child’s age and related physical, cognitive, and socioemotional abilities.

For instance, infants and toddlers may be more susceptible to risks related to emotionally stressful procedures because their coping abilities are less well developed and depend, in part, on support from caregivers. By contrast, older children may be better able to regulate emotions and exert their autonomy, although they might be at an increased risk in other domains. For instance, owing to their growing self-awareness and other awareness, preschool- and school-aged children may be increasingly susceptible to experiencing shame and embarrassment, heightened concerns about privacy, and other related risks to the child’s self-concept.

With developmental differences in risk assessment in mind, we assessed LittleBeats user experience among children representing a large age range (infancy through middle childhood). Although we did not interview children about their study experiences, we considered the children’s age in our analysis of parents’ open-ended responses and parents’ perspectives regarding how their children felt about and responded to the research procedures. Research with children requires parental consent and, depending on the child’s age, the child’s assent to affirm their willingness to participate in the research. The consent process related to the LittleBeats technology has been addressed in a prior report [26]; therefore, we did not consider issues related to the provision of parental consent before participating in this research. Instead, our focus here was on parents’ perceptions of and reflections on their own and their children’s experiences following the use of LittleBeats at home.

Although child development research incorporating the use of wearable devices is rapidly expanding [28-32], systematic assessment of parents’ perspectives and experiences (or ethical
considerations more broadly) of such research has been sparse. A notable exception is a report by Levin et al [33], which outlines several key concerns parents may have about participating in research using wearable or remote sensing devices. These concerns focus on privacy expectations, particularly regarding audio or video data (considered “high fidelity data streams”), data management, and data use (eg, for scientific vs commercial purposes). Although we know of no study that assessed parents’ perceptions and experiences of using wearable devices at home after data collection, Levin et al [33] provided valuable insights into parents’ general willingness to participate in such research. Among a nationally representative sample of 210 parents (n=105, 50% mothers) with at least 1 child aged ≤5 years, 71.4% (n=150) of parents responding to hypothetical scenarios indicated at least some willingness to participate in studies involving motion or physiological sensors (low fidelity), whereas a significantly lower percentage of parents (n=99, 47.1%) endorsed willingness to participate in studies gathering audio recordings at home. It remains unknown whether the concerns expressed in the study by Levin et al [33], in which parents hypothetically considered participating in different types of remote sensing research, would also be voiced among parents who participated in research in which their children wore a wearable device with multiple sensor types (eg, motion, physiology, and audio).

Figure 1. Four-domain framework of the Digital Health Checklist for researchers. The Digital Health Checklist for researchers depicts the 4 ethical principles undergirding the 4 key domains of access and usability, privacy, risks and benefits, and data management. Source: this figure is published with permission and reflects an adaptation of the Digital Health Checklist Developed for Researchers (DHC-R) [34,35].

This Study
Guided by the domains of the Digital Health Checklist [20], we assessed parents’ experiences with and perceptions of using LittleBeats at home using a mixed methods approach. In study 1, we conducted a qualitative (thematic) analysis of parental responses to a semistructured interview following the completion of 2 day-long LittleBeats recordings at home; children in this study were aged between 1 month and 9.5 years. In study 2, we collected data on parents’ perspectives of using LittleBeats (again, following the completion of several day-long recordings at home) from a separate, larger sample. In this second study, we administered close-ended questionnaire items developed considering the qualitative themes identified in study 1. The parents in study 2 also had the opportunity to provide open-ended comments. In study 2, we narrowed our developmental focus to children aged 1 month to 5 years because our substantive interests focused on early childhood, and analytic tools are currently being developed for LittleBeats data collected among children aged ≤5 years.

Study 1

Methods

Participants
A total of 47 families with children aged 1 month to 9.5 years were recruited through web-based forums (eg, Facebook [Meta Platforms, Inc] parenting groups) and flyers distributed to local organizations (eg, libraries and day care centers) in a small Midwestern city. Because the larger study from which data were drawn included assessments of child stress physiology, families were excluded if their children had any known cardiac abnormalities. Of the 47 families that participated in the larger study, 42 (89%) completed the follow-up interview about their experience of using LittleBeats at home. Interviews were not completed with 5 (11%) families because of losing contact with them or because interview procedures were not finalized at the time of their study participation.

From these 42 families, 43 children (n=20, 47% female) participated. In 1 instance, 2 (5%) children (aged 13 and 71 mo) were from the same family. Children were aged 1.1 month to
9.5 years (mean 44.9, SD 38.36 mo) and represented 6 age groups: young infants (aged 1-5 mo; 7/43, 16%), older infants (aged 6-17 mo; 10/43, 23%), toddlers (aged 18-35 mo; 7/43, 16%), preschool-aged children (aged 36-59 mo; 6/43, 14%), early school-aged children (aged 5-7 y; 7/43, 16%), and school-aged children (aged 8-10 y; 6/43, 14%). Overall, 22 (51%) children were first born, 11 (26%) were second born, and 9 (21%) were third or later born. Mothers were aged, on average, 35.04 (SD 4.09) years, and fathers were aged, on average, 37.42 (SD 4.48) years. Across mothers and fathers, the highest level of education reported included a high-school degree (1/79, 1%), some college or 2-year degree (18/79, 23%), a bachelor’s degree (22/79, 28%), or an advanced degree (38/79, 48%). Parents identified as Black (2/79, 3%), Asian (3/79, 4%), White non-Hispanic (70/79, 89%), Hispanic (2/79, 3%), or >1 race (2/79, 3%). These demographic data were missing for 2 (5%) of the 42 mothers and 3 (7%) of the 42 fathers. The mean family income was US $79,500 (SD US $25,000).

Ethical Considerations

This study was approved by the institutional review board at the University of Illinois Urbana-Champaign (protocol #21032).

Overview of LittleBeats Procedures

LittleBeats collects 3 streams of data (ECG, motion, and audio data) simultaneously while participants go about their everyday routines (Figure 2). Owing to COVID-19 protocols, all participant engagement was remote. LittleBeats kits (ie, LittleBeats device and shirt, ECG leads, disposable ECG electrodes, alcohol swabs to remove residue from electrodes, medical tape to secure wires on the child’s chest, charging cable and block, and setup instruction cards) were either mailed or delivered by a research coordinator to the family’s home. After receiving the kit, the mother and child met with the study coordinator through Zoom, a secure video web-conferencing platform. During this 40-minute Zoom visit, the study coordinator guided the mother through the LittleBeats setup (described in more detail subsequently), and the mother-child dyad participated in a series of tasks (video recorded for subsequent coding), including a baseline assessment of child stress physiology at rest and a mother-child play session. For child participants aged <7 years, the mother-child dyads were also asked to complete a brief series of age-appropriate motion interaction tasks, such as the mother picking up her child (aged 1-4 mo), the mother and child (aged 11 mo) clapping together, or the mother and child (aged 6 y) playing “Simon Says.” Toward the end of the Zoom visit, the study coordinator provided instructions for completing the LittleBeats home recordings. Families were asked to complete 2 day-long recordings (approximately 8 hours per day). All adults present at home during the recordings (eg, parents, grandparents, and babysitters) were required to provide consent to the LittleBeats recordings using a secure web-based form provided by the research team. If any nonconsenting adults were at home, parents were asked to turn off the device while these individuals were present. At the end of each day of recording, parents (usually mothers) completed a brief questionnaire about the day’s recording (eg, recording start and stop times). To compensate the families for their time, parents were sent a US $100 e-gift card.

With regard to setting up LittleBeats, the research coordinator walked the mother through the following setup steps at the beginning of the Zoom visit: (1) threading a set of ECG lead wires (20 cm) through the back of the shirt pocket, (2) connecting ECG leads jack (2.5 mm) to the LittleBeats device, (3) turning the device on by sliding the switch to the “on” position (confirmation that the beginning of the recording is indicated by a red flashing light displayed on the device), (4) placing the device in a snug, specially designed shirt pocket, which is secured using 2 snaps, (5) snapping leads to 3 repositionable latex-free gel electrodes, (6) putting the LittleBeats shirt on the child, (7) cleaning the skin (where the electrodes will be placed) with an alcohol prep pad and then placing the electrodes on the child’s skin, and (8) applying a small strip of 3M Micropore medical tape to each ECG wire approximately 5.1 cm below each ECG sticker to help secure the wires in place.

At the end of the Zoom visit, the research coordinator also walked the mother through how the LittleBeats device should be removed. The removal steps include (1) removing electrodes from the child’s skin and using provided alcohol wipes, as needed, to remove residual gel from the electrodes; (2) unsnapping the electrodes from the ECG wires; (3) taking off the LittleBeats shirt; (4) removing the device from the shirt pocket; (5) sliding the slide switch to the “off” position; and (6) plugging the device into the provided charging cable (microUSB cable).

**Figure 2.** (A) LittleBeats device case; (B) LittleBeats supplies, including electrocardiogram leads, electrodes, charger, and shirt; and (C) an infant wearing LittleBeats at home.
LittleBeats Device Design and Study Implementation for End Users

LittleBeats was developed with parents and children (ie, end users) in mind. To provide a context for parents’ interview responses about their study experiences, we noted several aspects of the device design and study implementation intended to proactively increase usability and decrease concerns about privacy. With respect to usability, we provided participants with clear, illustrated instructions in several formats (eg, hard copy and on the web). The device was also designed to be simple to use, with an on-off switch and a charging port, and we provided parents with the all the materials in the LittleBeats kit (refer to the Overview of LittleBeats Procedures section) that they would need to set up and use LittleBeats at home. For the child’s comfort, the device is compact (55×57x14 mm) and lightweight (1.48 oz), with foam padding lining the inside of the shirt pocket in which the device is to be placed. The shirts are adorned with a variety of pocket designs (eg, hearts, animals, trucks, and dinosaurs) to appeal to toddlers and preschool-aged children, and, as part of the LittleBeats kit, families received 2 shirts with different designs. For older children, we provided more age-appropriate solid shirt pockets.

With respect to privacy, audio recordings provide high-fidelity information regarding participants’ lives and require special considerations related to participant privacy, data confidentiality, and recording bystanders (for review, refer to the study by Cychosz et al [13]). Our approach to protecting participant privacy aligns with user-centered privacy protections recommended for mobile health research [36] and a “rights-based” approach adopted increasingly in the United States and used by the European Union (ie, General Data Protection Regulation), according to which individuals have the right to control their personal data, including but not limited to consent, erasure, secure data management practices, and transparency. For example, an important strategy to minimize privacy risks includes giving participants control over recordings [37,38]. In this vein, parents were told at several points during their participation (eg, consent process and consent form, verbally during the Zoom visit, and written instruction card) that they were free to turn off or pause the device at any time and that they could request that their recordings be partially or fully destroyed and not be used in the research. With respect to third-party individuals, parents were also instructed to use the device at home when only immediate family members or other consenting adults are present. Parents were informed that all data files were marked only by identification numbers, machine learning algorithms would be used to process the audio data, research personnel would listen to only snippets of the audio data files as part of checks on algorithm development and accuracy, and research personnel were trained to protect participant privacy and would immediately cease listening to audio snippets in instances where personal information (discussion of medical, financial, or other personal issues) is being relayed. To minimize the risk of data being intercepted during transfer (ie, uploading data via wireless or Bluetooth networks), data were stored directly on a microSD card on the physical device, and files were configured in such a way that only study personnel could access the data in a human-readable format (eg, wav files for audio) using a data processing pipeline developed specifically for LittleBeats. Because LittleBeats is not a commercial device, simple modifications can be made to the device firmware (eg, “turning off” ≥ 1 of the sensors) to suit research goals (refer to the study by Islam et al [19] for details about technical specifications).

Parent Interview and Coding Procedures

Upon the completion of the LittleBeats recordings, parents (41 mothers and 1 father) completed a brief phone interview about their experiences of using LittleBeats in the home. To help minimize social desirability biases in parental responses, such as parents’ reports of positive experiences with LittleBeats instructions received during the Zoom visit, these interviews were conducted by a second study coordinator who was not present during the Zoom visit. Guided by the dimensions outlined in the Digital Health Checklist [20], as well as special considerations related to research with children [27], our semistructured interview was designed for the purpose of this research to capture information about parents’ experiences and perspectives regarding access, usability, privacy concerns, and risks and benefits with respect to the use of the LittleBeats device and the process of carrying out home recordings. Participants rarely provided information specific to the fourth domain of the Digital Health Checklist, data management, which encompasses how data are collected, stored, and shared and the extent to which the data are accessible to other systems or interoperability. Given the nature of the LittleBeats data (ie, they are not shared outside the research team, not accessible or integrated with other systems, and not transferred via a wireless or Bluetooth network that might be susceptible to security breaches), the data management theme is somewhat less relevant to LittleBeats than to health applications that might be accessed by multiple users (eg, patients, health care providers, and insurance providers). When parents expressed their views on the processes of data collection, storage, and security in the interviews, they almost exclusively focused on the audio recordings and privacy considerations. Therefore, we coded these responses under the privacy domain.

The interview included 11 open-ended questions, and the study coordinator conducting the interviews used standard probes to gain more insight into parents’ experiences, perceptions, concerns, and questions (Multimedia Appendix 1). The interview questions allowed for feedback from all family members’ perspectives (ie, the participating child, participating parents, and any other children or adults in the home). All parent interviews, conducted by the same study coordinator to ensure consistency, were audio recorded with the participant’s permission. Interview recordings were manually transcribed, and identifiable information (eg, names and birth dates) and conversational placeholders (eg, “uh-huh”) were omitted from the transcripts.

We used Taguette [39], an open-source web-based tool for coding textual qualitative data, to capture prevalent themes in our interview data and followed the 6-step approach to thematic analysis defined by Braun and Clarke [40]. At step 1, a review of the transcripts provided preliminary ideas for codes. At step 2, initial codes were generated based on the data from 5
interview transcripts of parents with children from different age groups. Through a series of team discussions, we developed an initial codebook focusing on areas that fell into the larger categories outlined in the Digital Health Checklist [20]. Three transcripts were then used for training purposes, and 3 researchers individually coded the transcripts. Discrepancies were discussed, and additional changes were made to the codebook. Upon the completion of the training, 1 researcher (who was not informed of the specific study objectives) coded all the transcripts using the refined codebook. Reliability was assessed by having the fourth author code 8 randomly chosen transcripts, and among the parent responses that both coders deemed codable, agreement was excellent (Cohen \( \kappa = 0.967 \)). At step 3, the research team met on a regular basis throughout the coding process to identify and discuss potential themes. At step 4 and after the completion of coding, final themes were reviewed by checking themes in relation to the entire data set to ensure an accurate representation of the data. At step 5, themes were refined and finalized by providing descriptive labels and definitions. At the final step, we organized the results based on the key domains of the Digital Health Checklist and created a summary table of themes with selected interview excerpts to illustrate the findings.

Results

Overview

Themes identified under the major categories of access and usability, privacy, and risks and benefits are summarized in the subsequent sections. Overall, similar themes were identified across developmental periods, although specific examples illustrating a given theme often differed depending on whether the parent reported on their infant, toddler, preschool-aged child, or school-aged child.

Access and Usability

According to the Digital Health Checklist, the domain of access and usability prompts researchers to consider whether the participant will be able to use the device as intended. This may involve evaluating whether the product has infrastructure requirements, such as internet access, as well as whether the device has been successfully used in the target population. In this study, usability refers to parents knowing how and being able to successfully use the LittleBeats device and materials (eg, ECG leads). Furthermore, usability encompasses families’ experience of and ability to adhere to the study procedures more generally (ie, participant burden, eg, completing multiple day-long recordings), beyond the use of the device itself (Table 1).

A majority of parents expressed sentiments regarding their ability to easily operate the device (ie, turning the device on-off and charging the device). Some parents indicated feeling comfortable given their previous experience with comparable equipment, yet other parents with no such prior experience expressed similar views about the ease of use. Parents also commented that the instructions were helpful and appreciated having a variety of resources to refer to, if needed (eg, written instruction card, website, and study personnel contact). Aside from operating the device itself, parents had varying views on the materials needed to place the device on their child. Some parents noted that the design was well thought out and that setting up the electrodes was not complicated. However, other parents indicated some challenges with the materials, such as with threading the electrodes through the back of the shirt pocket.

Parents also expressed differing perspectives about the ease of setting up (and removing) the device. Although many parents felt comfortable placing LittleBeats on their children, some parents noted that gaining their children’s cooperation was sometimes a challenge. For instance, some parents reported difficulty putting the device on their “wiggly, squiggly” infants. Other parents reported reluctance on the part of their toddlers or preschool-aged children, who could express their opinions and desires verbally. Typically, if challenges related to child cooperation were experienced, it was during the setup phase, and parents suggested that once their child was wearing LittleBeats, it was quickly forgotten. Parents expressed that the placement of the device on the upper anterior torso (ie, chest) may be disruptive to some activities, such as napping for a child who is a tummy sleeper. Relatedly, the device being concealed in the shirt pocket, with the ECG leads underneath the shirt, was viewed as a disadvantage by some parents who wanted to know whether the device was recording properly or whether there was a malfunction (eg, device turned off or ECG electrodes fell off).

With respect to participant burden, parents expressed a mix of perspectives. Many parents described day-long recordings (ie, \( >8 \text{ h/d} \)) as feasible but challenging. However, parents noted factors that mitigated this challenge, such as the need to record for only a limited number of days spaced across multiple weeks, the ability to schedule their recordings when it worked for them, and the reduction in other competing activities due to the COVID-19 pandemic. In the same vein, parents expressed wanting more features to help them fulfill project expectations. Currently, the device provides no information to the user beyond an indicator light showing that the device is powered on. Parents found it difficult to know how long they had recorded for or how much battery charge was left when using the device.

In addition, many parents described the project as convenient, indicating that the remote data collection procedures were appealing. Being able to collect data at home, on their family’s own schedule, made it relatively easy to participate. Parents were not burdened by the need to travel to a research laboratory, and they could set up the device and start recording when it fit their schedule. Concerns about being able to keep the device on securely or ensure that the device was collecting data were voiced by some parents of older and more active children (eg, increased unsupervised time and gel adhesive weakening owing to perspiration). Other parents expressed their worry that their children would damage the device during data collection.
### Table 1. Themes, subthemes, and example excerpts related to the access and usability of the LittleBeats device and study procedures (study 1).

<table>
<thead>
<tr>
<th>Themes and subthemes</th>
<th>Example excerpts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating the device</td>
<td>• “Everything was pretty easy. It was easy to charge, it was easy to you know put the stickers on and attach, and like I said I don’t think she really felt like it was on. The first day after she asked after an hour ‘how long have I had it on’ I was like ‘why is it uncomfortable’ she was like ‘no I was just wondering’ and I was like ‘oh okay.’ I don’t think she even realized she had it on half the time.” (Parent of a school-aged child)</td>
</tr>
<tr>
<td>Instructions</td>
<td>• “They [the instructions] were very clear. I mean they made it so that I felt confident putting it on her and doing what I was supposed to do.” (Parent of an older infant)</td>
</tr>
<tr>
<td>Support materials (specially designed t-shirt, wires, and electrodes)</td>
<td></td>
</tr>
<tr>
<td>Ease of use</td>
<td>• “I think the t-shirt definitely made it easier to use. That was a nice little set up, and it made it, you know, stay in place and like see where it [the device] needed to be for it to be hooked up and stay in place...And then even with the hole on the inside [of the shirt] to make it easy to get all the cords. That was really a unique design tool but effective.” (Parent of a school-aged child)</td>
</tr>
<tr>
<td>Challenges</td>
<td>• “It was a little hard getting the black metal piece through the back of the shirt. Like I needed that hole to be a little bigger. So, I’m sure I ripped mine just a little bit...But I just made it a little bit looser.” (Parent of an older infant)</td>
</tr>
<tr>
<td>Setting up and removal of LittleBeats</td>
<td></td>
</tr>
<tr>
<td>Comfort with set up</td>
<td>• “I’m pretty comfortable getting it set up and turning it on. It seems pretty straightforward.” (Parent of a school-aged child)</td>
</tr>
<tr>
<td>Child cooperation</td>
<td>• “It was mostly the initial putting it on. She didn’t want to cooperate with letting us get it on...but after a little bit she forgot it was there because she didn’t have any issues messing with it and then when it was time to take it off she was fine.” (Parent of a preschool-aged child)</td>
</tr>
<tr>
<td>Location of device</td>
<td>• “I wish the device itself was a little more discreet. Well, he’s a stomach sleeper so for naps I had to take it off but if it was a little more discreet or was not in front of the t-shirt but maybe on the arm it would be more convenient.” (Parent of an older infant)</td>
</tr>
<tr>
<td>Participant burden</td>
<td></td>
</tr>
<tr>
<td>Time commitment</td>
<td>• “You know once we broke it up a little bit we could [complete recordings]. I was more worried about you know were rarely all home just the four of us especially now that quarantine is over...We’re just more on the go than we were a year ago.” (Parent of an early school–aged child)</td>
</tr>
<tr>
<td>Convenience</td>
<td>• “It was really easy for me as a parent. I drive my other son like I said to [research lab in different city] a bunch...and so that is just a drag, a lot of back and forth. But for I would say from a parent’s standpoint, this was very easy for me to do.” (Parent of an early school–aged child)</td>
</tr>
<tr>
<td>Worry about recordings</td>
<td>• “My son’s pretty active, so he sweats a lot over the course of the day. The little stickers would kind of migrate a little bit...So, I worry a little bit that the first recording like the second half of the day might not be as accurate as it was supposed to be.” (Parent of a preschool-aged child)</td>
</tr>
<tr>
<td>Worry about device</td>
<td>• “It would be nice if there were some kind of indicator of battery more visible. And it was also, you know, since I had to take it on and off then count the time, that was also kind of challenging...so some kind of indication of time would also be awesome but I don’t know how complicated it would be to make it.” (Parent of an older infant)</td>
</tr>
<tr>
<td>Privacy</td>
<td>• “A lot of the activities that she wants to do involve painting or drinking water...those kinds of worrying me every time she picks them up. I was more concerned about the hardware.” (Parent of a preschool-aged child)</td>
</tr>
</tbody>
</table>

**Privacy**

The privacy domain focuses on the types of personal information that are or will be collected about participants. In this study, privacy relates to participants’ expectations about and understanding of the process of data collection, in general, and the audio recordings, specifically. Furthermore, this category encompasses the control that participants had over the data collected (Table 2).

Many parents commented that they were initially apprehensive about the home audio recordings but that their worries subsided when provided with more details during the initial informational call with the study coordinator. Other parents noted feeling more comfortable with the audio recordings over time as they participated in the study. Some parents discussed that although
they had no concerns, their spouse or partner did. Typically, only 1 parent (usually the mother) was present for the initial informational call with the study coordinator, and this parent then conveyed information to the other parent, which often sufficed to relieve privacy concerns.

By contrast, for some parents, positive views of research, such as having trust or placing value in research, negated concerns about privacy. Other participants described not being concerned with the audio recordings because they “had nothing to hide.” From this view, the audio would capture a typical day in their life, and participants elaborated by describing that the recordings would include everyday family discussions as well as arguments, which participants conveyed as just part of ordinary family life. Others’ lack of concern regarding the audio recordings stemmed from their ability to control when they were recording and, consequently, what was being recorded. They described the process of turning the device on and off as relatively easy and, therefore, reported turning the device off when they were discussing private matters. Some participants mentioned developing ground rules ahead of time to ensure that private information was not discussed when recordings were taking place and, if needed, would alert or remind other family members of the recordings.

The possibility of recording other individuals beyond immediate family members was considered. In working to respect others’ privacy, the participants mentioned several challenges. Some participants expressed that they altered their typical day to avoid interacting with others so that they would not have to worry about unintentionally recording a nonconsenting individual. Other participants stated that although they had planned to record at convenient times when no nonconsenting individuals were around, unexpected situations arose. In addition, although parents had the ability to control when the device recorded, some parents acknowledged that remembering to turn off the device when others were around could be challenging.

Table 2. Themes, subthemes, and example excerpts related to privacy concerns about the LittleBeats audio recordings (study 1).

<table>
<thead>
<tr>
<th>Themes and subthemes</th>
<th>Example excerpts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial apprehension about audio recordings</td>
<td>“Cause that was my husband’s big question like ‘are they just going to sit and listen to our day?’ So, he was a little worried about that but once it was explained [that machine learning algorithms would be used to analyze the audio data] he was more comfortable and on board.” (Parent of an early school-aged child)</td>
</tr>
<tr>
<td>Unconcerned about audio recordings</td>
<td></td>
</tr>
<tr>
<td>Former views or experiences of research</td>
<td>“She [study coordinator] also told me that it is only used for research purpose and nothing else...I actually love to participate [in research studies]. It is only used for research purposes, so that’s okay.” (Parent of a preschool-aged child)</td>
</tr>
<tr>
<td></td>
<td>“I was in a study when I was pregnant and we did something similar...my understanding was that the recordings just gets run through the software so we really don’t have anything any interesting happening here so I wasn’t terribly concerned about that [the audio recordings].” (Parent of an older infant)</td>
</tr>
<tr>
<td>Just an ordinary family context</td>
<td>“I explained everything to everybody [family members, including older children in home]. I do remember there was one particular situation where my 10-year-old was getting into trouble and afterwards he said, ‘Well, they’re gonna hear that!’ And I said this is just a regular family, there’s nothing to be embarrassed about or whatever.” (Parent of a toddler)</td>
</tr>
<tr>
<td>Ability to control the recordings</td>
<td>“My husband’s a veteran, and he works at the V.A..., so we had to make sure we turned it off before he came home from work because a lot of times he talks about his day.” (Parent of a toddler)</td>
</tr>
<tr>
<td>Respecting others’ privacy</td>
<td></td>
</tr>
<tr>
<td>Adjusting routines or activities to accommodate the study</td>
<td>“I think the only thing is that we didn’t go play with some friends across the street those days where we would’ve otherwise. Like it impeded a little bit of our typical routine, but it felt pretty unobtrusive.” (Parent of a preschool-aged child)</td>
</tr>
<tr>
<td>Unexpected situations</td>
<td>“When something was happening that I wasn’t expecting, like when I would get a phone call or something like that, and I was just a little concerned about remembering to turn off the device.” (Parent of an older infant)</td>
</tr>
</tbody>
</table>

**Risks and Benefits**

Evaluating the risks of possible harms in relation to the possible benefits resulting from the knowledge to be gained from the research is linked to the principle of beneficence. Study benefits should outweigh the possible harm to participants and the groups they represent. Risk assessment includes evaluating the type of harm, psychological, physical, reputational, or economic. In addition, researchers must consider the duration, severity, and intensity of the possible harm. Specific to the risks associated with the use of LittleBeats at home, parents expressed varying views along several dimensions, including safety, child comfort, and understanding of the research and its direct outcomes for participants (Table 3).

Many parents expressed that they thought the device was safe for their children to wear. These parents described not being concerned about safety because of the design of the device and the protective features built into it (eg, device was enclosed, tape-covered wires, fitted shirt, and pocket with secure snaps).
Some parents indicated that they initially had safety concerns (eg, the device being close to the skin and use of Bluetooth to transfer data) before learning more about the device and its setup (eg, the device itself is not in contact with the skin but is placed in a padded pocket, data are stored directly on the device, and Bluetooth is not used for data transfer). In some instances, parents detailed concerns about their children wearing the device in unsupervised contexts, such as during naptime, and they preemptively removed the device before naps.

Parents also commented on their children’s level of comfort or discomfort. Several parents mentioned that they observed their child functioning normally, such as engaging in typical routines and activities. Parents also stated that their children did not express any discomfort and did not seem to notice that they were wearing the device after a while. Other parents noted their children’s discomfort in putting on or removing the electrodes and medical tape used to secure the wires on the chest. Some parents worried about how comfortable it would be if the child were to hit the device on another object, such as the edge of a table.

Finally, parents’ understanding of the research and its direct outcomes for their families may confer risks and benefits. Some parents revealed a limited understanding of how the data would be used (ie, the ultimate outcome of the research process) or wanted direct feedback on their children’s development, which could pose unintended risks (eg, unfulfilled expectations of direct benefits). Other parents voiced the benefits attributed to participating in the research project itself. For instance, participation provided dedicated time spent together as a family, or completing the surveys was an opportunity to reflect on their children’s activities and development. Several parents expressed their desire to contribute to the project because they recognized the importance of the research. Some parents indicated that they had enjoyed participating in previous studies, and others stated that this project’s description seemed interesting and fun. Other parents of older children revealed that when they initially talked to their children about the study, their children seemed interested in participating, so they signed up. Some participants communicated that their children enjoyed participating in the project, with one parent acknowledging that their children felt special for a day while wearing the LittleBeats shirt.

<table>
<thead>
<tr>
<th>Themes and subthemes</th>
<th>Example excerpts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Safety</td>
<td>“No [safety concerns] because all of the wires were covered by her shirt and taped down.” (Parent of an older infant)</td>
</tr>
<tr>
<td></td>
<td>“Not really [any safety concerns]. I mean the wires were short enough that I wasn’t worried about them.” (Parent of an early school-aged child)</td>
</tr>
<tr>
<td></td>
<td>“I thought it will get like hot because I recorded for the 8 hours straight, I didn’t stop it at all, I was worried maybe it’s gonna be hot or something, but it wasn’t hot at all. That was my main concern only.” (Parent of a younger infant)</td>
</tr>
<tr>
<td></td>
<td>“And then I did have an initial concern...about the safety of having that device running on Bluetooth. I’m not sure how it communicates data and that being so close to skin.” (Parent of an older infant)</td>
</tr>
<tr>
<td>Child’s comfort or discomfort</td>
<td>“I guess putting them [electrocardiogram electrodes] on wasn’t the hard part. The hard part was taking them off, especially the was a little bit hard, and my son is also not very fond of changing clothes.” (Parent of an older infant)</td>
</tr>
<tr>
<td></td>
<td>“I’d probably take it off especially because my little one is about 10 1/2 months and she’s a tummy sleeper so that would be uncomfortable.” (Parent of a preschool-aged child)</td>
</tr>
<tr>
<td></td>
<td>“I mean it seemed it was fine. My sons were playing outside you know riding their bikes and everything and they didn’t...say anything was uncomfortable.” (Parent of an early school-aged child)</td>
</tr>
</tbody>
</table>

Outcomes of participating in the research

| Limited understanding | “I would love to know what kind of information. I know what kind of information they collected with the device and I’m just curious what they are going to use it for in the future.” (Parent of an older infant) |
| Understanding gained | “[Filling out] this survey, I found that I am pretty lucky that my son is more adaptable. The question, was for example, ‘when you want him to go to bed, he just cried or tantrum’ but he never does that.” (Parent of a preschool-age child) |
| Parent’s enjoyment or satisfaction | “I just like participating in research and helping out the scholars. In my undergrad, I was doing some research and I know how important it is and how hard it can be so...I think it’s good to help.” (Parent of an older infant) |
|                      | “I actually like to spend time with my son. He goes to school every day, so I like to do something with him like the zoom interview. And also I want to show him new technologies.” (Parent of a preschool-aged child) |
| Child’s enjoyment | “I didn’t mind the surveys or anything, and my son loved wearing the LittleBeats. He kept asking if he could put them on. So, I think it captured the kid’s interest too.” (Parent of an early school-aged child) |
|                      | “We had fun doing it [the study], and I think [my son] enjoyed being special, wearing his special shirt for a day.” (Parent of a toddler) |
Study 2

Overview

Building on the key themes of access and usability, privacy, and risks and benefits identified in study 1, we administered a brief survey among a larger sample of parents participating in a different LittleBeats study with children aged 0 to 5 years. Although our main interest was to complement the qualitative findings of study 1 with a quantitative assessment of parents’ perceptions using close-ended rating scales, parents were also able to provide open-ended comments. Therefore, we have also summarized the main themes reflected in these open-ended comments.

Methods

Participants

In study 2, a total of 110 parents (n=108, 98% mothers and n=2, 2% fathers) completed a user experience survey after completing 3 days of LittleBeats recordings at home. Recruitment procedures were similar to those described in study 1. Children (60/110, 54.5% female) were aged, on average, 23.4 months (SD 16.87 mo; range: 2-65 mo) and were identified by parents as Black (n=5, 4.7%), Asian (n=8, 7.5%), White non-Hispanic (n=67, 63.2%), Hispanic (n=15, 14.2%), or >1 race (n=11, 10.4%). Children were first born (n=50, 47%), second born (n=39, 38%), and third or later born (n=17, 15%). Parents were aged, on average, 34.85 (SD 5.01) years, and their highest level of education reported included some high school or high-school degree (4/106, 3.8%), some college or 2-year degree (9/106, 8.5%), a bachelor’s degree (33/106, 31.1%), or an advanced degree (60/106, 56.6%). Parents identified as Black (7/106, 6.6%), Asian (13/106, 12.3%), White non-Hispanic (75/106, 70.8%), Hispanic (8/106, 7.5%), or >1 race (3/106, 2.8%). The mean family income was US $83,250 (SD US $26,470). Of the 110 parents, 4 (4%) were missing responses on the demographic survey but did complete the LittleBeats user experience survey described subsequently.

Ethical Considerations

This study was approved by the institutional review board at the UIUC (protocol #22631).

Procedure

Families were mailed a LittleBeats kit and participated in a Zoom visit, during which a study coordinator walked the parent through the LittleBeats setup and a visit procedure consisting of a baseline assessment of child stress physiology and parent-child interaction tasks (eg, play). At the end of the visit, parents received instructions about completing the day-long recordings and were asked to complete 3 day-long recordings over the course of 2 weeks. Parents also completed a series of web-based questionnaires about family demographics, child behavior, and family functioning. Parent questionnaires were administered either via Qualtrics or REDCap (Research Electronic Data Capture; Vanderbilt University [41,42]) hosted at the UIUC, with the support of the Interdisciplinary Health Sciences Institute and Research IT—Technology Services at the UIUC. Both web-based software platforms are designed to support secure data capture for research studies. Once parents returned the LittleBeats kit by mail, 1 parent in the household (who had been involved in setting up and carrying out the LittleBeats recordings) was asked to rate 5 items about their experience of using LittleBeats, including setting up LittleBeats, along with their perceptions of safety, privacy, and participant burden. Each item was rated on a 5-point scale ranging from 1 (strongly agree) to 5 (strongly disagree). Following each item, parents had the opportunity to add comments or elaborate on their rating. A final open-ended item also asked parents whether there was anything else they would like to share about their experience or anything they would tell someone who was considering joining a LittleBeats study.

Data Analytic Plan

Descriptive statistics, including the frequency distribution, for parental ratings on each of the LittleBeats user experience items were examined. For each close-ended item, we conducted a single-sample t test (2-tailed) to determine whether the mean rating significantly differed from the midpoint of the scale (ie, value of 3=“neutral”). Finally, using the coding scheme developed in study 1, we assessed themes from parents’ responses to the open-ended items.

Results

Parents’ Ratings on User Experience Items

Percentage frequency distributions of parents’ ratings on the user experience items are shown in Table 4. Single-sample t tests indicated a significant difference between the item average (lower ratings indicated greater agreement; higher rating indicated greater disagreement) and the midpoint of the rating scale (3=“neutral”) for 4 (80%) of the 5 items. Compared with a “neutral” response, parents were significantly more likely to agree that (1) the LittleBeats instructions were helpful and clear (mean 1.21, SD 0.41; t109=-45.98; P<.001), (2) they felt comfortable setting up LittleBeats on their child (mean 1.42, SD 0.75; t109=-22.22; P<.001), and (3) they felt their child was safe while wearing LittleBeats (mean 1.33, SD 0.51; t109=-34.48; P<.001). Compared with a “neutral” response, parents were significantly more likely to disagree that they worried about being recorded by the LittleBeats device (mean 3.62, SD 1.06; t108=6.14; P<.001). The final item tapped parents’ perceptions of burden (“I felt that completing LittleBeats recordings for full 3 days was challenging”), and the item average (mean 2.98, SD 1.17) did not significantly differ from “neutral” (t109=-.16; P=.87).
Parents’ Responses to Open-Ended Items

A review of parents’ responses to the optional item to add further comments following each of the rating scales revealed themes that closely mirrored study 1 findings. Regarding the ease-of-use item, 29 (26.4%) of the 110 parents added comments. Most parents noted that having an instruction card included in the kit, as well as a QR code to easily link to the website for more detailed instructions, increased usability.

Regarding comfort in setting up the device, 23 (20.9%) of the 110 parents added comments. Parents noted that they felt comfortable and that the setting up of the device was easy. However, parents also noted that the process of setting up the device was difficult when their child moved around. Other parents mentioned the comfort level of their child (eg, noting that their child felt discomfort when removing the ECG electrodes).

Regarding safety, 25 (22.7%) of the 110 parents added comments. Parents noted few concerns because the device was concealed in a pocket and not easily accessible to the child. Parents who expressed a concern commented on the placement of the device on their child’s chest.

Regarding concerns about being recorded, 32 (29.1%) of the 110 parents added comments. Some parents noted feeling self-conscious about their parenting or other family members’ language choices. Typically, these comments were followed by comments about feeling relieved that the audio would be processed by a machine (vs a human coder). By contrast, many parents explained that they went about their day as usual, which typically contained some sort of sibling argument or other family disagreements.

Regarding participant burden, 68 (61.8%) of the 110 parents added comments. Unlike in study 1, where participants were asked to use the device for 2 days, study 2 participants were asked to use the device for 3 full days (or a total of about 24 hours) over the course of 2 weeks. Several parents commented on their families’ busy schedules and difficulty finding full days when only immediate family members were present.

Finally, a number of parents (46/110, 41.8%) responded to the final open-ended question asking whether they had any other comments they would like to share. Responses mirrored study 1 themes in several respects, including parents’ and children’s enjoyment in participating in the study (eg, “fun and easy” and “I would recommend to my friends”), children’s ability to forget about the device and go about their usual day (eg, “did not interfere with our day”; “[Child] did not notice the device...he was able to nap with it on and so it was really pretty simple to participate!”), and “once the shirt was on, she forgot it was there and so did I!”), and suggestions for ways to minimize burden and improve the experience (eg, adding a display on the device that provides more information about battery charge, power status, and recording length).

Discussion

Summary

Digital health technologies have largely been developed with adults in mind. Interest in and attention to the use of wearable devices among infants and young children, however, has been growing, and data collection using wearable devices provides several advantages over traditional data collection methods, including continuous assessment, greater ecological validity, and the automated detection of behaviors using machine learning algorithms. Given these advantages, combined with rapid technological advances, it is likely that the use of wearables in child development research will burgeon in the coming years. Therefore, assessing how such devices and related data collection protocols are perceived and experienced by parents and their children is critical. User experience studies not only address ethical considerations but can also lead to important changes in research protocols that address parents’ concerns and increase the benefits for future families who participate. Indeed, our mixed methods investigation across 2 studies yielded consistent findings that shed light on parents’ experiences and perceptions of LittleBeats’ usability and safety, the privacy of the audio recordings, and potential risks and benefits of participating in research of this kind. A large majority of parents indicated that device instructions were helpful and clear, the device was easy to use and safe, and remote visits were convenient. Parents’ views about privacy, risks, and benefits were more varied, although, on average, parents reported feeling comfortable with the audio recordings. In summarizing the major themes identified within the major categories, we consider ways in which the findings can inform the future design and implementation of wearable platforms in child development research.
Key Findings

Results across all themes underscored the variability in parents’ (mostly mothers’) perspectives and experiences. With respect to access and usability, some parents expressed interest in having access to information that indicated the cumulative time recorded as well as the battery charge remaining. Such additions to the platform would eliminate parents’ need to track the recording length and minimize parents’ concerns about whether the device was sufficiently charged and recording. Some parents also noted difficulty with threading the ECG lead wires through the back of the shirt or were worried that their child would tug on the wires. These challenges can be remedied by changing the shirt design such that the ECG wires would be more fully integrated into the shirt fabric or design. Although parents indicated that day-long recordings (ie, >8 h/d) were feasible, some parents noted challenges. To alleviate the burden of day-long recordings, the time requirements can be adjusted to be more flexible. For instance, parents can be asked to complete recordings for fewer hours per day across multiple days (ie, 3 to 4 h/d across 4 to 5 d), although the optimal length and frequency of recordings needed to reliably capture the constructs of interest will vary as a function of the research questions being addressed. Importantly, such burdens were balanced by parents’ comments regarding the convenience of remote visit procedures and the ease of using LittleBeats.

Privacy was a theme that also garnered a variety of responses. Some parents indicated few concerns about the privacy of the home audio recordings, whereas other parents worried that the recordings captured private conversations. In the latter case, some families used rules or reminders to control or limit when audio recordings were collected. It is also notable that parents within the same family sometimes expressed differing levels of concern or comfort with the audio recordings. When this pattern emerged, it was largely fathers who voiced concern about invasion of privacy, perhaps because they were not present for initial conversations with the study coordinator, who detailed how the data would be collected and used.

We consider 2 main ways to address parents’ privacy concerns about the home audio recordings (also refer to the study by Cychosz et al [13]). First, providing specific and concrete examples of how the audio recordings are processed and analyzed, perhaps by illustrating a hypothetical example of the data collection, processing, and analysis steps, may help ease privacy concerns. Indeed, some parents noted that the use of machine learning algorithms to analyze the data alleviated their concerns about the audio recordings and privacy-related issues. Thus, describing the machine learning algorithms in a detailed yet accessible manner for nontechnical users and stating ways in which the data will not be used or analyzed (eg, no transcriptions of speech) may help reassure parents. Such information should be provided to all family members participating in the home recordings, including older siblings, and should be presented in various formats (eg, brief informational videos, hard copy pamphlets, interactive web page), along with multiple ways to contact study personnel for questions or comments. As part of this solution and building on some parents’ perspectives that the recordings were just capturing “typical family life,” researchers conducting day-long recordings may also explicitly highlight the family as an important context for development, coupled with appreciation for the fact that all families are different, and that, as researchers, we want to capture what life is like for each family and infant.

A second solution to alleviate parents’ concerns about privacy could involve technological innovations, such as collecting audio recordings in which speech content is not intelligible (refer to the study by Levin et al [33]) or data processing (eg, machine learning algorithms) that occurs on the device or hub in the home so that the audio recordings are not stored or released to the researcher. However, these solutions require further technological advances in audio signal processing and raise issues regarding data-quality assurance. That is, without high-fidelity recordings, the validation and quality checks of machine learning algorithms become difficult. Furthermore, when parents were presented with several hypothetical scenarios for collecting child sensor data in the home environment, parent-reported willingness to participate did not significantly differ between study scenarios in which lower resolution audio data were collected (eg, recording 1-min snippets every 20 min and processing audio data automatically so that raw audio data are not stored) and study scenarios in which higher resolution data (eg, continuous audio recordings) were collected [33]. Taken together, although technological solutions aimed at increasing privacy protection seem to be a reasonable avenue to pursue, future studies on users’ experiences of child wearables, particularly home audio or video recordings, should systematically assess parents’ concerns, needs, and desires when it comes to balancing the privacy of day-long home recordings with the benefits of participation.

Third-party or bystander privacy is also a complex issue [37,38]. In this study, there were two categories of potential third parties: (1) nonparental caregivers or relatives at home who were part of the child’s regular routine and (2) individuals who were not part of the home environment (eg, delivery persons and neighbors). In the first case, nonparental caregivers can be included in the recording if they provided consent. In the second case, the parent would need to turn off the device while the individual is present or change their routine to avoid third parties, which may have consequences for ecological validity. Concerns about third-party recordings can also be resolved by the same types of technical solutions outlined earlier.

The principle of beneficence yielded a variety of responses regarding the risks and benefits of the study procedures. First and foremost, safety was a key theme, and across both samples, parents predominantly expressed views that LittleBeats was safe. When concerns about safety were mentioned, parents often presented hypothetical concerns (eg, the device being close to the skin, the device radiating heat, and the child accidentally falling on the device; the last scenario is mentioned as a potential risk in the parental consent form), which were usually alleviated once the parent learned more about the study. Some parents also mentioned concerns about the child wearing the device during unsupervised times, such as naps, and removed the device during these times. Because infants and young children are much more likely to take ≥1 naps over the course of the day, this subtheme differed across age groups, with parents of children in younger age groups being more likely to mention
device use with respect to nap times. Another set of risks is related to the child’s discomfort, particularly around the application and removal of the ECG electrodes. This potential risk is also mentioned in the parental consent form, and we aimed to ameliorate this risk using latex-free electrodes designed specifically for pediatric populations.

Potential or perceived risks were balanced by parents’ perceived benefits, including increased understanding of their child’s development through the completion of the parent surveys, parents’ satisfaction in contributing to the scientific process, children’s enjoyment of the study procedures (eg, play session with parents), and wearing the novel LittleBeats shirt and device. We note that we did not ask directly about perceived benefits in study 2 close-ended items, although parents in this study did indicate the benefits of participation in the final open-ended question asking whether they had any other comments they would like to share. These responses often paralleled the positive sentiments that study 1 parents expressed. Nevertheless, items that assess the perceived benefits of study participation will be important to include in future studies.

With respect to increasing direct benefits to participants, we gave families personalized books summarizing information that we have collected about their children (eg, height and weight at different ages) in prior studies. Such summaries have been well received and appreciated. Similar types of summaries can be made from data extracted from day-long recordings (eg, frequency and duration of infant babbling or crying). Providing this type of study feedback to parents may also promote effective participant recruitment and retention, particularly among studies that involve high-fidelity data, such as audio recordings. As noted by Levin et al [33], individuals are likely to evaluate intrusiveness and data privacy, on the one hand, and direct benefits to themselves and their children (such as receiving useful, personalized information or feedback from the data collected), on the other hand, when making decisions about whether to participate in such research.

Study Limitations and Future Directions

We note several limitations of our user experience studies. First, we did not ask our older child participants about their experiences directly, although parents reported on a variety of child experiences, including compliance with putting on the device, excitement in wearing the shirt, feeling special while wearing the shirt, and comfort or discomfort. The device hardware was relatively compact and lightweight, and parents reported that children tended to forget about it once it was on. Nonetheless, these reflections clearly highlight the need to directly assess not only parents’ perspectives but also children’s perspectives. Thus, parental reports of their child’s experiences should be augmented by direct observations of infants and younger children while wearing the device as well as interviews with older children. Second, we tracked parents’ reported experiences based on the child’s developmental stage. Similar themes were found across developmental periods, although specific examples of how themes manifested often differed by the child’s age. However, because the subsamples of children in different age groups were relatively small, future research with larger subsamples is needed to more thoroughly investigate developmental considerations related to user experiences in the context of research using child wearables. However, an age-specific consideration that did clearly emerge relates to daytime sleep. Third, in both samples, parents reported high levels of educational attainment. Future research on parents’ perspectives of using child wearable devices in the home setting should include families with diverse demographic characteristics. Including samples characterized by sociodemographic factors in user experience studies is especially critical for child wearables developed for the purposes of mobile health interventions.

Conclusions

Wearable sensors designed for and validated with infants and young children present researchers and clinicians with tremendous opportunities to assess developmental processes and outcomes in more ecologically valid and potentially less burdensome ways than laboratory assessments. Furthermore, LittleBeats’ multiple modalities provide especially rich data to assess an array of constructs central to child development researchers and clinicians, including parent-child vocal turn-taking, regulation of stress, sleep-wake cycles, physical activity, and developmental disorders. At the same time, although we have validated LittleBeats sensors and machine learning algorithms to accurately capture some of these key constructs [17-19,43], the degree to which LittleBeats and similar child wearables deliver benefits (eg, high ecological validity and low burden) will largely depend on acceptance by the end users (eg, parents and children), making user experience studies critical to this research space. In short, if the technology is not acceptable to the end user, it is less likely to be adopted and used as intended. The user experience assessment presented in this paper goes hand in hand with technical validations of the device, and both are critical for successful implementation. The current results suggest that parents predominantly view LittleBeats as easy to set up and use at home, although views regarding privacy and burden were more varied. On the basis of parents’ thoughtful and specific feedback, several concrete changes can be implemented to improve the LittleBeats platform and, ultimately, parents’ and children’s experiences.

Acknowledgments

The authors are grateful to the families who participated in this research and to Jenny Baldwin for her assistance with participant recruitment and data collection. This work was supported by funding from the National Institute on Drug Abuse (R34DA050256), National Institute of Mental Health (R21MH112578), and USDA National Institute of Food and Agriculture (ILLU-793-368).
Conflicts of Interest
None declared.

Multimedia Appendix 1
Parent user experience interview.

References


7.UF


35. Health tools are licensed under a creative commons attribution-non-commercial 4.0 international license 2018-2024. ReCODE. URL: https://recode.health/tools/ [accessed 2024-01-04]


Abbreviations

ECG: electrocardiogram
REDCap: Research Electronic Data Capture
UIUC: University of Illinois Urbana-Champaign

©Nancy L McElwain, Meghan C Fisher, Camille Nebeker, Jordan M Bodway, Bashima Islam, Mark Hasegawa-Johnson. Originally published in JMIR Human Factors (https://humanfactors.jmir.org), 08.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Human Factors, is properly cited. The complete bibliographic information, a link to the original publication on https://humanfactors.jmir.org, as well as this copyright and license information must be included.
Usability and User Experience of an mHealth App for Therapy Support of Patients With Breast Cancer: Mixed Methods Study Using Eye Tracking

Carolin Anders1*, MSc; Preetha Moorthy2*, MSc; Laura Svensson1, MSc; Julia Müller1, MSc; Oliver Heinze1, Dr sc hum; Petra Knaup1, Dr sc hum; Markus Wallwiener3, Dr med; Thomas M Deutsch3, Dr med; Thao-Vy Le3; Lina Weinert1,4, MSc

1Institute of Medical Informatics, Heidelberg University Hospital, Heidelberg, Germany
2Department of Biomedical Informatics at the Center for Preventive Medicine and Digital Health (CPD), Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany
3Department of Obstetrics and Gynecology, Heidelberg University Hospital, Heidelberg, Germany
4Section for Oral Health, Heidelberg Institute of Global Health, Heidelberg University Hospital, Heidelberg, Germany

* these authors contributed equally

Corresponding Author:
Carolin Anders, MSc
Institute of Medical Informatics
Heidelberg University Hospital
Im Neuenheimer Feld 130.3
Heidelberg, 69120
Germany
Phone: 49 622156 ext 34651
Email: Carolin.Anders@med.uni-heidelberg.de

Abstract

Background: Early identification of quality of life (QoL) loss and side effects is a key challenge in breast cancer therapy. Digital tools can be helpful components of therapeutic support. Enable, a smartphone app, was used in a multicenter, prospective randomized controlled trial in 3 breast cancer centers. The app simultaneously serves as a therapy companion (eg, by displaying appointments), a tool for documenting QoL (eg, by enabling data collection for QoL questionnaires), and documentation of patient-reported side effects. The need for digital tools is continually rising. However, evidence of the effects of long-term use of mobile health (mHealth) apps in aftercare for patients with breast cancer is limited. Therefore, evaluating the usability and understanding the user experience of this mHealth app could potentially contribute valuable insights in this field.

Objective: A usability study was conducted to explore how patients with breast cancer receiving neoadjuvant, adjuvant, or palliative outpatient treatment rated their engagement with the app, the user experience, and the benefits of using the app.

Methods: A mixed methods approach was chosen to combine subjective and objective measures, including an eye-tracking procedure, a standardized usability questionnaire (mHealth App Usability Questionnaire), and semistructured interviews. Participants were surveyed twice during the study period. Interviews were transcribed verbatim and analyzed using thematic analysis. Analysis of the eye-tracking data was carried out using the tracker-integrated software. Descriptive analysis was conducted for the quantitative data.

Results: The mHealth App Usability Questionnaire results (n=105) indicated good overall usability for 2 different time points (4 wk: mean 89.15, SD 9.65; 20 wk: mean 85.57, SD 12.88). The qualitative analysis of the eye-tracking recordings (n=10) and interviews (n=16) showed that users found the Enable app easy to use. The design of the app, information about therapies and side effects, and usefulness of the app as a therapy companion were rated positively. Additionally, participants contributed requests for additional app features and suggestions for improving the content and usability of the app. Relevant themes included optimization of the appointment feature, updating the app’s content regularly, and self-administration. In contrast to the app’s current passive method of operation, participants expressed a desire for more active engagement through messaging, alarms, or emails.

Conclusions: The results of this study demonstrate the good usability of the Enable app as well as the potential for further development. We concluded from patients’ feedback and requests that mHealth apps could benefit from giving patients a more
active role (eg, being able to actively document side effects as they occur). Additionally, regular updates of app content could further contribute to encouraging continued use of mHealth apps. Our findings may also assist other researchers in tailoring their mHealth apps to the actual needs of patients undergoing breast cancer therapy.

**Introduction**

**Background**

Breast cancer is the most common type of cancer detected in women in the Western world. One in 8 women will develop breast cancer during her lifetime. In Germany, there are 69,000 new cases per year [1]. The diagnosis is a drastic event in the lives of those affected. Although the mortality rate has decreased in recent years, processing and dealing with the new life situation is a great challenge for patients and their social environment [2]. At the onset of therapy, patients can have a strong desire for education and information. Therefore, providing patients with reliable sources of information and support services is a major and important task for the treatment team. Digitalization in medicine offers great potential for supporting the exchange of information and communication between patients and health care providers [3-5]. These benefits can be realized through the use of mobile health (mHealth) apps, which can encompass several helpful functions for patients, such as the provision of educational materials, appointment or medication reminders, and diaries. For the cohort of patients with breast cancer, many of these mHealth apps are already available or are in development [6]. This cohort also shows a high readiness for using health technology, indicating that mHealth apps are an appropriate means of support in the early phase of breast cancer treatment.

A recent study by Chen et al [7] also found that remote monitoring of symptoms between clinical visits could not only improve patient-provider communication but also prepare patients for subsequent chemotherapy cycles and support symptom management. Within the joint Center for Innovative Care project, a network of 5 university hospitals in southwest Germany, a new mHealth app for patients with breast cancer was developed. This therapy support tool, called the Enable app, aims to combine known benefits of mHealth tools with an innovative reactive assessment of patient-reported outcomes (PROs). It was conceptualized as an iOS or Android mobile app for smartphones and developed by members of the research team with the support of software developers. It includes educational content, information about the side effects of therapies and medications, and information about other support services such as psycho-oncology or nutritional counseling in the form of static text and images. A progress bar illustrates the patient’s individual therapy status in terms of clinical treatment over time (ie, cycles of treatment). In addition to its role as a therapy companion, the app serves as a measurement tool to systematically record patient satisfaction, health-related quality of life (QoL), and patient-reported adverse events. It monitors the neoadjuvant, adjuvant, and follow-up situations in patients with indications for surgery, chemotherapy, radiation, or systemic therapy with primary or metastatic breast cancer. Figure 1 shows exemplary screenshots of the Enable app’s start page, the questionnaire display, and information about treatments. As studies have shown that physicians generally underestimate a large proportion of relevant side effects, patients are empowered to report PRO data and side effects directly through the app. In cases of significant treatment-related deterioration, the care team is alerted, and recommendations are sent to the patient. This more relevant treatment information, in turn, helps improve therapy monitoring, treatment quality, and patient satisfaction [8,9].
The clinical outcomes of the use of the Enable app were studied in the ENABLE randomized controlled trial (RCT). Other research questions addressed in the ENABLE RCT related to improving patients’ adherence to therapy, recognizing and treating critical side effects in a timely manner, and measuring the health-related QoL of different therapy strategies. All study participants underwent QoL assessments at 6 time points during and after adjuvant or neoadjuvant chemotherapy. In the intervention group, an additional short weekly EuroQol Visual Analogue Scale questionnaire was administered. In case of deteriorating results, further screening for side effects was triggered, alerting study staff and enabling immediate contact with the patient to provide support in all phases of breast cancer therapy (reactive PRO assessment). The control group received only the app without the reactive PRO assessment.

The body of scientific literature shows that good usability is an important factor for the success of an mHealth app. More specifically, usability can influence patients’ acceptance and adoption of mHealth [10,11]. Usability is defined by Nielsen [12] as a “quality attribute that assesses how easy interfaces are to use.” According to the International Organization for Standardization (ISO) 9241-1, usability is the “extent to which a system, product, or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use” [13]. Although usability focuses exclusively on the process of using an app or device, user experience involves the users’ subjective feelings that result from the use or anticipated use of a system or a product. For the evaluation of mHealth, both concepts are relevant to obtain a comprehensive view of influencing factors [14].

Good usability can help ensure that the app can be used intuitively by patients and health care providers, which in turn improves compliance and increases the effectiveness of the app. A review by Zapata et al [10] demonstrated the importance of adapting mHealth apps to patients’ needs. Relevant usability themes of similar apps were, for example, streamlining of the navigation paths, a clearer information architecture, or the desire for personalization [15,16]. Recent research has also shown that usability assessment is an essential step in the mHealth app development process [17,18]. It is important to ensure that the app is easy to use for the target group and provides the desired benefits [12]. However, a systematic review by Jongerius et al [6] showed that only 1 of 29 mHealth apps for breast cancer care that were studied in their work underwent and published a usability assessment. To address the aforementioned requirements and achieve sustainable and effective use of the Enable app, the investigation of usability and user experience is indispensable. Therefore, the study presented in this paper intended to gain an understanding of how patients use the app. The aim was to investigate how patients evaluate their engagement with the app, the user experience, and the benefits of using the app. These findings will serve as a basis for further optimization and adaptation of the app to the patients’ needs.

A mixed methods approach provides the opportunity to collect, triangulate, and analyze qualitative and quantitative data,
allowing for the possibility of interpreting the findings from one research approach (ie, qualitative and quantitative) to explain the data generated from the other research approaches. Furthermore, it allows for the use of a qualitative approach to illustrate quantitative findings or the integration of various research approaches to provide a thorough and comprehensive picture of the study [19,20]. Previous studies [15,21,22] have indicated that interviews and usability questionnaires are prevailing methods used for assessing the usability of mHealth apps. However, there are limited studies regarding the real-time capture of users’ visual interactions and the subsequent retrospective analysis of user engagement with mHealth apps through techniques such as eye tracking. Eye tracking, a sensor technology, is used to ascertain an individual’s presence and record their real-time eye movements. This approach is also used to assess the usability of technologies by showcasing decision-making processes through the analysis of eye movement patterns [23,24].

Objectives
Developing new mHealth apps can be time-consuming and requires several iterations of testing and evaluation. The ENABLE project aims to evaluate both the usability and clinical outcomes of the Enable app within the same RCT, which could be a promising approach to speed up development, testing, and planning for further implementation. This paper presents a usability study nested within the ENABLE RCT and following a mixed methods approach incorporating the eye-tracking method. The objective of this usability study was to explore how patients with breast cancer receiving neoadjuvant, adjuvant, or palliative outpatient treatment rated their engagement with the app, the user experience, and the benefits of using the app.

Methods

Study Design
This study was designed following a mixed methods approach combining real-world user experience and standardized observations in a laboratory setting. The study took place at the Department of Obstetrics and Gynecology, Heidelberg University Hospital, Germany.

Procedure

Study Population and Recruitment
The study participants were recruited from the intervention and control groups of the ENABLE RCT patient cohort (German Clinical Trials Register—DRKS ID: DRKS00025611). The ENABLE RCT had the following inclusion criteria: diagnosis of invasive or metastatic breast cancer and planning of neoadjuvant, adjuvant, or palliative therapy in an outpatient treatment setting (indications for surgery or chemo-, radio-, or systemic therapy); minimum age of 18 years; German language skills; and possession of a smartphone with internet access. Owing to technical requirements for eye tracking, patients wearing bifocals were excluded from participation. At study enrollment, patients were asked about their interest in participating in the usability study. All interested patients at the Department of Obstetrics and Gynecology, Heidelberg University Hospital, Germany, received written and verbal information regarding the content and aim of the study and the respective data protection regulations. On the informed consent form, patients could indicate whether they were interested in participating in the usability aspect of the ENABLE RCT. Patients who consented to participate in the nested usability study were contacted individually to schedule appointments for participation following a convenience sampling strategy. No reimbursement was provided. The target sample size was 100 questionnaires, 15 qualitative interviews, and 10 eye-tracking studies. Patient recruitment took place from March 2021 to September 2023.

Instruments
The German translation of the mHealth App Usability Questionnaire (MAUQ) [25] was chosen to quantitatively assess the usability of the Enable app [26]. The MAUQ enables the usability assessment of mHealth apps from the user’s perspective. The MAUQ stand-alone version was formulated to evaluate 3 constructs of usability—ease of use, interface satisfaction, and usefulness—as well as the overall usability score for the app through descriptive statistics. Each of the items of the MAUQ is rated on a Likert scale ranging from 1 (strongly agree) to 7 (strongly disagree), with the overall score ranging from 0 to 100. In addition, the questionnaires were complemented with a set of questions developed by the authors. Newly added questions concerned the use of other mHealth apps, smartphone ownership, sociodemographic information, and a free-text field to be able to describe the study sample more precisely. The target sample size was 100.

In addition to the questionnaire, open-ended, semistructured, and guide-based interviews with patients were conducted to explore their perspectives on the usability of the Enable app. The interviews were conducted by 2 female researchers (CA and LW) with a professional background in health services research and implementation science. Both researchers have profound experience with qualitative interviewing. The interview guide (Multimedia Appendix 1) was developed by a team of health services researchers (LW and JM) based on an extensive literature review and recommendations from the app developers. Afterward, the interview guide was pretested. This study is reported according to the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines (Multimedia Appendix 2 [27]).

Furthermore, to objectively assess how patients interact with the app and identify potential usability issues, an eye-tracking study was conducted. The eye-tracking study was conducted by a usability expert (PM) and a team of health services researchers (CA and LW). A total of 5 tasks were formulated for the eye-tracking study (Multimedia Appendix 3): app log-in, filling in a questionnaire, searching and reading an article, and logging out from the app. To determine the comprehensibility of the tasks, the duration of the study, and the workings of the Enable app, 2 pilot tests were conducted. Following the pilot test outcome, the eye-tracking studies were carried out for 60 minutes with each participant, including the eye tracker setup and the retrospective interview.

The chosen mixed methods approach is designed to systematically collect, cross-validate, and analyze both

https://humanfactors.jmir.org/2024/1/e50926
qualitative data (derived from semistructured interviews and eye tracking) and quantitative data (obtained through the MAUQ). The inclusion of the eye-tracking method in the usability study enriches the capacity to integrate subjective and objective metrics. The qualitative aspect of the eye-tracking analysis enhances the understanding of the user's app perception within the context of individual interactions and app usability. Simultaneously, semistructured interviews enable an assessment of the practicality of integrating the Enable app into daily routines. In contrast, the quantitative data derived from the questionnaire provide precise metrics related to usability measurements.

Hence, the mixed methods approach investigates the why and how aspects through qualitative inquiry, supplementing conventional quantitative and visual data analyses. The fusion of direct observations of user interactions with the app, poststudy retrospective interviews, semistructured interviews, and the usability questionnaire collectively supports the contextualization and comprehensive interpretation of the gathered data.

Data Collection and Analysis

Quantitative Measures

The MAUQ and sociodemographic questionnaire were mailed twice to all patients after inclusion in the RCT. Data collection lasted from May 2021 to October 2022. Study data were collected and managed using REDCap (Research Electronic Data Capture; Vanderbilt University) tools [28] hosted at Heidelberg University Hospital. REDCap is a secure, web-based software platform designed to support data capture for research studies. After completion, all data were exported from REDCap to the R statistical software (version 4.0.4; R Foundation for Statistical Computing). All data were checked for completeness and analyzed by study team members. A descriptive analysis of the questionnaires was performed using R. Means and absolute and relative frequencies were calculated.

Qualitative Measures

Interviews were conducted after participants had used the app for 8 weeks. The interviews took place partly face-to-face at the clinic and by telephone in consideration of current guidelines for preventing infections with SARS-CoV-2 (ie, participants and researchers wore appropriate masks and distance was kept at all times). Nonparticipants were not present during the interviews. No relationship with participants was established before taking part in the study. No repeated interviews were conducted. No field notes were taken. All interviews were audiotaped, pseudonymized, and transcribed verbatim. Transcripts were not returned to participants for verification. Data were transcribed, managed, and analyzed using MAXQDA Standard 2020 (version 20.4.1; VERBI GmbH). After 16 interviews, data saturation was discussed among the researchers. As no new themes emerged in later interviews, the researchers agreed that data saturation had been reached and no additional interviews were necessary. After completion of data collection, thematic analysis of the data was conducted independently by 2 researchers (CA and LW) [29]. First, the researchers reviewed the transcripts independently and identified themes from the literature and the interview guide and inductively from the data. Second, discrepancies were discussed in iterative cycles until a consensus on themes and the final coding scheme was reached. All themes were organized into main themes and subthemes. Each theme was clearly defined by a quote from the interview transcripts (Multimedia Appendix 4). Quantitative and qualitative data were analyzed separately.

For the eye-tracking data collection process, an assigned room where the Tobii Pro Nano (Tobii AB) was installed at the hospital was used; the Tobii Pro Nano is an eye-tracking device specifically designed for small screens, including smartphones. This hardware features a sampling rate of 60 Hz, measures 17 × 1.8 × 1.3 cm, and includes a USB type-A connector. The Tobii Pro Nano was securely affixed to the mobile phone stand, and the Enable app was installed on a smartphone. To facilitate data capture, both the smartphone and the eye tracker were connected to a laptop running the Tobii Pro Lab software (version 1.194) via USB cables. For the purposes of this study, both an Android device (Samsung Galaxy 10, Android version 11) and an iOS device (iPhone 11, iOS version 14.6) were available to users. The choice of smartphone was contingent upon the user’s preferred operating system. The eye tracker recorded the participants’ interactions with the Enable app, such as task completion time, participants’ navigation, gaze plots, and heat maps [30-32]. A heat map was used when fixation duration data were collected [30,31], and a gaze plot was used when location of eye movement data were collected [33,34]. For this study, after the completion of tasks, the study moderators composed post hoc questions pertaining to the interactions, participants’ experiences, and usability issues observed during the procedure. The post hoc questions were discussed with the participants in a short debrief. The debriefing sessions were held to gather direct feedback from participants after interacting with the Enable app, allowing for a deeper understanding of the participants’ behavior and interaction with the app. Through these debriefing sessions, participants could provide context and commentary on their behavior and interaction [35]. Engaging users using post hoc questions, such as using images or live content from recorded sessions, allowed for a better understanding of the real-life context with minimal disruption as it facilitated the recall of situational information prompted by data, sound, or visual imagery.

The data analysis was based on the recordings of the study sessions concurrent with the eye movements of participants. The retrospective analysis involved transcribing participants’ feedback from the audio recordings obtained during the debriefing sessions. Data analysis also included the completion of predefined tasks by the participants, task completion time, and completion status of the tasks. The analysis focused on task performance analysis and the problem analysis of eye-tracking metrics and participants’ feedback.

Ethical Considerations

The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of Heidelberg University Hospital (S-685/2020). All participants provided written informed consent for taking part, audio recording of the interviews, and video recordings during the eye-tracking
procedures. Confidentiality and anonymity were ensured throughout the study. The data was protected against unauthorized access. No incentives or compensation was provided to participants for study participation.

**Results**

**Overview**

The MAUQ was sent to 165 patients recruited from the ENABLE RCT. The response rate was 63.6% (105/165) for the MAUQ at week 4 and 56.4% (93/165) for the MAUQ at week 20. A total of 105 questionnaires for the MAUQ at week 4 (including sociodemographic data) and 93 questionnaires for the MAUQ at week 20 were analyzed. In total, 16 patients were recruited for the interviews, and 10 were recruited for the eye-tracking procedure. The mean duration of the interviews was 25 (SD 7.34) minutes.

**Sociodemographic Characteristics**

The sociodemographic data of the participants in the ENABLE usability study are shown in Table 1, and additional characteristics of the participants regarding smartphone and app use are shown in Table 2. The mean age of all participants (n=105) was 51.3 (SD 10.9) years.

**Table 1. Sociodemographic characteristics of the participants.**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Interview participants (n=16), n (%)</th>
<th>Eye-tracking study participants (n=10), n (%)</th>
<th>Questionnaire participants (n=105), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Woman</td>
<td>16 (100)</td>
<td>10 (100)</td>
<td>105 (100)</td>
</tr>
<tr>
<td>Age group (y)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;30</td>
<td>2 (12.5)</td>
<td>2 (20)</td>
<td>2 (1.9)</td>
</tr>
<tr>
<td>30-40</td>
<td>2 (12.5)</td>
<td>1 (10)</td>
<td>16 (15.2)</td>
</tr>
<tr>
<td>41-50</td>
<td>6 (37.5)</td>
<td>3 (30)</td>
<td>32 (30.5)</td>
</tr>
<tr>
<td>51-60</td>
<td>4 (25)</td>
<td>2 (20)</td>
<td>33 (31.4)</td>
</tr>
<tr>
<td>61-70</td>
<td>1 (6.3)</td>
<td>0 (0)</td>
<td>16 (15.2)</td>
</tr>
<tr>
<td>71-80</td>
<td>1 (6.3)</td>
<td>0 (0)</td>
<td>6 (5.7)</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Academic degree</td>
<td>9 (56.3)</td>
<td>6 (60)</td>
<td>37 (35.2)</td>
</tr>
<tr>
<td>High school education</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>13 (12.4)</td>
</tr>
<tr>
<td>Lower or intermediate secondary school</td>
<td>5 (31.3)</td>
<td>4 (40)</td>
<td>54 (51.4)</td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>2 (12.5)</td>
<td>0 (0)</td>
<td>1 (1)</td>
</tr>
<tr>
<td><strong>Employment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employed</td>
<td>11 (68.8)</td>
<td>9 (90)</td>
<td>66 (62.9)</td>
</tr>
<tr>
<td>Unemployed</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>17 (16.2)</td>
</tr>
<tr>
<td>Studying or vocational training</td>
<td>1 (6.3)</td>
<td>1 (10)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Retired</td>
<td>2 (12.5)</td>
<td>0 (0)</td>
<td>18 (17.1)</td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>2 (12.5)</td>
<td>0 (0)</td>
<td>3 (2.9)</td>
</tr>
</tbody>
</table>
Table 2. Additional participant characteristics on smartphone and app use.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Interview participants (n=16), n (%)</th>
<th>Eye-tracking study participants (n=10), n (%)</th>
<th>Questionnaire participants (n=105), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Use of smartphone (y)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤10</td>
<td>9 (69.2)a</td>
<td>5 (62.5)b</td>
<td>56 (53.3)</td>
</tr>
<tr>
<td>&gt;10</td>
<td>4 (30.8)a</td>
<td>3 (37.5)b</td>
<td>44 (41.9)</td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>5 (4.8)</td>
</tr>
<tr>
<td><strong>Use of other mHealth c apps</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>4 (28.6)d</td>
<td>5 (50)e</td>
<td>33 (31.4)</td>
</tr>
<tr>
<td>No</td>
<td>10 (71.4)d</td>
<td>5 (50)e</td>
<td>71 (67.6)</td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>1 (1)</td>
</tr>
<tr>
<td><strong>Frequency of app use</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daily or several days a week</td>
<td>5 (45.5)f</td>
<td>3 (37.5)b</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Once a week</td>
<td>5 (45.5)f</td>
<td>4 (50)b</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Once a month or less</td>
<td>1 (9.1)f</td>
<td>1 (12.5)b</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

Quantitative Measures

The MAUQ [25] was used to collect quantitative data on the usability of the Enable app. The data were collected at weeks 4 and 20 starting from the baseline of the study. Quantitative data gathered from the MAUQ were analyzed using descriptive statistics. Only complete questionnaires for which the MAUQ score could be calculated were evaluated. Hence, 32.4% (34/105) of incomplete questionnaires collected at week 4 and 29% (27/93) of incomplete questionnaires collected at week 20 were excluded from the analysis. According to Zhou et al [25], the usability of an app is calculated based on the average of the responses to all statements. The higher the overall average, the higher the usability of the app. In this study, the overall usability scores for weeks 4 and 20 were 89.15 (SD 9.65) and 85.57 (SD 12.88), respectively. The mean for each of the subscales from week 4 to week 20 was also calculated and is presented in Table 3. The results show that the usefulness score declined over time from week 4 (80.89) to week 20 (77.33). In addition, the interface and satisfaction score also decreased but not as much as that of the usefulness subscale. The ease of use score, in contrast, remained constant at both weeks 4 and 20.

Table 3. Quantitative analysis of the mHealth App Usability Questionnaire and subscales.

<table>
<thead>
<tr>
<th>Time point</th>
<th>Overall, mean (SD)</th>
<th>Ease of use, mean (SD)</th>
<th>Interface and satisfaction, mean (SD)</th>
<th>Usefulness, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wk 4 (n=71)</td>
<td>89.15 (9.65)</td>
<td>92.41 (11)</td>
<td>91.6 (10.15)</td>
<td>80.89 (15.67)</td>
</tr>
<tr>
<td>Wk 20 (n=66)</td>
<td>85.57 (12.88)</td>
<td>92.27 (11.91)</td>
<td>88.23 (13.6)</td>
<td>77.33 (16.63)</td>
</tr>
</tbody>
</table>

Qualitative Measures

Interviews

In total, 527 text passages were coded during the interviews. A total of 9 themes and 60 subthemes were identified, each of which could still be categorized under the superordinate themes of preconditions for app use, usability, and reflection. These themes are summarized in Figure 2.
Preconditions

Initial Expectations
As an opening question in the interview, patients were asked about their initial thoughts when they first heard about the Enable app. The most frequently mentioned expectations were related to the quality of information. Patients expected the information in the app to be updated regularly, understandable, and in line with the latest research. Another expectation was that the app would provide contemporary therapy support and be perceived as modern, including replacing printed brochures. Patients expected the app to provide guidance over the course of therapy, contact options, and easy access to relevant information. Approximately half of the participants had neutral expectations for the app:

Yes, I already thought that it [the app] would support me through everyday life and therapy, that I can also use it to organize myself a bit. [Interview 9; transcript position 2]

Onboarding
The aspect of the onboarding process was not part of the interview guide. However, individual participants reported that they felt well supported by the study staff at the beginning of their app use. Even if they were initially overwhelmed by the app or experienced technical difficulties, participants expressed that they received the necessary support and were able to handle the app:

Oh dear, now I have to dig into yet another app. I don’t know if I can handle it. But the more I got a grip on it, the better it worked. [Interview 3; transcript position 2]

Usability

Presentation and Design
Patients were asked to describe their impressions of specified design aspects. Overall, patients were content with the color scheme and perceived it as pleasant without being boring or flashy. For some patients, this cheerful esthetic contributed to a sense of joy when using the app and encouraged them to use it more often:

The design is very friendly. Very beautifully visualized. I always enjoy opening the app. It is also well designed, you always have the feeling that it is not draining in any way, it is more playful with all these images and visualizations. I find it very very clear. [Interview 11; transcript position 34]

Regarding the layout of the written information, patients appreciated how the most relevant parts were highlighted through the positioning of boxes. The font size, design, and structuring of the information were seen as adequate. The selection of accompanying images was described as empathetic and not too explicit. The app included a personalized visual representation of the therapy progress. This display was also rated as clear and useful. Patients explained that the presence of this display motivated them:

I found this progress bar, which shows me how long I will be in therapy for, especially beautiful. It...motivated me, showing me that there is always a path forward and that the therapy will soon be over. [Interview 13; transcript position 43]

App Interaction
Regarding usability, 4 important aspects emerged while interacting with the app. Neither the positioning nor the design of the app icons were perceived as entirely intuitive. However, patients grew acclimated to the icons, and thus, this did not further impede usability:

Yes, the icons that were down in this bar. In the beginning, I didn’t know the meaning of each icon. But when I took a closer look once, I knew it for the next time. [Interview 7; transcript position 39]

Log-in and log-out procedures were described as easy and quick and did not pose any problems for the patients in this study. Most patients had no issues working with the app’s structure. They could easily navigate within the app and were able to find what they were looking for:

I found my way around the app really quickly. I haven’t tried all the features yet, I haven’t clicked on everything because I don’t need it all. But I have always been able to find the things that I wanted very quickly, and everything is right there when you click on it. [Interview 11; transcript position 44]
Overall, patients liked using the app as it enabled them to access information on the go. Patients described that having their smartphones with them at all times allowed them to read information given the absence of other electronic devices such as laptops or tablets. However, a few patients mentioned the additional benefits of having a web-based version of the Enable app.

**Use Patterns**

This code encompasses descriptions of how and when patients used the app. Most patients experienced changes in the frequency of app use. In the beginning, they used the app often, and some patients used it multiple times per day:

*In the beginning, shortly after my diagnosis, I had a lot of questions—for my physicians, how things work and so on. During this time, it (the app) really helped me a lot.* [Interview 10; transcript position 24]

Over time, use declined. This development was mostly due to lower demand for support and information as patients became used to therapy proceedings. Patients also used the app less as they felt that they had already read everything.

After this initial phase, patients reported using the app whenever they needed to look up appointments, had free time (eg, during waiting times before physician’s appointments), had or experienced new side effects from their treatment, were prescribed new medications, or were prompted by push notifications:

*I always used it shortly before my [chemotherapy] appointments. Or when I had questions regarding diet and exercise. And sometimes there were questionnaires I had to fill in. And yes, as soon as the app said “there is news,” I opened it...And to look up times for my appointments.* [Interview 10; transcript position 12]

**Satisfaction**

Patients praised the general aspects of the app and liked the idea of having a digital tool accompanying them throughout their therapy; for example, the app provides a good overview of relevant topics, especially at the beginning of the disease. Except for 1 interviewee, all participants (15/16, 94%) would recommend the app to others:

*...because it really provides a great overview...because so many aspects are addressed. Not only the type of therapy, but also just different things about cancer. Especially at the beginning these keywords—Yes, these terms in the boxes from tiredness to fatigue and polyneuropathy and different things.* [Interview 7; transcript position 47]

**Reflection**

**Added Value of App Use**

When asked about the concrete benefits of the app in everyday life, several aspects were mentioned. The most important aspect for the participants was the information on therapies and side effects, which was perceived as helpful, especially in the initial phase of therapy. The quality of the information was praised as the app’s information was considered understandable and its origin was considered reliable:

*You feel informed, you feel—that gives you a form of security, because you say to yourself: Well, if I have the information from here [the app], then it was completely clear to me: I don’t have to look it up again. That’s true for me because these are reliable information providers who wrote this.* [Interview 12; transcript position 81]

The comprehensibility and language level were also perceived as adequate. Statements on the amount of information were heterogeneous according to individual information needs. However, the amount of information was predominantly perceived as sufficient in the context of the app. Furthermore, the appointment display, contact information, and progress bar were found to be helpful and clear. With regard to the contact information provided in the app, the fact that it was easy to find was rated positively.

Some patients reported that the questionnaires in the app gave them a positive feeling as they reflected on their condition and (in the intervention group) it was experienced positively that the questionnaires were read by the study staff and that staff could react proactively to them if necessary. Overall, patients perceived the app as a good therapy companion that guided and supported them through the various phases of the disease and therapies.

**User Appraisal**

Users’ opinions on the existing functions and features of the app were added to this category. Most patients complained about the appointment display as the date and time on the app did not always correspond to the actual clinic appointments (eg, in the case of last-minute postponements):

*It’s a shame that the—I don’t know how the appointments displayed in the app, how often those are matched. I’ve had frequent differences there. Especially when appointments had to be postponed, the chronology was no longer correct for me.* [Interview 9; transcript position 2]

Regarding the quantity of information, some patients wished for more in-depth information or links to other information platforms. It was remarked that the amount of information available varied depending on the topic. Regarding the quality of the content, patients noted that the listed side effects or drugs were grouped differently. For instance, the patients were unable to locate paclitaxel as it belonged to the taxane drug class. In total, 12% (2/16) of the patients in particular perceived errors in spelling, punctuation, and grammar as distracting. The presentation of the contact information on the app was described as difficult to find, especially in emergencies. The additional pop-up notifications of the app updates were rated negatively as it was not apparent to the user what exactly was new in the app. Furthermore, respondents ascertained that the menu navigation was not intuitive enough and, therefore, needed to be improved.
**Recommendations**

Statements about features of the app that are not yet offered were classified as recommendations or wishes. Most wishes were mentioned in relation to the appointment display. Patients would like to have additional information about appointments, such as directions, a reminder function, the ability to export appointments from the app to private calendars (eg, Google or Outlook calendars), or the ability to make appointments directly from the app. The desire for self-administration (ie, areas such as appointments, questionnaires, or therapy progress that can be actively managed by the patient) was also frequently voiced. In addition, some patients wished to view the questionnaires that had already been completed to be able to monitor their condition over the course of therapy:

> With the exception of filling in the questionnaires, you can’t work with the app yourself. Therefore, if you could manage things in the app by yourself, then of course I would think that would be great. [Interview 9; transcript position 2]

Patients also wanted the content of the app to be updated, expanded, and adapted to new scientific findings. In this context, there was a desire for more explanatory videos to be included in the app. Patients also suggested that the app should offer more information about current and upcoming clinical trials for patients with breast cancer. To see what content in the app has already been read, patients suggested a read status, where content that has already been read is highlighted. Emergency contacts should also be highlighted in the app to make them easier to find, for example, by displaying them on the home page:

> Especially the emergency numbers, I don’t know how to get something like that into the app, but that might be an idea, because I’ve been looking a lot for the right contact person. Maybe that would also be something that you could highlight a little bit or display as a button. [Interview 4; transcript position 32]

To be able to find certain topics more quickly, the need for a search function was mentioned several times. Furthermore, to improve the readability of the content, patients would like to be able to adjust the font size. It was also suggested that the app could be used on other devices, such as tablets.

**Eye Tracking**

*Overview*

The analysis of the data collected from the eye-tracking recordings as well as the retrospective interviews showed that the participants found the app easy to use. We observed that most participants completed the given tasks, although the time taken to complete a few tasks proved to be challenging. On the observations and retrospective interviews during the eye-tracking study, we discovered 3 noticeable patterns related to the design and layout of the app, content and navigation through the app, and additional features the participants would like to have in the app. Figure 3 shows exemplary heat maps from the eye-tracking analysis. The data collected during the task performance, such as the task completion rate and task completion times, are provided in Multimedia Appendix 5.
dates and the progress of a questionnaire, at the top of the screen. This finding indicates that patients expect important information to be located at the top of the app’s layout. Furthermore, the patients actively mentioned that the retrievability and visibility of the questionnaire were low. Although the questionnaires were available on the home screen of the app, patients believed that the questionnaires were available on the menu. In contrast, patients found the overall layout of the app to be acceptable.

Content and Navigation of the App

Regarding the content of the app, patients showed more interest in the titles of the articles (eg, topics such as symptoms or side effects) than in the images displayed. When asked during the retrospective interviews, patients mentioned that they did not pay attention to the images as they provided no information on what the article was about. Patients preferred to read the title of the article as it gave them information about its content, as shown by the red areas of the heat maps in Figure 3. Moreover, many participants explored the app to find the right information or icon to perform the tasks. However, this correlates with how frequently patients used the app. During the interviews, some patients said that they used the app frequently, for example, every day, to read articles on side effects or symptoms and fill out questionnaires regularly, whereas some patients used the app frequently at the start phase of the ENABLE RCT and later minimized the use of the app except to fill out questionnaires. The data showed that patients also had issues navigating through the app, especially related to the task of finding a specific article. Analysis of the recorded data of the participants’ navigation and gaze plots from the Tobii Eye Tracker showed that patients looked for a search function. Most patients clicked the menu icon; however, they did not proceed further to find the article nested under the Symptoms category on the menu. In addition, some patients searched for the article on the start page along with the other articles already displayed.

“Would Like to Have” (Wishes)

Participants identified a need for additional features in the Enable app as a consequence of the challenges they encountered during the eye-tracking study tasks. These suggested features were considered as nice-to-have options and were based on the specific problems faced by the participants during the study. The first was the availability of an option to mark an article as a favorite and be able to view the favorite article on the start page. Second, patients desired to have more articles or information about the symptoms and side effects of breast cancer and its treatments. Third, the icon currently representing contact information for health care providers (My Care Team) was misleading. Patients preferred to have another icon that indicates contact or communication as this would enable them to contact the study nurses more quickly. Finally, a search option was suggested by all participants.

Discussion

Principal Findings

The aim of this study was to investigate how patients with breast cancer rated their engagement with the Enable app, the user experience, and the benefits of using the app. In particular, the design, layout, navigation, content, and requests for new features were identified as important outcomes of interest for evaluating the app and further improving it to meet user needs. The interviews provided valuable suggestions for optimizing the app and the implementation process. The design and color scheme were rated very positively overall. In terms of use patterns, it was noticeable that the frequency of app use decreased over the therapy period.

Patients found the app easy to navigate. However, there was some criticism that the menu icons were not intuitive enough, especially at the onset of use. Perceived benefits were discussed extensively in the interviews. Patients found the information on therapies and side effects very useful. The appointment display and progress bar were also found to be helpful and motivating. At the same time, the appointment display was most often criticized, and it was the feature for which there were the most recommendations for change (eg, to be able to manage appointments autonomously in the app or set reminders). In terms of content, it was mentioned that there was a lot of information on some topics and not enough on others. Patients also wanted more content updates within the app (eg, on current topics such as the COVID-19 pandemic) and a search function to access specific content.

A study by Ansaar et al [36] showed that nearly 78% of all usability evaluation studies in their systematic review used a questionnaire-based method. However, using mixed methods approaches in usability evaluation studies provides benefits such as the possibility to balance the advantages and disadvantages of the different methods. Moreover, by applying the mixed methods approach, both subjective and objective aspects can be combined to assess usability [36]. In many aspects, such as the navigation, recommendations, and perceived benefits codes, the results of the different survey methods support each other. However, the interviews and eye-tracking study sometimes provided different findings. For example, the importance of images within the app was positively highlighted in the interviews. In contrast, the eye-tracking study and retrospective interviews revealed that images played a subordinate role for patients, with titles being more important for finding relevant content in the app. Although participants reported in the interviews that they were able to navigate easily within the app and find the content they were looking for, we observed in the eye-tracking study that there were difficulties with finding specific content. Furthermore, the interview inquiries primarily centered on the practicality of incorporating the Enable app as a follow-up intervention in daily life. Meanwhile, the use of eye-tracking technology allowed for direct, real-time observation of user behavior while engaging with the app through task performance. Despite patients reporting the ability to regularly use the app without difficulty, the eye-tracking study’s direct observation unveiled valuable insights into their actual use patterns within their everyday routines. In this context, disparities between the results obtained from the 2 methods emerged, possibly stemming from users’ lack of awareness regarding any issues until they were prompted with specific inquiries.
Comparison With Prior Work

Our results on the MAUQ indicate good usability. The results for the total scale showed that usability decreased from weeks 4 to 20. A decrease in usability over time has also been observed in previous studies [37-39]. Possible explanations for this decline in our study can be found in the interviews, indicating that the extent of app use also decreased over the course of therapy. Patients found the app to be particularly advantageous at the start of their therapy because of their great need for information. However, as they gained more knowledge about the disease and its treatment, their demand for information decreased. In addition, patients reported that the app lost its appeal once all the available articles had been read, often leading to a desire for new content to be added. Patients also expressed a need for additional features or improvements as they continued to use the app. As a result, the decrease in the app’s usability score could be attributed to patients perceiving it to be less useful after an extended period of use owing to the lack of content updates and unmet desires.

Looking more closely at the subscales of the MAUQ, *usefulness* had the lowest score compared with *ease of use* and *user interface and satisfaction*. These items assess whether the app is helpful and useful for patients’ health and well-being. This relationship is also apparent when looking at the *usage patterns* category from the interview analysis. It appears that patients are less likely to use the app because of the lack of new content. This is consistent with the findings of other studies on mHealth apps for patients with breast cancer [16,40,41]. As an implication for similar apps for other chronic conditions, it seems important to update the app content on a regular basis to provide patients with an incentive to continue using the app as well as strengthening patients’ satisfaction and information needs. Consistent with the findings from the interviews and eye-tracking study, only the *ease of use* subscale remained almost stable over the duration of app use.

In the context of other usability studies on mHealth apps, the importance of paying more attention to the user group of older adults is emphasized. The different age ranges of patients and the different levels of technical affinity for older patients are mentioned as possible factors causing usability problems. Some studies emphasize that these factors are often overlooked and need to be considered when developing mHealth apps [42,43]. In our study, these aspects were less evident. With an average age of 51 years, our study participants do not represent a predominantly older population but are close to the German population average for women, which is 46 years [44]. In contrast, the study participants were also far below the average age of 64 years for patients with breast cancer. Therefore, further research on app development and usability with a focus on older participants should be conducted to more adequately represent the typical population of patients with breast cancer.

Considering the preferred device for using the *Enable* app, most participants were content with using the app on their smartphones. However, there were isolated requests to be able to increase the font size of the content and use the app on a larger-scale device, such as a tablet or PC. This issue was also mentioned by participants in a usability study by Jessen et al [45], in which an mHealth app for self-management of chronic diseases was evaluated.

Although the onboarding process was not part of the interview guide, some patients actively recalled how they were introduced to the app as well as how they perceived the technical onboarding process. The patients did not experience issues with these steps and reported being content with the process, mostly because of the strong support of the study team. Previous research has pointed out that complex registration and log-in procedures can be perceived as especially cumbersome by patients and can lead to stopping app use [46-48]. Our study identified the strong interpersonal connection with and continued support from the study team as a positive influence on the perceived ease of onboarding. This support took place in the context of a research study and is not viable in a real-world implementation. However, the issue of technical support arose exclusively during the qualitative interviews. We did not collect any quantitative data on this topic. Thus, further streamlining of the onboarding process while being mindful of health care workers’ limited time resources should be an area for future research.

Strengths and Limitations

The chosen mixed methods approach can positively support the further development of the app. The expansion of the classic social science method spectrum to include technical methods such as eye tracking made it possible to combine the subjective patient perceptions reported in interviews and questionnaires during everyday use with objective measurements under laboratory conditions.

However, the integration of qualitative results and the objective measurement from the eye-tracking procedure introduced discrepancies. As noted previously, interviewees appreciated the use of images in the app, whereas eye-tracking results showed that more time was spent on the article titles than on the images. Another example is that the interviews and the questionnaire produced good ratings of usability, but the eye-tracking study showed that patients found it difficult to find defined content. Although difficult to analyze, these discrepancies are common in mixed methods studies [19]. In our study, these discrepancies could be explained by methodological differences. For example, reading a title naturally takes longer than glancing at an image, leading to a long fixation time. Therefore, this result does not allow for the conclusion that titles are more important than images. Here, the qualitative interviews were helpful in interpreting this finding. Regarding the second example—overall good usability scores in comparison with eye-tracking times—several interpretations appear plausible. First, it is possible that social desirability led patients to rate the usability more favorably in both the interviews and the questionnaire. Consequently, the objective measure via eye tracking revealed that usability was worse than in subjective measures. Second, the setting of the eye-tracking procedure (eg, unusual or uncomfortable sitting position, being observed by ≥2 researchers, or using a different device) could have led to changed patterns in (app use) behavior. Although we acknowledge these discrepancies, we conclude that the mixed methods approach and its results deepened the understanding...
of the studied topic and produced valuable insights, with discrepancies leading to vigorous and fruitful discussions among the researchers.

However, the generalizability of the study results is limited by several factors. To ensure that patients with lower digital health literacy could participate in the quantitative data collection without constraints, we decided to use printed surveys sent by mail. Patients returned them at their discretion. Hence, it cannot be verified whether the surveys were filled out at the correct time. In addition, some values were missing from the returned surveys, and manual data entry could have led to documentation errors. Incomplete or inconclusive questionnaires had to be completely excluded from the analysis as it was not possible to calculate the score. Although all necessary steps were taken to ensure high-quality and reliable data (eg, data entry was always checked by another researcher), using a web-based survey instead of a printed survey could have made data collection easier, faster, and more reliable. These trade-offs have to be balanced in future research projects.

This study population contained an above-average proportion of academics, especially among the subgroups of interviewees and eye-tracking study participants. This should be taken into account when interpreting these results. A systematic review by Niazkhani et al [49] showed that patients with lower educational attainment and limited health literacy were less likely to intend to use an electronic patient health record and were more likely to use it ineffectively. Moreover, previous experience with computers or health technology has been associated with increased acceptance, and acceptance increases with higher education [7]. Although these results refer to electronic health records, they indicate that this aspect should be further investigated in future studies. Given the median age at breast cancer diagnosis of 64 years and the relatively younger median age of this study cohort, conclusions from this study must be interpreted with caution as they may not represent the views and digital literacy of older women with breast cancer [50].

The Enable app was developed specifically for patients with breast cancer. Consequently, our study sample included only female patients with breast cancer. Some of our results and recommendations may have limited generalizability to other patient populations. Nevertheless, we think that aspects such as the relevance of content updates, the accuracy of displayed appointments, or the intuitiveness of the app navigation might also be relevant beyond the target group. This should be verified in further research.

As part of the ENABLE RCT, reasons for dropping out were documented where available. These reasons were examined to see whether there were any indications of usability problems. A small proportion of the included study participants in the RCT dropped out because of physical exertion or feelings of being overwhelmed by the app. In this respect, further research is needed to understand how patients in later stages of the disease or with greater disease burden perceive the usability and benefits of the intervention. Furthermore, mHealth apps should be designed to be usable and helpful for these patient groups as well, especially in the context of patients living with cancer. As the mean age of participants in this study was relatively low, it can be assumed that there is a risk of selection bias. It is possible that younger patients decided to participate in the study and use the app because of a higher affinity for smartphones [11].

In addition, using the eye-tracking device led to further limitations. Potential participants in the eye-tracking study had to undergo an additional screening process to exclude patients wearing bifocal glasses. Although patients were recruited for the study, this criterion did not allow us to cast a wider net for the participant recruitment process. Furthermore, we also had the challenge of asking patients to sit still so that the eye-tracking data could be captured without breaks. However, this request is generally against the natural way in which users sit and interact with mobile devices. Another point to note is that the execution of the tasks on the app by the patients was deviated as the tasks were presented on paper and this retracted some of the gaze points of the patients. This is, in general, a common problem when tasks are not integrated into mobile apps during development for testing purposes.

Conclusions

The results of this usability study demonstrate good usability of the studied app and potential for purposeful development. The design and color scheme were rated very positively overall. However, there was some criticism that the menu icons were not intuitive enough, especially at the onset of use. Noticeably, the frequency of app use decreased over the therapy period. Perceived benefits of the app were information on therapies and side effects. The appointment display and progress bar were also found to be helpful and motivating. Still, participants offered recommendations for changing the appointment display (eg, to be able to manage appointments autonomously in the app or set reminders). In terms of content, it was mentioned that there was a lot of information on some topics and not enough on others. Patients also wanted more content updates within the app (eg, on current topics such as the COVID-19 pandemic) and a search function to access specific content. The interviews and eye-tracking study revealed valuable suggestions for improvement as well as requests for additional app features. An important point is that the app currently provides information to the patient mainly passively. The patients’ wishes indicate that the app needs to be further developed so that they can actively enter information into the app and work with it. The overlap between decreasing usability and decreasing usefulness also suggests that the app needs to be regularly updated with new content to maintain its usefulness over time. These findings will be incorporated into the further development of the Enable app. We concluded from patients’ feedback and requests that similar mHealth apps could benefit from giving patients a more active role (eg, being able to actively document side effects as they show up instead of being prompted to do so). In addition, regular updates to app content (eg, adding new informational pieces) could further contribute to and, thus, encourage the continued use of mHealth apps.
Acknowledgments

This study was funded by the Baden-Württemberg (Germany) Ministry of Science, Research, and Arts under reference 42-04HV.MED (19)/15/1 as part of the Center for Innovative Care project. The authors would like to thank all patients in this study for their participation and valuable input. They would also like to thank Corinna Hartmann for her support in recruiting patients for this study. Thank you to Bendix Harms for his support with the quantitative analysis.

Data Availability

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

Authors’ Contributions

CA, PM, and LW collaborated on the draft of the manuscript. PK contributed. MW is the principal investigator of the ENABLE project. LW, JM, and PM were responsible for the study design and protocol, and OH contributed. MW and TMD prepared and submitted the study protocol, and LW, JM, and PM contributed. LW, JM, and PM collaborated on the construction and testing of the interview guides and the quantitative data collection tools, and LS supported the finalization of these instruments. CA and LW conducted the interviews, analyzed transcripts, and interpreted the interview data. PM, CA, and LW conducted the eye-tracking study and retrospective interviews, TL contributed. PM analyzed and interpreted the eye-tracking data. PM analyzed and interpreted the survey data, and CA and LW supported data interpretation. OH contributed to the acquisition of funding. All the authors provided substantial comments and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File, 27 KB - humanfactors_v11i1e50926_app1.docx ]

Multimedia Appendix 2
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 480 KB - humanfactors_v11i1e50926_app2.pdf ]

Multimedia Appendix 3
Eye-tracking tasks.
[DOCX File, 18 KB - humanfactors_v11i1e50926_app3.docx ]

Multimedia Appendix 4
Definition of themes.
[DOCX File, 34 KB - humanfactors_v11i1e50926_app4.docx ]

Multimedia Appendix 5
Task performance data.
[XLSX File (Microsoft Excel File), 15 KB - humanfactors_v11i1e50926_app5.xlsx ]

References


Abbreviations

- COREQ: Consolidated Criteria for Reporting Qualitative Research
- ISO: International Organization for Standardization
- MAUQ: mHealth App Usability Questionnaire
- mHealth: mobile health
- PRO: patient-reported outcome
- QoL: quality of life
- RCT: randomized controlled trial
- REDCap: Research Electronic Data Capture

©Carolin Anders, Preetha Moorthy, Laura Svensson, Julia Müller, Oliver Heinze, Petra Knaup, Markus Wallwiener, Thomas M Deutsch, Thao-Vy Le, Lina Weinert. Originally published in JMIR Human Factors (https://humanfactors.jmir.org), 05.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Human Factors, is properly cited. The complete bibliographic information, a link to the original publication on https://humanfactors.jmir.org, as well as this copyright and license information must be included.
Original Paper

Feasibility and Design Factors for Home-Based Pulmonary Rehabilitation of Patients With Chronic Obstructive Pulmonary Disease and Chronic Lung Diseases Based on a People-Object-Environment Framework: Qualitative Interview Study

Shih-Ying Chien¹,², PhD; Alice May-Kuen Wong³, MD; Winston Tseng³, PhD; Han-Chung Hu⁴,⁵, MD; Hsiu-Ying Cho⁵, MA

¹Department of Industrial Design, Chang Gung University, Taoyuan, Taiwan
²Department of Physical Medicine and Rehabilitation, Chang Gung Memorial Hospital, Taoyuan, Taiwan
³Division of Community Health Sciences, School of Public Health, University of California, Berkeley, CA, United States
⁴Department of Thoracic Medicine, Chang Gung Memorial Hospital, Taoyuan, Taiwan
⁵Department of Respiratory Therapy, Lin-Kou Chang Gung Memorial Hospital, Taoyuan, Taiwan

Corresponding Author:
Shih-Ying Chien, PhD
Department of Industrial Design
Chang Gung University
259 Wen-Hwa 1st Rd, Kweishan
Taoyuan, 333
Taiwan
Phone: 886 3 2118800 ext 5669
Email: shihying@mail.cgu.edu.tw

Abstract

Background: The feasibility of implementing home-based pulmonary rehabilitation (PR) can be assessed from the perspectives of patients with chronic lung disease and health care professionals involved in PR.

Objective: Through a qualitative inquiry using interviews and the adoption of the people-object-environment framework, this study aims to understand the influences of interpersonal, environmental, and situational factors on the perceptions and considerations of individuals involved in home-based PR for patients with chronic lung disease.

Methods: One-on-one interviews were conducted with 20 patients with chronic lung disease and 20 health care professionals for investigating their attitudes and opinions based on their experiences regarding home-based PR as well as for identifying the key factors affecting the benefits and drawbacks of such therapies. This study further evaluates the feasibility of using digital tools for medical diagnosis and treatment by examining the technology usage of both parties.

Results: The 4 key issues that all participants were the most concerned about were as follows: distance to outpatient medical care, medical efficiency, internet connectivity and equipment, and physical space for diagnosis and treatment. Interviews with patients and health care professionals revealed that the use of technology and internet was perceived differently depending on age and area of residence. Most participants reported that digital tools and internet connectivity had many benefits but still could not solve all the problems; moreover, these same digital tools and network transmission could lead to problems such as information security and digital divide concerns. This study also emphasizes the significant impact of human behavior and thinking on shaping the design of health care interventions and technologies. Understanding user perspectives and experiences is crucial for developing effective solutions for unmet needs.

Conclusions: The results of this study indicate that despite the different perspectives of patients and health care professionals, their considerations of the key issues are very similar. Therefore, the implementation of plans related to telemedicine diagnosis, treatment, or rehabilitation should take the suggestions and considerations of both parties into account as crucial factors for telehealthcare design.
Introduction

As the third leading cause of morbidity and mortality worldwide, chronic obstructive pulmonary disease (COPD) is a significant public health issue [1-4]. In 2019, the number of individuals diagnosed with COPD exceeded 328 million worldwide [5-8]. A significant correlation between physical activity and lung function [9-12] emphasizes the importance of regular exercise for individuals with COPD who require pulmonary rehabilitation (PR) [13-19]. However, patients with COPD often report reluctance to engage in physical activities due to dyspnea, the effects of which include chronic cough, exacerbations, reduced exercise capacity, and impaired quality of life [20-25]. PR is a tailored and comprehensive intervention conducted via thorough assessment of the patient. In individuals with chronic pulmonary diseases, the primary objective of the pulmonary intervention is to improve not only their overall health but also their psychosocial well-being in the long term [26-28]. Typically, PR programs are customized for personal symptomatic conditions [29-31]; hence, PR interventions entail tailored exercises and educational sessions aimed at enhancing activity tolerance, mitigating symptoms, and augmenting skills that aid in managing chronic respiratory diseases [31,32]. The majority of PR treatments usually require one-on-one sessions and the assistance of a therapist [33-35]. However, the one-on-one care approach is limited due to shortages in health care personnel, elevated work-related stress, and prolonged working hours [36]. Moreover, when the COVID-19 pandemic hit, lockdowns and personnel restrictions forced the interruption of PR for many patients with chronic lung disease, which posed a threat to their lives [37-40].

Due to the COVID-19 pandemic, telehealth has become increasingly attractive owing to its functionality, importance, and prospects [41,42]. In addition to reducing human contact and easing the burden on health care workers, telehealth leverages technology communication and transmission to alleviate the workload of respiratory therapists and improve the accuracy of respiratory rehabilitation records [43-45]. Using telehealth, patients can undergo rehabilitation at home and be monitored remotely by medical personnel [46,47]. Home-based PR can also mitigate the difficulties of outpatient care for patients living in remote areas and those with physical disabilities [48-50]. Furthermore, it can be used as an auxiliary means of physical PR to assist in self-management and precisely modify behavior, thereby reducing hospitalization and medical costs [51,52].

Traditional PR usually relies on one-on-one human monitoring through observation or physiological monitors to examine a patient’s health condition. Remote health care has the advantage of prescribing home-based PR, enabling patients who are unable to leave their homes due to physical conditions such as disability or living in rural areas to partake in rehabilitation programs at home [53,54]. However, there are also many limitations and considerations of remote health care, as follows:

1. Lack of security and limited interpersonal interaction: The safety of patients is the primary concern of clinical physicians [55,56]. The biggest challenge of home rehabilitation is emergency treatment, which has been the main hurdle for remote health care since many years [57]. In addition, remote therapy can only provide limited physical and mental assessments [58,59]. Due to the lack of face-to-face interpersonal interactions, patients may develop loneliness, helplessness, and frustration, which may reduce the effectiveness of treatments and the speed of recovery [60,61].

2. Privacy and security issues: Most remote health care is performed through network transmission. Many clinical physicians believe that network transmission may lead to data leakage or theft of medical records or personal information of patients [62,63].

3. Technological and equipment limitations: The implementation of remote health care requires specific technological equipment such as smartphones or computers with network functions. However, for many remote users or special groups such as older persons, lack of equipment, poor network communication quality, or unfamiliarity with network-related technology hinder utilization [64].

4. Insurance payment limitations: Different regions or countries have different standards for remote health care services. Therefore, many insurance companies do not have a remote health care reimbursement system or only cover specific services [65-68].

Despite its limitations and by taking people, object, and environment into consideration, telemedicine remains a valuable tool for the provision of health care services, especially for patients who have difficulty visiting medical facilities in person or those affected by infectious diseases and related restrictions such as lockdowns and quarantine. Telemedicine enables uninterrupted treatment and continued assistance for patients in their recovery. However, in establishing a home-based PR, it is essential to consider the various environments of participants to effectively maximize the benefits of this medical service.

Methods

Ethics Approval

This study was approved by the institutional review board of Chang Gung Memorial Hospital (approval 202200070BO). The participants were patients with chronic lung disease and respiratory health care professionals who had provided written informed consent from both urban and rural areas. Due to the COVID-19 pandemic, all one-on-one interviews were conducted by videoconferencing.
Participants and Procedures

The 20 patients recruited for the interviews included those who had participated in PR programs and those who had not. During the interviews, the patients provided insights into the implementation of PR programs from a patient-centric standpoint. All interviewees had a medical history of 5 years or more.

The 20 respiratory health care professionals included registered thoracic surgeons, respiratory therapists, physical therapists, and PR specialists. Most of these professionals had experience treating patients with chronic lung disease and had participated in designing exercise prescriptions, patient tracking and monitoring, and disease progression research in PR programs. Furthermore, the majority of the interviewees had treated a specific proportion of patients with chronic lung disease within the past 3 years (Table 1).

Table 1. Characteristics of the health care professionals (n=20).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (male:female)</td>
<td>8:12</td>
</tr>
<tr>
<td>Age (years), min-max; mean (SD)</td>
<td>27-65; 46 (11)</td>
</tr>
<tr>
<td>Experience in pulmonary rehabilitation (years), mean (SD)</td>
<td>5.2 (8.47)</td>
</tr>
<tr>
<td>Type of health care professional in pulmonary rehabilitation, n (%)</td>
<td></td>
</tr>
<tr>
<td>Thoracic surgeons</td>
<td>3 (15)</td>
</tr>
<tr>
<td>Respiratory therapists</td>
<td>12 (60)</td>
</tr>
<tr>
<td>Physical therapists</td>
<td>2 (10)</td>
</tr>
<tr>
<td>Pulmonary rehabilitation specialists</td>
<td>3 (15)</td>
</tr>
</tbody>
</table>

Prior to the interviews, all participants were required to complete a survey questionnaire, which included demographic information and details of their use of smart devices and the internet. Daily use was defined as regular usage. The patients provided information about their pulmonary disease status, duration of illness, and a self-assessment of their health status (on a 5-point scale ranging from excellent to poor) as well as recalled their activity frequency over the past 7 days. Health care personnel were required to answer questions related to their primary clinical responsibilities. Each participant took part in a 1.5- to 2-hour interview session conducted by the primary author, who was also a clinical researcher and an assistant professor affiliated with the Chang Gung Medical Foundation. In-depth interviews were primarily used to collect the data. After collecting the interview data, all identifiable personal information was removed from the transcripts. The data were then coded, organized, and analyzed using NVivo 12.0 software (Lumivero) for qualitative data analysis. Accurate and detailed data interpretation, the transcripts were provided to the interviewees for review and cross-checked with relevant researchers to confirm the accuracy of data interpretation.

Results

Characteristics of the Participants

This study consisted of 40 participants: 20 health care professionals specializing in PR and 20 patients with chronic lung diseases. The background characteristics of the 20 health care professionals are shown in Table 1; nearly 60% (12/20) were respiratory therapists, and the remaining health care professionals were pulmonary surgeons, physical therapists, and rehabilitation physicians. Their mean age was 46 years, and all had more than 3 years of experience in PR and treatment (mean 5.2 years). The background characteristics and activity habits of the 20 patients interviewed are shown in Table 2; the majority of the patients had COPD (12/20, 60%), and 25% (5/20) were lung transplant recipients. The majority of the participants (15/20, 74%) had never participated in a PR program, and 70% (14/20) of the patients rated their physical condition as poor. Regarding exercise over the past week, 65% (13/20) of the patients chose a 10-minute walk as their exercise indicator, followed by strength training (5/20, 25%). Notably, 55% (11/20) of the patients reported preferring to sit rather than stand and to stand rather than move.
Table 2. Characteristics of the patients (n=20).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (male:female)</td>
<td>10:10</td>
</tr>
<tr>
<td>Age (years), min-max; mean (SD)</td>
<td>51-85; 68 (9.8)</td>
</tr>
<tr>
<td>Participation in pulmonary rehabilitation programs, n (%)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>5 (26)</td>
</tr>
<tr>
<td>No</td>
<td>15 (74)</td>
</tr>
<tr>
<td>Chronic lung diseases, n (%)</td>
<td></td>
</tr>
<tr>
<td>Chronic obstructive pulmonary disease</td>
<td>12 (60)</td>
</tr>
<tr>
<td>Asthma</td>
<td>3 (15)</td>
</tr>
<tr>
<td>Lung transplantation</td>
<td>5 (25)</td>
</tr>
<tr>
<td>Self-assessment of their health, n (%)</td>
<td></td>
</tr>
<tr>
<td>Excellent</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Very good</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Good</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Fair</td>
<td>5 (25)</td>
</tr>
<tr>
<td>Poor</td>
<td>14 (70)</td>
</tr>
<tr>
<td>Exercise frequency and quantity over the past week, n (%)</td>
<td></td>
</tr>
<tr>
<td>I have engaged in high-intensity strength training, including aerobic exercise, fast cycling, and swimming.</td>
<td>2 (10)</td>
</tr>
<tr>
<td>I have participated in moderate physical activities such as stretching exercises and flexibility training.</td>
<td>5 (25)</td>
</tr>
<tr>
<td>I have walked for at least 10 minutes every day.</td>
<td>13 (65)</td>
</tr>
<tr>
<td>What statement best characterizes my exercise habits? n (%)</td>
<td></td>
</tr>
<tr>
<td>Given the option, I will opt to sit rather than stand.</td>
<td>11 (55)</td>
</tr>
<tr>
<td>I frequently require standing but not for the purpose of lifting heavy objects.</td>
<td>5 (25)</td>
</tr>
<tr>
<td>Climbing slopes and stairs is a common necessity for me.</td>
<td>4 (20)</td>
</tr>
<tr>
<td>I often transport heavy objects and engage in manual labor.</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

Survey Results

In order to better examine the potential of telehealth, we conducted a survey targeting contemporary electronic communication tools, specifically computers, cellphones, and tablets, which were widely utilized by both patients and health care professionals, as shown in Figure 1. The majority of the patients and health care professionals reported using desktop computers as their most frequently used electronic communication device, constituting the largest proportion at 40% (8/20), followed by smartphones at 30% (6/20). Notably, health care professionals reported a higher usage rate (by 10%) of tablet computers compared to patients. Note that neither group of participants reported habitually using laptops. Overall, the patients were less proficient with technology compared to the health care professionals, which is a crucial determinant in implementing telehealth programs.
People-Object-Environment Framework

This study adopts the people-object-environment framework as the focal point for the interview investigation to improve the understanding of the feasibility of home-based PR. This study analyzes the advantages and disadvantages of remote PR in the current context, with the aim of bridging the gap between ideal use and reality (Table 3). Activities involving various elements such as individuals, entities, and environmental factors often result in the emergence of diverse concerns among different participants. Through the analysis presented in Table 3, we identified the gaps in home-based rehabilitation services from the perspectives of health care professionals and patients. Subsequently, this facilitated a thorough discussion of potential solutions to meet the needs and expectations of all involved parties. Our research findings reveal that the use of telehealth for home-based PR programs had both advantages and disadvantages. Using a people-object-environment framework to analyze the results, we describe 4 dimensions: reduced time and transportation constraints to access medical care; improved medical efficiency; changes in equipment, network, and physical space; and information transmission security, about which health care professionals particularly raised concerns in telehealth.
### Table 3. Analysis of the pros and cons of telerehabilitation with the people-object-environment framework.

<table>
<thead>
<tr>
<th>People (patient)</th>
<th>Object (patient)</th>
<th>Environment (patient)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pros</strong></td>
<td><strong>Cons</strong></td>
<td><strong>Pros</strong></td>
</tr>
<tr>
<td><strong>People (health care professionals)</strong></td>
<td><strong>Bridging communication barriers:</strong></td>
<td><strong>Reducing health care burden:</strong></td>
</tr>
<tr>
<td>Promoting health care access in remote areas: Telehealth facilitates convenient health care services, which enhance medical care for patients in remote areas and promote community health.</td>
<td>Telehealth reduces physical interactions and social contact between patients and physicians, which may have long-term effects on patients’ psychological and social well-being.</td>
<td>Telehealth reduces the health care burden for long-term patients or those requiring regular follow-ups; this minimizes the time and effort associated with transportation and waiting as well as provides cost-effective health care options.</td>
</tr>
<tr>
<td>Boosting patient involvement: Telehealth enables interactions with health care professionals via web-based platforms, providing medical information and guidance as well as fostering active engagement of patients in their own health management.</td>
<td><strong>Operational and communication barriers:</strong> Older or technologically inexperienced patients may encounter difficulties in understanding instructions from remote health care professionals through telehealth.</td>
<td>Dealing with cross-infection risks: Telehealth minimizes contact between patients and health care professionals, thereby lowering the risk of cross-infection and promoting the health and safety of both health care professionals and patients.</td>
</tr>
<tr>
<td><strong>Cons</strong></td>
<td><strong>Pros</strong></td>
<td><strong>Cons</strong></td>
</tr>
<tr>
<td><strong>Object (health care professionals)</strong></td>
<td><strong>Advantages of telehealth:</strong> Telehealth provides a convenient health care model, particularly beneficial for regions facing constraints related to time, geographical location, and transportation. The utilization of basic computer equipment enables the provision of medical consultations, making health care more accessible and efficient for patients in such areas.</td>
<td><strong>Digital health care infrastructure:</strong> Telehealth accelerates the transmission and exchange speed of health care information, leading to improved overall health care efficiency.</td>
</tr>
<tr>
<td>Wireless transmission: Wireless transmission significantly reduces the workload of health care professionals and makes health care services more efficient by transitioning from a one-on-one service to a one-to-many format.</td>
<td>Digital divide: Individuals who are from lower socioeconomic backgrounds or have limited access to digital resources may face barriers to participation in telehealth due to the lack of appropriate technological equipment or internet connectivity. This highlights inequalities in the distribution of health care resources.</td>
<td><strong>Disparity in health care resources between urban and rural areas:</strong> Despite the convenience of telehealth for remote consultations, operational difficulties may still exist for areas lacking proper equipment.</td>
</tr>
<tr>
<td>Technological dependency: Users with limited technological skills or resources may encounter difficulties in operating telehealth, which requires adequate knowledge of technology, suitable equipment, and stable internet connectivity.</td>
<td><strong>Health care quality and patient experience:</strong> Although telehealth provides convenient remote health care options for certain diseases or conditions, in-person consultations or measurements from medical instruments may offer more accurate health care services.</td>
<td>Equipment and infrastructure requirements: Telehealth relies on high-speed internet and appropriate equipment, which can still pose challenges in certain rural areas.</td>
</tr>
<tr>
<td><strong>Environment (health care professionals)</strong></td>
<td><strong>Equipment and infrastructure requirements:</strong> Telehealth relies on high-speed internet and appropriate equipment, which can still pose challenges in certain rural areas.</td>
<td><strong>Environmental limitations:</strong> Home environments often impose spatial constraints that may limit various rehabilitation, diagnostic, and treatment activities.</td>
</tr>
</tbody>
</table>
Due to the absence of physical contact, telehealth may not provide opportunities for face-to-face contact with patients, which can make it challenging for physicians to conduct comprehensive physical examinations or assessments. Limitations in comprehensive treatment: Some diagnoses and treatments may require physical contact and assistance from specific equipment, which cannot necessarily be substituted by telehealth.

Both patients and health care professionals acknowledged the significant benefits of telehealth in addressing the challenges of distance in accessing medical care. Reductions in travel and wait times due to telehealth allowed patients to actively participate in their health care decision-making through web-based platforms. The digitization of medical records for better disease management was also facilitated. For residents in remote areas, telehealth eliminated geographical barriers to health care, promoted more efficient allocation of medical resources, and prevented the closure of regional hospitals and health care institutions.

### Dimensions

#### Dimension 1: Distance, Time, and Transportation Issues

Both patients and health care professionals acknowledged the significant benefits of telehealth in addressing the challenges of distance in accessing medical care. Reductions in travel and wait times due to telehealth allowed patients to actively participate in their health care decision-making through web-based platforms. The digitization of medical records for better disease management was also facilitated. For residents in remote areas, telehealth eliminated geographical barriers to health care, promoted more efficient allocation of medical resources, and prevented the closure of regional hospitals and the “medical deserts” phenomenon. Nevertheless, health care professionals identified potential risks and quality-of-care issues associated with telehealth. Due to the absence of physical interactions, physicians were limited to relying on surface-level symptoms for diagnosis. Comprehensive physiological examinations and physical evaluations were also limited due to the lack of suitable equipment, which could in fact jeopardize the safety of critically ill patients who require urgent or emergency treatment. Moreover, prolonged social isolation resulting from telehealth may have adverse effects on patients’ psychological and social well-being.

#### Dimension 2: Enhancing Medical Efficiency Through the Use of Telehealth

Health care professionals described that the popularization of telehealth was due to its advantages in improving the efficiency of disseminating medical information. Through online platforms, health care professionals can access patients’ medical history instantaneously and collaborate to provide optimized treatment. This approach reduced constraints on patients’ time and space, expanded the scope of medical services, reduced the workload of health care professionals, and transformed the traditional one-on-one respiratory treatment mode into a one-to-many model. Health care professionals also noted that the advantages

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expansion of health care service areas: Through telehealth, physicians can diagnose and treat patients remotely without being limited by geographical location while providing real-time medical services.</td>
<td>Lack of physical contact: Telehealth may not provide opportunities for face-to-face contact with patients, which can make it challenging for physicians to conduct comprehensive physical examinations or assessments.</td>
</tr>
<tr>
<td>Expansion of professional scope: Telehealth enables physicians to engage in remote meetings and collaborations with other health care experts, enhancing medical efficiency.</td>
<td>Limitations in comprehensive treatment: Some diagnoses and treatments may require physical contact and assistance from specific equipment, which cannot necessarily be substituted by telehealth.</td>
</tr>
<tr>
<td>Enhancement of diagnosis and treatment efficiency: Telehealth reduces time and space limitations between physicians and patients, improving the efficiency of the overall health care services.</td>
<td></td>
</tr>
<tr>
<td>Increased convenience: Telehealth offers patients greater convenience, particularly for those residing in remote areas or facing mobility challenges, thereby reducing the time and costs associated with hospital visits.</td>
<td></td>
</tr>
<tr>
<td>Enhanced diagnostic and treatment capabilities: Through remote imaging and information sharing, physicians can access additional support and assistance, which improve diagnostic accuracy and treatment outcomes.</td>
<td>Technical requirements: The use of telehealth requires stable internet connectivity and appropriate device support, which may be challenging for users who are not familiar with technology.</td>
</tr>
<tr>
<td>Improvement of health care resource utilization: Telehealth aids physicians in managing and allocating regional health care resources more effectively, enhancing utilization efficiency and reducing unnecessary health care costs.</td>
<td>Medical responsibility and risk management: Telehealth may involve issues of medical responsibility and risk management, such as misdiagnosis, treatment errors, or incomplete medical records, which may result in medical disputes and litigation. Physicians and health care institutions need to ensure compliance with relevant medical responsibility and risk management principles in telehealth as well as maintain a high level of medical practice.</td>
</tr>
<tr>
<td>Reducing reliance on physical space: Telehealth reduces the need for physical space such as clinics and hospitals, thereby lowering costs and burdens associated with facilities and resources for health care institutions.</td>
<td>Health care security and privacy risks: The use of telehealth may involve health care security and digital privacy risks such as patient identity verification and medical record protection. Physicians should exercise caution in handling such issues.</td>
</tr>
</tbody>
</table>

### Table

<table>
<thead>
<tr>
<th>People (patient)</th>
<th>Object (patient)</th>
<th>Environment (patient)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expansion of health care service areas: Through telehealth, physicians can diagnose and treat patients remotely without being limited by geographical location while providing real-time medical services.</td>
<td>Enhancement of diagnostic and treatment capabilities: Through remote imaging and information sharing, physicians can access additional support and assistance, which improve diagnostic accuracy and treatment outcomes.</td>
<td>Reducing reliance on physical space: Telehealth reduces the need for physical space such as clinics and hospitals, thereby lowering costs and burdens associated with facilities and resources for health care institutions.</td>
</tr>
</tbody>
</table>
of data and imaging arising from telehealth actually improved diagnostic and treatment efficiency, which could as a result achieve precision medicine. However, health care professionals also raised concerns about telehealth. The mode of transmitting medical information through data still harbored many risks and considerations such as diagnostic and treatment errors. Additionally, injuries (such as falls or respiratory distress) that could occur during treatment raised issues related to medical responsibility and risk management. Therefore, handling patient identity verification and medical records with caution when administering telehealth is crucial in order to safeguard patient privacy during medical treatments.

**Dimension 3: Leveraging Internet Connectivity and Device-Based Solutions for Rehabilitation and Health Monitoring**

Patients and health care professionals both identified that digital health care had the potential to significantly reduce medical wait times and enhance efficiency, which complemented the services of regional hospitals, lowered the medical burden of chronic patients, and eliminated limitations due to transportation, geography, and time. However, many challenges still remain with the use of digital tools for rehabilitation and monitoring systems. Most patients who require PR are older, aged ≥65 years, and unfamiliar with digital devices and networks, and they often lack the knowledge and understanding of how to install and configure such devices and applications. In addition, remote areas lack stable networks, technology, and equipment. This digital divide inhibits a subset of patients in certain areas from fully utilizing relevant medical services, highlighting the problem of an uneven distribution of medical resources.

**Dimension 4: Advantages and Disadvantages of Converting Medical Spaces**

The changing medical environment has brought many advantages to patients and health care institutions through telehealth, particularly during the COVID-19 pandemic, by reducing hospital-acquired infections and patients’ reliance on medical space and resources, thereby alleviating the burden of health care costs. However, as previously mentioned, some health care professionals reported that not all diagnoses and treatments could be properly conducted remotely due to the availability or operation of equipment and the limitation of patients’ home space and environment, which restrict the implementation of many treatment regimens. All the 4 dimensions supported by quotes from patients and health care professionals are shown in Table 4.
This study explores the perspectives and barriers of respiratory health care professionals and patients toward telehealth for rehabilitation for respiratory diseases. Based on the participant interviews, the use of home-based telerehabilitation for patients with lung diseases was perceived to have both advantages and disadvantages, which could be categorized under 4 domains: location, digital technology, internet connectivity, and physical space requirements. Unlike previous research [69], we adopted the people-object-environment framework and interviewed patients as well as health care professionals, with the aim of obtaining feedback from all participants in the same context. The 4 aspects mentioned above were found to be the most important concerns for health care institutions and patients.

### Table 4. Verbatim quotes supporting the main dimensions by patients and health care professionals.

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Patient</th>
<th>Health care professional</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dimension 1: Distance to outpatient care</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>…During the pandemic, being able to have online consultations reduced a lot of my stress. I heard many of my friends around me were infected while at hospital. [Female, 48 years old]</td>
<td>…Telemedicine has reduced a lot of transport-related issues. Through online connections, we can access all of the patient’s data and make more accurate assessments. [HCP #12]</td>
</tr>
<tr>
<td>Negative</td>
<td>…I live in a very rural area where there are no taxis, so every time I see a doctor, I have to take four different buses. The journey alone takes me over three hours, so I avoid seeing a doctor if I can help it. [Female, 64 years old]</td>
<td>…To be honest, not every patient is suitable for telemedicine. For example, older patients may have difficulty understanding what I ask them to do. Also, some patients’ conditions cannot be determined solely by questioning and require examination using medical instruments and devices, so it is difficult for me to make a diagnosis without a proper examination. [HCP #8]</td>
</tr>
<tr>
<td><strong>Dimension 2: Medical efficiency</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>…Because I get breathless when I walk, I try not to go out if I don’t have to. I’m also afraid of falling when I go out, and I don’t want to bother my children. So if I can see a doctor through a computer, I prefer that. [Male, 66 years old]</td>
<td>…The biggest advantage of telehealth is saving a lot of time and manpower. Of course, this refers to medical work that is more repetitive and lower risk. But I hope that in the future, online systems will have warning functions that can quickly let me know which patient has an issue that needs special attention. [HCP #3]</td>
</tr>
<tr>
<td>Negative</td>
<td>…I would rather see a real doctor. Just talking on the phone doesn’t give me a feeling that I’ve really seen a doctor. [Female, 72 years old]</td>
<td>…To be honest, although the internet is convenient, I feel that its effectiveness is sometimes limited. Perhaps respiratory therapy needs to be divided into stages, and not every stage is suitable for being done at home. It may need to be classified/graded. [HCP #17]</td>
</tr>
<tr>
<td><strong>Dimension 3: Internet connectivity and equipment</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>…Every time I go out to see a doctor, I’m always in a rush and get so nervous that I forget to ask the doctor any questions. By seeing the doctor through a computer, I have more time to chat with the doctor. [Female, 70 years old]</td>
<td>…The internet connection is very convenient. As long as the health insurance card is inserted, all the patient’s information can be accessed. Telehealth has not only changed a patient’s medical treatment mode but also prevented many regional hospitals from closing down. [HCP #11]</td>
</tr>
<tr>
<td>Negative</td>
<td>…To be honest, I don’t really understand the internet. If no one helps me set it up, I won’t know how to see a doctor online. And if the doctor doesn’t see me, how will they know what’s wrong with me? [Male, 76 years old]</td>
<td>…Sometimes, the reason why I cannot wait for a patient is because the foreign caregiver has not set up the computer properly. When communicating with the patient through the computer, sometimes the elderly cannot understand, and it is also difficult to communicate with the caregivers. If I were there in person, I could still teach them how to do it. [HCP #8]</td>
</tr>
<tr>
<td><strong>Dimension 4: Space for diagnosis and treatment</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>…I am old and unable to move around easily. It would be best for me to see the doctor at home. [Male, 86 years old]</td>
<td>…A patient receiving online medical care at home will require much less space for us, such as waiting rooms and registration areas. It will also significantly reduce the demand and burden on staff. [HCP #1]</td>
</tr>
<tr>
<td>Negative</td>
<td>…There are many things that I cannot do at home. I need to have my blood pressure measured, but there is no one to help me at home. Also, I like to chat with people, but at home, there’s only me. [Female, 70 years old]</td>
<td>…Online medical care now is quite good, with many complete functions such as registration, appointment progress, and electronic medical records. However, I personally have reservations about having many medical records stored in the cloud, as there is no absolute security. Also, if a patient falls at home, how to allocate responsibility and the risks involved are also concerns. [HCP #14]</td>
</tr>
</tbody>
</table>

aHCP: health care professional.

**Discussion**

This study explores the perspectives and barriers of respiratory health care professionals and patients toward telehealth for rehabilitation for respiratory diseases. Based on the participant interviews, the use of home-based telerehabilitation for patients with lung diseases was perceived to have both advantages and disadvantages, which could be categorized under 4 domains:
Previous studies have reported that distance to outpatient care has a profound impact on chronic patients; in other words, the farther away from home, the lower is a patient’s willingness to seek medical care [70]. However, Bhatt et al [71,72] highlighted that patients often exhibit a reluctance to engage in PR, regardless of proximity, for several reasons. Both groups of interviewees in our study reported that the provision of alternative options would reduce the number of hospital visits, which would benefit patients and health care institutions. Although digital health care has its limitations, leveraging the internet to expand regional hospital services is not only beneficial to the public but also makes medical services more effective [73]. In special circumstances such as the outbreak of a pandemic, issues such as patients being unable to attend in-person treatments cannot be ignored. It is undeniable that digital health care, in particular, spawns numerous benefits in such situations. However, there are still limitations to digital health care for people (doctors or the public), equipment and network, and the environment (urban or rural). For example, most older patients feel that only consultations in person with doctors generate the feeling of being treated. Furthermore, without physical examinations, it could be difficult for doctors to diagnose the cause of symptoms. Nonetheless, digital technology remains a good choice for patients with respiratory disease who do not want to venture outside or exercise; however, not every patient’s home is equipped with remote medical devices or equipment, especially in rural areas [74]. Most patients with respiratory disease are also older, and without caregiver assistance, operating such devices can be difficult. In addition, due to limited professional knowledge, equipment, network, and living space, home care cannot replace all hospital diagnoses and rehabilitation. Therefore, in consideration of the findings from our study and a previous study [75], the implementation of hierarchical medical care requires that patients first undergo video consultations and then be referred to nearby medical institutions for appropriate treatment based on the severity of their condition. Patients can be referred to a larger medical center when necessary for treatment through an electronic referral platform between institutions. This approach not only effectively improves the utilization efficiency of medical resources but also significantly reduces medical expenses and transportation costs for patients.

This study has some limitations. First, as our study was conducted in a specific health care institution, our findings may not be generalizable to other regions or institutions. Second, most patients had poor health conditions and no prior experience with PR; thus, they relied only on limited experience and information. Lastly, the majority of the participants were older, which may have influenced their responses regarding computer use or internet issues. Additionally, health care service needs likely vary between urban and rural areas, and our study does not distinguish between the challenges and differences in home-based PR between these 2 types of areas. Future research should consider this aspect in their study design.

Most study participants reported that telehealth could greatly benefit patients with chronic pulmonary diseases; however, these benefits were not without limitations. Reflections on these limitations by patients or health care professionals revealed that telehealth is not suitable for all patients. For example, diagnosis and treatment via telehealth can only accomplish certain tasks and merely serve as a tool for preliminary diagnostic assessments. Nonetheless, preliminary assessments can determine whether a referral to a regional hospital or a large teaching hospital is necessary. This classification and referral system will also be applicable to rehabilitation therapy. Not all patients are suitable for home-based PR, considering patient safety, the required space and equipment, or the need for further precision testing, among other factors. In addition, although telehealth brought many conveniences to patients and health care professionals, both parties still faced significant psychological pressure. Patients noted that digital medicine lacked warmth, and they tended to prefer human care, while physicians had doubts about medical decision-making without the ability to perform physical examinations. The degree of control over digital technology was also an issue. Both parties lacked confidence that effective treatment could be achieved solely through the internet. Even though digital care has the advantage of long-term patient monitoring, some patients were unfamiliar with internet devices, and health care professionals were concerned that patients may not always respond correctly to instructions. Moreover, patients often neglect their physician’s advice (such as following prescribed exercise schedules) due to lack of motivation and the need to physically meet with the physician. Both groups of study participants indicated that significant improvements in telerehabilitation technology were still needed, particularly for patients in rural areas or those who were older and living alone, who require more support and services.

Acknowledgments

This research was supported by the National Science and Technology Council (grant 111-2222-E-182-002-MY2). We also thank all the interviewees who provided insight and expertise that greatly assisted this research.

Conflicts of Interest

None declared.

References


45. Radzikowska U, Ding M, Tan G, Zhakparov D, Peng Y, Wawrzeniak P, et al. Distribution of ACE2, CD147, CD26, and other SARS-CoV-2 associated molecules in tissues and immune cells in health and in asthma, COPD, obesity, hypertension,


Abbreviations

COPD: chronic obstructive pulmonary disease
PR: pulmonary rehabilitation
Perspectives on the COVID-19 Vaccination Rollout in 17 Countries: Reflexive Thematic and Frequency Analysis Based on the Strengths,Weaknesses,Opportunities, and Threats (SWOT) Framework

Vanja Kopilaš1*, PhD; Khrystyna Nasadiuk2*, MD, PhD; Lucia Martinelli3, PhD; Lenka Lhotskà4,5, PhD; Zoran Todorović6,7, MD, PhD; Matjaz Vidmar8,9, PhD; Helena Machado10, PhD; Anna Lydia Svalastog11,12, PhD; Srećko Gajović13, MD, PhD

Department of Psychology, Faculty of Croatian Studies, University of Zagreb, Zagreb, Croatia
2Department of Biochemistry, Danylo Halyský L’viv National Medical University, Lviv, Ukraine
3Science in Society Program, MUSE-Science Museum, Trento, Italy
4Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, Prague, Czech Republic
5Faculty of Biomedical Engineering, Czech Technical University in Prague, Prague, Czech Republic
6University Hospital Medical Center "Bežanijska kosa", Belgrade
7Department of Pharmacology, Clinical Pharmacology and Toxicology, Faculty of Medicine, University of Belgrade, Belgrade
8Institute for the Study of Science, Technology and Innovation, The University of Edinburgh, Edinburgh, United Kingdom
9School of Engineering, The University of Edinburgh, Edinburgh, United Kingdom
10Institute for Social Sciences, University of Minho, Braga, Portugal
11Østfold University College, Halden, Norway
12Centre for Research Ethics and Bioethics, Uppsala University, Uppsala, Sweden
13BIMIS-Biomedical Research Center Šalata, University of Zagreb School of Medicine, Zagreb, Croatia

* these authors contributed equally

Corresponding Author:
Srećko Gajović, MD, PhD
BIMIS-Biomedical Research Center Šalata
University of Zagreb School of Medicine
Šalata 4
Zagreb, 10000
Croatia
Phone: 385 1 4566 948
Email: srecko.gajovic@hiim.hr

Abstract

Background: As the SARS-CoV-2 virus created a global pandemic and rapidly became an imminent threat to the health and lives of people worldwide, the need for a vaccine and its quick distribution among the population was evident. Due to the urgency, and on the back of international collaboration, vaccines were developed rapidly. However, vaccination rollouts showed different success rates in different countries and some also led to increased vaccine hesitancy.

Objective: The aim of this study was to identify the role of information sharing and context sensitivity in various vaccination programs throughout the initial COVID-19 vaccination rollout in different countries. Moreover, we aimed to identify factors in national vaccination programs related to COVID-19 vaccine hesitancy, safety, and effectiveness. Toward this end, multidisciplinary and multinational opinions from members of the Navigating Knowledge Landscape (NKL) network were analyzed.

Methods: From May to July 2021, 25 completed questionnaires from 27 NKL network members were collected. These contributors were from 17 different countries. The responses reflected the contributors’ subjective viewpoints on the status and details of the COVID-19 vaccination rollout in their countries. Contributors were asked to identify strengths, weaknesses, opportunities, and threats (ie, SWOT) of the respective vaccination programs. The responses were analyzed using reflexive thematic analysis, followed by frequency analysis of identified themes according to the represented countries.
Results: The perspectives of NKL network members showed a link between organizational elements of the vaccination rollout and the accompanying societal response, both of which were related to strengths and weaknesses of the process. External sociocultural variables, improved public communication around vaccination-related issues, ethical controversies, and the spread of disinformation were the dominant themes related to opportunities and challenges. In the SWOT 2×2 matrix, Availability and Barriers emerged as internal categories, whereas Transparent communication and promotion and Societal divide emerged as key external categories.

Conclusions: Inventory of themes and categories inspired by elements of the SWOT framework provides an informative multidisciplinary perspective for effective implementation of public health strategies in the battle against COVID-19 or any future pandemics of a similar nature.

(JMIR Hum Factors 2024;11:e44258) doi:10.2196/44258

KEYWORDS
SARS-CoV-2 virus; COVID-19 vaccination; pandemic; hesitancy; safety; vaccination; COVID-19; tool; implementation; vaccine hesitancy; effectiveness; sociocultural; communication; disinformation

Introduction
On March 11, 2020, the World Health Organization declared the COVID-19 pandemic [1]. COVID-19 first appeared in December 2019 in Wuhan, China. This disease, caused by the SARS-CoV-2 virus, led to an unprecedented challenge for health institutions that required most countries to integrate their efforts to globally mitigate the spread of the disease [2-4]. Various policies to control the spread of the virus have been adopted in different countries. Some of them were drastic, such as national lockdowns, as well as initiating the widespread use of individual protection devices and means [5]. The individual protective measures included recommending frequent hand washing and application of sanitizers, maintaining social distance, and mandatory wearing of face masks or respirators. However, even a simple measure of covering the face proved to have psychological, cultural, religious, and behavioral implications at both the individual and communal levels [6]. Moreover, the policies aimed to stop the spread of the virus impacted the psychological well-being of the population [7]. These implications should be considered in public campaign strategies aimed at achieving effective public consent toward the adoption of protective measures.

The publication of the genetic sequence of SARS-CoV-2 on January 11, 2020, resulted in the explosion of comprehensive studies on the virus and stirred global research and development activity to develop vaccines against the virus [8]. To accelerate this work, next-generation vaccine technology platforms have been deployed and the first COVID-19 vaccine candidate entered human clinical testing as early as March 16, 2020. In December 2020—in record time and following collaborative efforts of the global scientific community, pharmaceutical companies, and governments—several types and brands of vaccines based on different technologies and mechanisms of action became available for mass deployment [9].

The importance of mass vaccination has been established in the context of previous epidemics such as in the case of smallpox eradication and the incidence reduction of measles and polio [10]. The goal of mass vaccination programs is to interrupt person-to-person disease transmission by surrounding infected people with a high proportion of vaccinated individuals who have developed protective antibodies against the infection (ie, reaching “herd immunity”) [11]. Public health experts have prioritized increased vaccination delivery with the hope to resume socioeconomic activities [12]. According to one study, to reduce the number of confirmed COVID-19 cases and deaths, it was estimated that, on average, the administration of 80 vaccine doses per 100 people was necessary [13]. However, the efficacy of vaccination is challenged by an increasing number of mutated strains, clinically proven possibilities of reinfection, and globally uneven rates of vaccination [14,15].

The vaccination process depends on various societal factors such as vaccine hesitancy, vaccine refusal, practical aspects of its application, and uneven/unequal vaccine rollout [16]. Even prior to the COVID-19 pandemic, the World Health Organization identified vaccine hesitancy as one of the top 10 global threats to public health [17]. After the appearance of COVID-19, this issue has gained a completely different dimension, and several people showed different degrees of vaccination acceptance from total refusal to hesitation, including health workers [18]. Levels of COVID-19 vaccine acceptance and obstacles to its rollout are country- and context-dependent [19]. Research has shown that most people are neither absolutely for nor against COVID-19 vaccines [20,21]. Hence, to begin to understand vaccine acceptance, it is important to gain insight into the reasons behind individual and collective decision-making [22].

SARS-CoV-2 is a novel virus, and various questions about dealing with this threat by mass vaccination emerged during the pandemic, including the efficacy of vaccines, the duration of the vaccine’s effect, and the impact of new virus variants [23]. The rapid vaccine development raised questions regarding safety, availability, and efficacy [24]. This is not surprising given the fact that vaccine development usually takes 10-15 years, whereas COVID-19 vaccines were developed in less than 1 year [25]. In addition, there are various factors that can increase disease spread and mortality rates that seem incoherent with the proposal for a uniform global vaccination rollout. The mortality rates were lower in countries investing more in the health system and vice versa [15,26]. Research from the United States showed that prosperous states with a higher population of older people and a higher number of physicians had a lower rate of vaccine hesitancy compared to that of other states [12].
The availability of vaccines, both in terms of the number of doses and equal distribution, has been an important issue within various countries, involving technical as well as socioeconomic aspects [27]. Timing is also very important since seasonal and environmental factors play important roles in the reduction of COVID-19 symptomatology [15,28]. Due to the numerous factors involved, interdisciplinary collaboration appears to be an appropriate solution to tackle vaccine hesitancy [29].

To facilitate a discussion on a successful vaccination rollout process, in this study, an analysis inspired by the strengths, weaknesses, opportunities, and threats (SWOT) framework was performed to explore perceptions and establish an informative perspective of the vaccination campaigns in 17 countries during the first phase of the mass vaccination programs in the first half of 2021. To facilitate this research, the scholars from the interdisciplinary Navigating Knowledge Landscape (NKL) research network were surveyed between May and July 2021. They were asked to provide information and their own opinions about the vaccination rollout programs in their respective countries. The participating scholars belonged to different disciplines, creating a specific combination of sociological and cultural analytical competences merged with medical and public health expertise. The aim of this interdisciplinary and transnational analysis was to better understand how information-sharing practices and social context were intertwined to coproduce public opinion on vaccination as a response to the COVID-19 pandemic.

SWOT, as a strategic planning framework, is usually used in evaluation of an organization, plan, project, or business activity. It is therefore a significant tool for situation analysis that helps managers identify organizational and environmental factors affecting performance and operations [30]. The framework can be used to identify favorable and unfavorable factors and conditions, solve current problems in a targeted manner, recognize the challenges and obstacles faced, and formulate strategic plans to guide scientific decisions [30-33]. The SWOT framework strives to offer a comprehensive, systematic, and accurate description of the scenario in which a topic is located [34]. SWOT analysis has two dimensions: internal and external. The internal dimension includes organizational factors focusing on strengths and weaknesses, whereas the external dimension includes environmental factors, namely opportunities and threats [30]. Since SWOT analysis is primarily used in organizational studies, our goal was to use its elements as a conceptual and narrative analysis tool where focus was placed on the intertwining viewpoints of social, political, and public health practices. A similar approach has already been applied as a research method in which aspects of the SWOT framework were used to yield more precise and organized data [35]. However, to date, a SWOT-based analysis of the COVID-19 vaccination campaign has only been reported for India and Zimbabwe [36-38]. Therefore, with this study, we aimed to offer a new transdisciplinary and multinational viewpoint of the vaccination process.

### Methods

#### Study Design

The data set included 25 contributions from 27 members of the NKL research network, collected from May to July 2021. These members contributed their viewpoints through a questionnaire aimed at mapping, in a representative manner, the rollout of the vaccination campaigns against SARS-CoV-2 during the early stages when vaccines were available to the general public.

All contributions were collected in a public data set, which is available with open access in Mendeley Data [39].

#### Study Sample

The 27 contributors were from 17 different countries: Australia, Austria, Croatia, Czech Republic, Germany, Hungary, Italy, Norway, Portugal, Romania, Serbia, Slovenia, South Korea, Sweden, Turkey, Ukraine, and the United Kingdom (including England and Scotland). Three contributions from the same country were received from Slovenia, Sweden, and Portugal; two from Croatia and the United Kingdom; and one contribution from each of the rest of the countries. Two contributions were coauthored (from Australia and Sweden). The contributors come from different academic backgrounds, but most of them are experts in the fields of life sciences, sociology, philosophy, and medicine. However, it is important to note that the contributors were expressing their own opinions and perceptions.

#### Measures

The questionnaire contained three parts asking about the status and details of COVID-19 vaccination in the respondent’s country. Contributors were asked to return a short-text (ie, narrative) answer of 200-300 words per part. In this study, we focused only on the SWOT-related aspect of the responses (ie, Part 1) and the responses to the other parts of the questionnaire (Parts 2 and 3) were considered only to identify the eventual contribution to the SWOT-inspired analysis. SWOT elements were selected among the entire response text during the analysis process. The specific questions are presented in Textbox 1.
Textbox 1. Questionnaire items.

- **Part 1: The national vaccination program**
  Describe the COVID-19 vaccination program in your country: what were its strengths, main weaknesses, opportunities to improve it, and threats to its success?

- **Part 2: Public discourse and ethics**
  How would you describe public responses to your country’s vaccination program? What is your impression on the various collective attitudes toward the vaccination program in your country? Were there any ethical issues or concerns around the vaccine program in your country?

- **Part 3: Personal experience**
  What is your personal experience, opinion, or attitude regarding COVID-19 vaccination?

**Ethical Approval**

Ethical approval for this study was obtained from The University of Edinburgh, Scotland, United Kingdom.

**Data Analysis Procedure**

To fulfill the study’s aims and obtain results, reflexive thematic analysis [40] and descriptive statistics (frequency analysis) of the themes were performed. This method is considered appropriate for exploratory research such as our study. Moreover, flexibility of the thematic analysis and opportunity for theme development seemed a great fit and application for our data set [40,41]. The open-ended questions allowed for formulating responses that enabled the respondents to frame the description of the vaccination process in their countries according to their own personal views.

For the purposes of reflexive thematic analysis, we divided the responses into four categories according to the elements of the SWOT framework. The subcategories of each category were identified and a list of the themes for each SWOT element was established. In a subsequent step, we analyzed the data for patterns and recurring topics. We looked for country-specific differences and similarities in regulations and practices. In addition, close attention was paid to how the experts made sense of their experiences with the vaccination process and how the issues addressed were expressed. In presentation of the research results, focus was placed on themes identified throughout the reflexive thematic analysis. The results were then contextualized based on the existing literature.

Following that, frequency analysis of the identified themes was performed in relation to the corresponding countries. In the case of multiple contributions from the same country referring to the same theme or subtheme, only one data point was counted. The obtained results are presented in the form of tables and graphs.

**Results**

**Thematic Analysis**

**Overview of Themes**

Reflexive thematic analysis of collected contributions was performed independently by two researchers (VK, KN). Through the process of the reflexive thematic analysis [40], the numbers of themes respectively belonging to the elements of strengths, weaknesses, opportunities, and threats were established (Figure 1).

**Figure 1.** Graphical presentation of the established themes within each of the strengths, weaknesses, opportunities, and threats (SWOT) elements.
Thematic analysis of the vaccination process yielded a nearly even distribution of the four SWOT elements across all included countries and contributors, with 7 themes identified for strengths, 5 themes identified for weaknesses, 6 themes identified for opportunities, and 7 themes identified for threats. In total, analysis of the SWOT elements covered 25 different themes.

The contributors shared their subjective perceptions of the effectiveness of the vaccination campaigns in their countries, which ranged from claims of success to voices of criticism. The United Kingdom was the first country in the world to start the COVID-19 vaccination program in December 2020. Shortly afterward, the vaccine rollout was launched in the United States and the countries of the European Union, albeit with some delay (3 months) in Ukraine. In many countries represented in this study, the vaccination rollout started with some constraints, poor planning/management, and delays with vaccines delivery, but improved over time. In Portugal, an efficient organization of the vaccination process was achieved with the change of the Head of the Vaccination Task Force. In the countries of the European Union, the vaccination process was coordinated with that of other member countries (Croatia), although this collaboration was not always perceived as efficient, as pointed out by a contributor from Sweden.

A successful vaccination program was achieved in the United Kingdom, being respectively described as “overall…a large success” and “an overwhelming success” [39]. The contributions from Portugal and Serbia highly rated the results of the vaccination programs in their respective countries in relation to the high vaccination rate and being ahead of plans/schedule. A relatively successful vaccination process was also reported in Croatia, Hungary, Italy, and Norway. Efficient implementation was noted in Turkey, and an active vaccination process was described in South Korea with major public facilities offering vaccinators discounts or exemptions from paying admission or usage fees. For some other countries, the collected contributions reported low vaccination coverage in the survey period (May to July 2021), including Australia, Romania, and Ukraine. The respondent in Romania specifically reported low coverage for high-risk groups and people over 65 years old. Moreover, very low coverage of rural areas occurred due to lack of local community involvement, especially mayors, policy makers, and family doctors, with some of the latter refusing to dispense vaccines.

Slow rollout of the vaccines was noted in Australia, Austria, and Germany. In Australia, the delayed vaccine rollout has been described as a “vaccine stroll out,” as by July 2021, only 6% of the Australian population had been vaccinated [39]. Moreover, some individuals in priority groups such as older people or those with disabilities living in long-term care homes were still waiting for their second or even their first dose of the vaccine. In addition, in some countries, the vaccination points were hard to access in remote, rural areas (Australia).

If we are to judge vaccination rollout success by looking at the percentage of people who had received at least one dose of the vaccine during the time period corresponding to our data collection, the most successful country in our sample was the United Kingdom, with approximately 70% of the population receiving at least one dose (Multimedia Appendix 1) [42]. The lowest percentage was reported in Ukraine, where only approximately 8% of people had received a single vaccine dose [42].

**Strengths**

The primary themes related to strengths included (1) societal discussion/consensus on priorities to get the vaccine, (2) defined vaccination strategy/plan, (3) vaccine availability, (4) positive attitudes toward vaccines and the vaccination process, (5) practical aspects of the vaccination solved (eg, medical personnel satisfied, sites easy to access, fast process, no long queues), (6) well-designed public communication campaign on the vaccination process, and (7) flexibility to provide vaccines.

High availability of vaccines was reported in Hungary, Italy, Sweden, and the United Kingdom. Following the controversies around the possible side effects of AstraZeneca vaccines, stocks of the European Union–approved vaccines were excessive in Slovenia. The wide availability of vaccines to whoever wanted them was considered a strength of the vaccination campaigns. In Romania, free vaccination has been offered to everybody who wanted one, including those with Romanian or European citizenship. In Portugal, vaccination was available independent of legal status, including to undocumented migrants. Free vaccination was also offered in Serbia to people from abroad, primarily citizens of neighboring countries, with no restrictions.

Medical workers played a key role in achieving successful vaccination campaigns. Family doctors contributed to the success of the vaccine rollout in Croatia and a helpful approach was reported by the medical staff of the Czech Republic. For Portugal, strong commitment of health care professionals and communication initiatives of the medical doctors to clarify doubts related to the vaccine’s side effects were noted.

Transparent planning and strategies, as well as prioritization of people with higher infection risk or greater vulnerability, were commonly reported strengths of the vaccination programs. In most countries, the prioritization was perceived as fair, although in some countries controversial cases of people from nonpriority groups being vaccinated early also occurred (Portugal, Slovenia). The priority groups in most countries included older adults, those with underlying health conditions, and workers exposed to a high infection risk. By contrast, vaccination of health care professionals has not been prioritized in Sweden. In all countries, the vaccine was provided free of charge, dispensed on a voluntary basis; however, mandatory vaccination was reported for medical workers in Italy and South Korea and for people in high-risk jobs in Australia. Moreover, an easy registration process, owing to easy-to-access platforms such as apps, web pages, or via the phone, was described for Turkey and Ukraine. Automatic enrollment based on medical records via general practitioners (eg, family doctors) was available in the United Kingdom. An efficient registration process in the Czech Republic was also claimed as a strength.

**Weaknesses**

The primary themes related to weaknesses were as follows: (1) social divide due to the vaccine distribution and side effects,
(2) unclear vaccination strategy/plan, (3) lack of vaccines, (4) negative attitudes or hesitancy toward vaccines and the vaccination process, and (5) barriers to access vaccines.

Lack and shortage of vaccines were emphasized in Ukraine and Turkey, as well as at the beginning of the vaccination rollout in some other countries, where delayed deliveries were also reported. Delayed rollout to the remote Aboriginal communities was noted in Australia. The registration process was essential in achieving an effective rollout of the vaccines. Poor functioning of the distribution organization was highlighted by many participants from different countries, especially in the early stages of vaccination programs. Getting a vaccination appointment was rated as difficult in Sweden.

Trust was pointed out as an important issue in several contributions. A low level of trust in the medical science (Croatia), in the effectiveness and safety of the vaccines (Ukraine, Romania, Slovenia), and in the official authorities (Slovenia) were reported. An overall high level of skepticism in society at large was observed (Germany). Lack of enthusiasm and willingness to be vaccinated or vaccine hesitancy were widely reported (Austria, Norway, Romania, Slovenia, Ukraine). Despite the very successful vaccination process in Serbia, only a small percentage of younger people and health workers were vaccinated in the country. High hesitation among young people was also reported in Slovenia. In contrast, in Ukraine, young people were rather eager to be vaccinated, despite the high level of general hesitancy noted in the country. In Australia, vaccine hesitancy was exacerbated by the risks of side effects reported for the AstraZeneca vaccine.

Lack of clear and coherent communication on how to receive the vaccination was considered an important barrier to access in Slovenia. The need for suitable and unequivocal guidelines about vaccination was stressed in Italy, as constant changes have confused the population and discouraged vaccination, while different rules in different parts of the country and frequent regulation changes were noted to have discouraged vaccination in Germany. Unavailability of scientific information to foreigners/migrants, especially for those not fluent in the local language (eg, the home workers caring for the older population) was stressed for Italy. The digitally based vaccination approach was considered an important barrier to those not having adequate digital abilities. In Sweden, despite having one of the highest internet coverage rates in the world, people living in socially disadvantaged areas, including asylum seekers and migrants, and older adults or people with cognitive impairment who did not master the digital skills required were at risk to be excluded from accessing important information. A low level of digitalization was also mentioned as an obstacle to vaccination success in Romania.

Opportunities

The primary themes related to opportunities were as follows: (1) adding more flexibility; (2) increasing vaccine availability and multiple options for registration; (3) active outreach to marginalized groups, vulnerable citizens, refugees, and ethnic minorities; (4) improving the role of the media (better communication) and national awareness campaigns; (5) information sharing about the usefulness of the vaccination process; and (6) provisions for vaccinated individuals.

The freedom to choose to make an appointment for vaccination, no matter where people were registered (Sweden), and adding more flexibility to accessing vaccination (Croatia) were considered among the opportunities to improve the vaccination rollout.

To motivate people to be vaccinated, financial support (approximately US $30) was offered in Serbia. Vaccination coupons or exemptions from admission or usage fees of public facilities (approximately US $900) were introduced by a National Vaccine Injury Compensation Program in South Korea. In addition, this country also allowed a one-day “vaccination leave” from work to be taken the day after receiving the vaccine, along with an additional one-day leave in the case of experiencing some subsequent side effects [39]. In Ukraine, in the unlikely case that vaccination would cause disability or death, a compensation allowance (approximately US $21,000-27,000) was promised by the government.

Among the opportunities to improve the vaccination process, the freedom to choose among the available vaccine brands/types was recognized as a good strategy to counteract the arising doubts about a certain brand of vaccine (Slovenia, Ukraine). The choice of vaccine brand was also available in Turkey and in Serbia, contributing to successful vaccination campaigns. Finally, a more responsible role of media was mentioned as an opportunity to improve people’s attitude toward vaccination, as pointed out for Croatia and Ukraine. Moreover, in various countries, the respondents suggested that improvement of communication strategies and specific information programs might be crucial to reach vaccine-hesitant citizens and facilitate the vaccine rollout.

In addition to traditional media, social media were noted to play a role. Social media influencers were identified to positively contribute to motivating people to be vaccinated, producing a “crowd effect,” as reported for Croatia and Ukraine, where public figures, such as the President and the health minister gave declarations through the media. Vivid promotions in favor of vaccination by persuasive political and medical discourses, accompanied by enthusiastic arguments in favor of science and against conspiracy theories and vaccination skepticism, were described for Slovenia.

Themes

The primary themes related to threats were as follows: (1) appearance of new virus strains/lower efficacy of the vaccines, (2) unforeseen side effects of the vaccines, (3) spreading disinformation, (4) ethical controversies, (5) legal controversies, (6) religious controversies, and (7) a change in the behavior of vaccinated individuals that facilitates the spread of infection.

Low trust in the efficacy and safety of the vaccines (Romania, Slovenia, Ukraine); a negative influencing role played by some media communications, especially when stressing the side effects (Serbia, Sweden), and alleged corruption related to the vaccine prioritization (Slovenia) were regarded as relevant threats to be considered for achieving successful mass vaccination campaigns. Insufficient information, disinformation,
or misinformation in the media and on the internet were reported for the Czech Republic, Sweden, and Romania, while development of conspiracy theories about vaccines was pointed out for Slovenia and Ukraine. Disputable communication from the government regarding vaccines and other public health measures such as lockdowns was described for Germany. Lack of adequate public communication strategies was also noted in Slovenia. Failures in communication with people from different cultural groups were reported in Australia.

Ethical concerns associated with the use of leftover doses were pointed out by respondents from Sweden and Portugal, referring to a lack of planning for how to handle leftover vaccines that could not be administered the next day or to the overall mismanagement of vaccine administration. In contrast, the opportunity to get a leftover dose was marked as a strength at the beginning of the vaccination campaign in Ukraine, where this was the only option to be vaccinated for those in nonpriority groups. Confusing messages from religious leaders and local

community priests were reported in Romania. Concerns of disobeying the Islamic conduct codes raised by vaccination opponents was described for Turkey, as during the month of Ramadan fears were prompted that vaccination during the fasting period was not acceptable.

**Frequency Analysis**

**Overview**

To explore the distribution of the responses by countries, frequency analysis was performed (Figures 2-5). Responses reporting a certain theme are marked in the figures in green color and assigned a value of 1, whereas those that did not mention the theme are marked with light yellow and assigned a value of 0. The total score corresponds to the sum of values of all related responses. Additionally, the average percentage of responses distributed for each element and theme was calculated (Multimedia Appendices 2-5).

![Figure 2](https://humanfactors.jmir.org/2024/1/e44258)

**Figure 2.** Overview of the opinions covering strengths-related themes by country. Green indicates presence of a theme (assigned a value of 1) and yellow indicates absence of the theme (assigned a value of 0). The total score corresponds to the sum of values of all related responses.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Australia</th>
<th>Austria</th>
<th>Croatia</th>
<th>Czech Republic</th>
<th>Germany</th>
<th>Hungary</th>
<th>Italy</th>
<th>Norway</th>
<th>Portugal</th>
<th>Romania</th>
<th>Serbia</th>
<th>Slovenia</th>
<th>South Korea</th>
<th>Sweden</th>
<th>Turkey</th>
<th>UK-England</th>
<th>UK-Scotland</th>
<th>Ukraine</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Societal discussion/consensus on priority to get the vaccine</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Defined vaccination strategy plan</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Vaccines’ availability</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Positive attitudes toward vaccine and vaccination process</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Practical aspects of the vaccination process</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Well-designed public communication or the vaccination process</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Flexibility to provide vaccines</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>6</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

![Figure 3](https://humanfactors.jmir.org/2024/1/e44258)

**Figure 3.** Overview of the opinions covering weaknesses-related themes by country. Green indicates presence of a theme (assigned a value of 1) and yellow indicates absence of the theme (assigned a value of 0). The total score corresponds to the sum of values of all related responses.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Australia</th>
<th>Austria</th>
<th>Croatia</th>
<th>Czech Republic</th>
<th>Germany</th>
<th>Hungary</th>
<th>Italy</th>
<th>Norway</th>
<th>Portugal</th>
<th>Romania</th>
<th>Serbia</th>
<th>Slovenia</th>
<th>South Korea</th>
<th>Sweden</th>
<th>Turkey</th>
<th>UK-England</th>
<th>UK-Scotland</th>
<th>Ukraine</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social divide due to the vaccine distribution and side effects</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Unclear vaccination strategy plan</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Lack of vaccines</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Negative attitudes toward vaccine and vaccination process</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Barriers to access the vaccine</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>13</td>
</tr>
</tbody>
</table>
Figure 4. Overview of the opinions covering opportunities-related themes by country. Green indicates presence of a theme (assigned a value of 1) and yellow indicates absence of the theme (assigned a value of 0). The total score corresponds to the sum of values of all related responses.

<table>
<thead>
<tr>
<th>Country</th>
<th>Australia</th>
<th>Austria</th>
<th>Croatia</th>
<th>Czech Republic</th>
<th>Germany</th>
<th>Hungary</th>
<th>Italy</th>
<th>Norway</th>
<th>Portugal</th>
<th>Romania</th>
<th>Serbia</th>
<th>Slovenia</th>
<th>South Korea</th>
<th>Sweden</th>
<th>Turkey</th>
<th>UK-England</th>
<th>UK-Scotland</th>
<th>Ukraine</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adding more flexibility</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Increasing availability of vaccines and multiple options for registration</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Active reach of marginalized groups, vulnerable citizens, religious, ethnic minorities</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Improving role of media [better communication], national awareness campaigns</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Information spreading about the usefulness of the vaccination process</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Provisions for vaccinated individuals</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5. Overview of the opinions covering threats-related themes by country. Green indicates presence of a theme (assigned a value of 1) and yellow indicates absence of the theme (assigned a value of 0). The total score corresponds to the sum of values of all related responses.

<table>
<thead>
<tr>
<th>Country</th>
<th>Australia</th>
<th>Austria</th>
<th>Croatia</th>
<th>Czech Republic</th>
<th>Germany</th>
<th>Hungary</th>
<th>Italy</th>
<th>Norway</th>
<th>Portugal</th>
<th>Romania</th>
<th>Serbia</th>
<th>Slovenia</th>
<th>South Korea</th>
<th>Sweden</th>
<th>Turkey</th>
<th>UK-England</th>
<th>UK-Scotland</th>
<th>Ukraine</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appearance of new virus (unknown/lower efficacy of the vaccine)</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Unknown side effects of the vaccines</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Spraying disinfection</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Ethical controversies</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Legal controversies</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Religious controversies</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Strengths

Three themes dominated the analysis of strengths, each being covered in 7 reports: societal discussion/consensus on priorities to get the vaccine, defined vaccination strategy/plan, vaccines’ availability, and flexibility to provide vaccines.

The strengths theme “societal discussion/consensus on priorities to get the vaccine” was mentioned in Croatia, Romania, Sweden, Turkey, UK Scotland, UK England, and Ukraine. “Defined vaccination strategy/plan” was reported in Croatia, Italy, Slovenia, Sweden, Turkey, UK (England), and Ukraine. “Wide vaccines availability” was indicated in Croatia, Hungary, Serbia, Slovenia, Sweden, and the United Kingdom. “Positive attitudes of the society toward vaccines and vaccination process” were reported for Croatia, Italy, Portugal, and Sweden. Logistic aspects of the vaccination being solved (including satisfaction of the medical personnel, sites easy to access for registration, fast process, no waiting in line) were noted for Romania, while well-designed public communication on the vaccination process was described for the Czech Republic, Hungary, and Serbia. Flexibility to provide vaccines was highlighted as a potential strength in the contributions from Croatia, Czech Republic, Portugal, Serbia, Slovenia, Turkey, and Ukraine. Relatively even distribution was identified across strengths categories with the exception of practical aspects of the vaccination solved that was reported by only one contributor.

Weaknesses

Most frequently reported weaknesses were barriers to access the vaccination (13 reports) and negative attitudes or hesitancy toward vaccines and the vaccination process (9 reports).

Social divide due to the vaccine distribution and side effects were considered weaknesses in Australia, Austria, Croatia, Germany, Hungary, Italy, Portugal, and Sweden. Unclear vaccination strategy/plan was described in Germany, Hungary, and Portugal. Lack of vaccines was noted in Australia, Croatia, Germany, Italy, Portugal, Slovenia, Turkey, and Ukraine (note that the questionnaire addressed these issues only related to the first 6 months of the vaccination campaigns). Negative attitudes or hesitancy toward vaccines and the vaccination process were described in Australia, Austria, Hungary, Portugal, Slovenia, Sweden, Turkey, UK England, and Ukraine, while barriers to access vaccination, including problems with prioritization, registration, and unfair/nontransparent distribution of the vaccines were noted in Australia, Croatia, Germany, Hungary, Italy, Czech Republic, Portugal, Romania, Slovenia, Sweden, Turkey, UK Scotland, and Ukraine. In addition, 72.2% of the contributions reported barriers to access vaccines as a weakness. Conversely, only 16.6% of our sample reported an unclear vaccination strategy/plan as weakness.

https://humanfactors.jmir.org/2024/1/e44258
Opportunities

The frequencies of the selected opportunities to improve the vaccination process were rather low including a maximum of 5 countries. Adding more flexibility to the vaccination process was mentioned in Croatia and Ukraine; increasing availability of the vaccines and multiple options for registration were mentioned in Croatia, Germany, Romania, and Ukraine; active reach of marginalized groups, vulnerable citizens, refugees, and ethnic minorities was mentioned in Australia and Sweden; improving the role of the media (better communication) and national awareness campaigns were indicated in Portugal, Romania, and Slovenia; information spreading about the usefulness of the vaccination process was highlighted in Croatia, Portugal, Slovenia, Sweden, and Ukraine; and monetary provisions for vaccinated individuals were mentioned in Serbia and Ukraine. Contributors did not report opportunities in large numbers. The highest percentage (27.7%) of responses related to the opportunities-related themes was attributed to spreading information about the usefulness of the vaccination process.

Threats

Concerning the possible threats to a vaccination campaign’s success, the contributors from Australia, Serbia, and Ukraine remarked the possible appearance of new virus strains and lower efficacy of the vaccine; unforeseen side effects were noted as possible threats in the contributions from Serbia and Sweden; spreading disinformation were noted or could be concluded from the abstracts from Australia, Croatia, Czech Republic, Romania, Slovenia, South Korea, Sweden, and Ukraine. Other possible threats to vaccination success mentioned were ethical (Czech Republic, Germany, Italy, Portugal, Romania, Slovenia, South Korea, UK Scotland, and Ukraine), along with legal (Italy) and religious controversies (Romania). Ethical controversies (50%) and spreading information (44.4%) were the most highly represented threats-related themes.

In this study, the mainly acknowledged threat feature for achieving successful vaccination campaigns reported by the respondents was related to the likely occurrence of viral mutations, resulting in new virus strains with the ability to escape the immunizing effects of the present available vaccines. This fact has been pointed out as a relevant source of uncertainties and doubts about the vaccines’ effectiveness as well as about their overall reliability and utility.

Discussion

Principal Findings

Since its introduction in the 18th century to the present day, vaccination has been one of the most effective tools in the battle against infectious diseases [10,43]. Owing to the high efficacy of vaccines, the public health burden of infectious diseases has been significantly reduced throughout the years [10]. However, despite their proven track record, the phenomenon known as “vaccine hesitancy” has been around almost as long as vaccination itself. This reluctance to accept an injection of an “unknown substance” into the body is exasperated by a need to vaccinate a large number of healthy individuals, including in the case of COVID-19 [44].

This study, based on an analysis of interdisciplinary experts’ viewpoints in 17 different countries inspired by the SWOT framework, allowed us to identify 25 themes distributed across the four SWOT elements. To our knowledge, this is the first study to analyze and compare the vaccine rollout process in various different countries. The frequency of the appearance of these themes and their distribution across the countries allowed us to select those that stand out. As the contributions were inspired by the SWOT framework, the presented analysis could be easily synthesized into the four main overarching SWOT categories (Figure 6). With respect to the strengths of the vaccination process, the identified seven themes correspond to a single category referred as Availability, being the major strength of the successful vaccination program. When weaknesses were described, the five themes identified could be best described by a single category termed Barriers, which were either not recognized or not addressed by the vaccination programs. The external aspects of opportunities described via the six themes identified fit under category Transparent communication and promotion, which allows other societal forces to contribute to the vaccination process. Finally, the seven themes describing threats correspond to the Societal divide category, where a polarized society has the potential to spoil even well-thought-out initiatives.

We believe that these categories offer the best representation of the most frequently reported themes in each of the SWOT elements. However, due to the intertwining factors present in the vaccination rollout process, it is important to not look at this distribution as a binary (presence/absence) phenomenon. This is particularly relevant when splitting the identified themes into “internal” and “external” categories. In the current SWOT 2×2 matrix (Figure 6), Availability and Barriers are labeled as internal categories, whereas Transparent communication and promotion and Societal divide are suggested as external categories [45]. However, within the Societal divide category labeled as a threat, there are ethical, religious, and legal controversies reported as important themes. Therefore, one cannot classify a controversy per se as a threat, as controversies can serve just as much as a source of debate with the potential to improve the vaccination process.
The specific time window when the study was performed corresponds to a relatively early phase of the vaccination process (on average half of the population had been vaccinated in the analyzed countries). This leads to a very specific bias in the submitted data: urgency to tackle an important and pressing issue. We reiterate that our study analyzed the subjective viewpoints of the respondents; hence, some of the themes across these categories were dependent on the various individual, psychological, emotional, and societal aspects specific to the given time window. The sudden appearance of COVID-19 and its rapid spread called for appropriately rapid responses. Considering psychological factors of egocentrism, information availability, social/group confirmation, individual motivation, and emotional affect as foundations of that rapid decision-making process, it is easily possible to misjudge and/or misperceive the key elements of the reasoning arising from the complexity of the situation [46]. However, although there were 27 individual respondents from 17 different countries, our results did not show country-specific differences. Hence, our findings can contribute to the development of strategies that will maximize the promotion of strengths and opportunities while minimizing weaknesses and threats globally.

The identified themes are consistent with the research on this topic [36,38]. The most prominent theme in the existing literature, which was also present in our study, explores the effective medical and public health system measures mapping on the key strengths identified herein. This shows how preparation and prevention strategies work, and how they can be used as a base of the powerful pushback against the spread of COVID-19. Moreover, a positive attitude toward vaccination has been defined as a strength in similar studies in India and Zimbabwe [36,38].

The application of the SWOT framework to complex societal processes can also be seen as a source of confusion. For example, a “strength” is considered as an internal aspect of the process, which can be understood to relate to the vaccination campaign itself. From this perspective, the attitude toward vaccination does not seem to be an internal component but rather an external aspect of SWOT and thus should be more appropriately classified as an opportunity rather than a strength. However, application of the SWOT framework in such complex scenarios requires consideration of the vaccination campaign as part of a sociotechnical system, thus incorporating vital elements of the social environment within the situated practice of vaccination. Combining our findings obtained from individuals from 17 different countries with previous research, it can be concluded that good organization that addresses the availability of vaccines coupled with an engaging societal discussion would represent a key strength/opportunity of the vaccination process.

A lack/shortage of vaccines combined with various logistical challenges have been reported as major issues for the success of vaccination campaigns within previous research [47,48]. The demand-supply gap combined with lack of knowledge and supporting infrastructures have been reported as particular weaknesses [36,38]. Compounding unequal vaccine distribution with unknown disease progression and an uncertain response to the vaccine seems to be the biggest barrier in the vaccine rollout [13]. Similarly, the respondents of this study recognized the practical issues of availability and fairness of distribution, and coupled these issues with the related attitudes and social division. This points to the fact that social distrust needs to be addressed within a vaccination plan as a major barrier. For both strengths and weaknesses, no clear geographical divide was present.

Increasing the public awareness about the vaccine effects through transparent communication and promotion stood out as a key opportunity-related theme. Communication reports on
the widespread acceptance of COVID-19 vaccines have shown to be effective tools to further increase vaccine acceptance [49]. Moreover, in an attempt to promote vaccination, some public figures have been vaccinated on television [49]. It is interesting to see that people who used mainstream media outlets as their major source of information on health were more likely to get vaccinated [50]. Our data support the notion that transparent information-sharing about biological mechanisms, efficacy, as well as side effects of the vaccines motivates people to join vaccination programs. Previous research has identified the potentially influential role of media in increasing people’s trust in vaccines when they hear politicians, celebrities, or other famous people talking positively about them [36]. Trust in vaccines, medical science, and medical professionals—together with other involved stakeholders, including government and policy makers—was highlighted in the analyzed contributions as an important factor. These findings align with previous research that found lack of communication from trusted providers and community leaders as one of the main reasons for low COVID-19 vaccination rates [44]. Communication of vaccine information and promotion of its uptake in the digital era includes the use of social networks [51]. However, the use of social networks is also associated with risks due to the wildfire-like dynamics of rumors in the digital environment and issues with unknown algorithms used by for-profit entities filtering information [52,53]. Social networks are expected to drive healthy public debates; however, they instead frequently reinforce like-minded “bubbles” and increase polarization [53,54].

Discussions on matters of autonomy (an individual’s right to choose) and state power have always been at the center of public health ethical dilemmas [29]. In the specific case of COVID-19 vaccines, besides the tensions between public health and individual interest/autonomy, other ethical challenges relate to the rapid design and testing of vaccines and who gets the vaccines (first) [55]. Public health authorities need to implement efficient, flexible, responsive, and resilient strategies to successfully fight the pandemic and raise awareness of all of the dangers arising with this disease [56]. Surprisingly, in our findings, the question of one’s autonomy did not crystalize as a theme. Instead, other ethical controversies and spreading of disinformation were found to be the most frequently reported themes within the threats element. In the present digital era, information accessibility is at its peak; however, it is important to be aware of the source of the information given that rumors and fake news are rampant [17,57].

When discussing the threats element in the SWOT framework, it was interesting that unforeseen side effects of the vaccine have not been considered as the most prominent threat theme, whereas other research shows that the most common reason for vaccine hesitancy or refusal is due to the concerns related to the side effects/safety [50,58,59]. The emergence of new virus strains was mentioned as a threat, since they decrease the efficacy of the vaccine and hence can contribute to the further spread of COVID-19. As people were already worried about the lack of information about safety, testing, and efficacy of COVID-19 vaccines, the new variants were seen as a contributor to the negative perception of the vaccines in society [15,60]. The synthesizing category for the threat element of Societal divide implies that social polarizations have the potential to paralyze a society when facing a complex public health crisis. Here, it should be stated that silencing the controversies is certainly not the path to avoid such an outcome. A society where controversies are not openly discussed is not without these controversies, but rather this situation would give rise to potentially dangerous and isolation subcultures. Although Societal divide was recognized as a threat in our sample, there were no clear examples where this has significantly directly influenced the vaccination process. Consequently, although the awareness of controversies as a potential threat was voiced, if the social environment is developed within the context of Transparent communication and promotion (opportunity), the Social divide may never reach the level of polarization to create adverse effects on public health campaigns.

Study Limitations

The collected responses represent the subjective viewpoints of experts who volunteered to take part in the study. Therefore, extrapolation to the national level must be drawn out with caution. In addition, due to the lack of research using this same methodology and implementing it on a multinational level, there were no relevant studies to make direct comparisons with and contrast conclusions. Moreover, SWOT analysis was not performed in its original form addressing organizational dynamics. Instead, this thematic analysis of expert viewpoints was only inspired by the SWOT framework. Therefore, the results of this study should be further examined and more research is needed on this topic in general. Further studies could consider interdisciplinary and multinational frameworks to find the best practice in public health policies that could yield improved vaccination rollout results globally.

Conclusion

This study was based on a collection of short responses to a specifically designed questionnaire, written by researchers from different countries and fields of expertise, thus bringing together multidisciplinary and cross-national opinions on vaccination rollout. This represents the first analysis of the vaccination process in 17 different countries inspired by the SWOT framework. The obtained results highlight the connection between organizational aspects of the vaccination rollout and corresponding societal response, both being related to the strengths and weaknesses of the process. The opportunities and threats corresponded to external societal factors, better public communication of vaccination-related issues, ethical controversies, and the spread of disinformation. The inventory of 25 SWOT-related themes and the resulting 2x2 SWOT matrix represents an approximate best-practice viewpoint for the successful implementation of public health policies—as represented by this multidisciplinary team—in the fight against COVID-19.
Acknowledgments
This study represents an activity of the interdisciplinary network Navigating Knowledge Landscapes. The study was supported by the European Union through the European Regional Development Fund, as the Scientific Centre of Excellence for Basic, Clinical and Translational Neuroscience, under grant agreement KK.01.1.07.0071, project “Synergy of molecular markers and multimodal in vivo imaging during preclinical assessment of the consequences of the ischemic stroke (SineMozak)” (to SG).

Data Availability
The data sets supporting the results presented in this study can be found in the online repository [39].

Authors’ Contributions
VK, KN, MV, and SG designed the study. VK, KN, and LM performed data acquisition, organization and analysis, and wrote the first version of the manuscript. VK, KN, LM, LL, ZT, MV, HM, ALS, and SG contributed to the interpretation of the data collected, framed the results, and critically revised the manuscript. All authors approved the submission to the journal.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Share of people from countries in our sample who received at least one dose of COVID-19 vaccine (taken from Our World in Data. Coronavirus (COVID-19) Vaccinations, 2021 [42]).

[ PNG File , 438 KB - humanfactors_v1111e44258_app1.png ]

Multimedia Appendix 2
Average percentage of opinions covering strengths-related themes.

[ PNG File , 151 KB - humanfactors_v1111e44258_app2.png ]

Multimedia Appendix 3
Average percentage of opinions covering weaknesses-related themes.

[ PNG File , 126 KB - humanfactors_v1111e44258_app3.png ]

Multimedia Appendix 4
Average percentage of opinions covering opportunities-related themes.

[ PNG File , 140 KB - humanfactors_v1111e44258_app4.png ]

Multimedia Appendix 5
Average percentage of opinions covering threats-related themes.

[ PNG File , 134 KB - humanfactors_v1111e44258_app5.png ]

References


Abbreviations

NKL: Navigating Knowledge Landscape
SWOT: strengths, weaknesses, opportunities, and threats

©Vanja Kopilaš, Khrystyna Nasadiuk, Lucia Martinelli, Lenka Lhotska, Zoran Todorovic, Matjaz Vidmar, Helena Machado, Anna Lydia Svalastog, Srećko Gajović. Originally published in JMIR Human Factors (https://humanfactors.jmir.org), 19.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Human Factors, is properly cited. The complete bibliographic information, a link to the original publication on https://humanfactors.jmir.org, as well as this copyright and license information must be included.
The Effect of a Video-Assisted Health Education Program Followed by Peer Education on the Health Literacy of COVID-19 and Other Infectious Diseases Among School Children: Quasi-Randomized Controlled Trial

Xiaojuan Zhang\textsuperscript{1,2}, Yingkun Justin Wen\textsuperscript{1,3}, Ning Han\textsuperscript{4}, Yawen Jiang\textsuperscript{1*}, PhD

\textsuperscript{1}School of Public Health (Shenzhen), Sun Yat-sen University, Shenzhen, Guangdong, China
\textsuperscript{2}School of Public Health, Sun Yat-sen University, Guangzhou, China
\textsuperscript{3}Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden
\textsuperscript{4}Institute of Public Health Supervision of Longgang District, Shenzhen, China

* these authors contributed equally

Corresponding Author:
Yawen Jiang, PhD
School of Public Health (Shenzhen)
Sun Yat-sen University
Room 533, West Wing of Medical Complex #1, Sun Yat-sen University
66 Gongchang Road, Guangming District Shenzhen
Shenzhen, Guangdong
China
Phone: 86 13632974660
Email: jiangyw26@mail.sysu.edu.cn

Abstract

Background: To improve the engagement and effectiveness of traditional health programs, it is necessary to explore alternative models of health education including video-assisted lectures and peer education.

Objective: This study aimed to evaluate the effects of a combination of video-assisted lectures and peer education on health literacy related to infectious diseases among students.

Methods: Third-grade classes from 11 pilot schools in Longgang District of Shenzhen, China, were randomized to the intervention and control groups. In the intervention group, a video-assisted interactive health education program was conducted twice over a time span of 5 months. Each of the 2 sessions included a 40-minute lecture on COVID-19 and other common infectious diseases in schools and a 5-minute science video. In addition, 5 “little health supervisors” at the end of the first session were elected in each class, who were responsible for helping class members to learn health knowledge and develop good hygiene habits. Students answered the same quiz before the first and after the second session. Models based on item response theory (IRT) were constructed to score the students’ knowledge of infectious diseases based on the quiz.

Results: In total, 52 classes and 2526 students (intervention group: n=1311; control group: n=1215) were enrolled. Responses of the baseline survey were available for 2177 (86.2%; intervention group: n=1306; control group: n=871) students and those of the postintervention survey were available for 1862 (73.7%; intervention group: n=1187; control group: n=675). There were significant cross-group differences in the rates of correctly answering questions about influenza symptoms, transmission, and preventive measures; chicken pox symptoms; norovirus diarrhea symptoms; mumps symptoms; and COVID-19 symptoms. Average IRT scores of questions related to infectious diseases in the intervention and control groups were, respectively, –0.0375 (SD 0.7784) and 0.0477 (SD 0.7481) before the intervention (\textit{P}=.01), suggesting better baseline knowledge in the control group. After the intervention, the average scores of the intervention and control groups were 0.0543 (SD 0.7569) and –0.1115 (SD 0.7307), respectively (\textit{P}<.001), suggesting not only significantly better scores but also greater improvement in the intervention group.

Conclusions: After the health education project, the correct answer rate of infectious disease questions in the intervention group was higher than that of the control group, which indicates significant effects of the combination of video-assisted lectures and peer education for the promotion of health literacy. In addition, the intervention effect of the first session persisted for at least 4
months up to the second session. As such, the proposed program was effective in improving the health literacy of school children in relation to infectious diseases and should be considered for massive health promotion campaigns during pandemics.

**Trial Registration:** ISRCTN ISRCTN49297995; https://www.isrctn.com/ISRCTN49297995

*JMIR Hum Factors 2024;11:e43943*  doi:10.2196/43943

**KEYWORDS**

infectious diseases; primary school students; quasi-randomized controlled trial; video-assisted health education; peer education; item response theory; IRT

**Introduction**

Primary school students are vulnerable to emerging and common infectious diseases such as COVID-19, influenza, mumps, and intestinal infectious diseases [1]. A survey on the reasons for sick leaves in primary and secondary schools in Shenzhen showed that the top 5 causes were common cold, gastrointestinal diseases, unexplained or other illness, influenza, and chicken pox [2]. In addition, the importance of acquiring essential knowledge regarding the prevention and control of COVID-19 cannot be overstated during the pandemic. Accordingly, it is critical to embed health promotion into the school education of primary school students. To that end, the outline of “Healthy China 2030” emphasizes the importance of fortifying health education among school children. In particular, primary schools were integral to the life cycle of the health education curriculum to the extent that early-life exposure to information on diseases and health behaviors is associated with improved future health outcomes [3].

Despite its importance, health education was highly restricted in its delivery forms. Conventionally, the most prevalent approach of health education of infectious diseases for school students was, arguably, classroom lectures aided with paper-based materials, in which the teaching contents are usually compiled by school teachers and researchers [4]. Traditional health education is also reported to have a limited duration of effects. Hampered by the collective challenges faced in traditional health education, most schools lack systematic health education programs [5]. To increase students’ interest in healthy behaviors and to extend the duration of education effects, researchers have been exploring alternative media for health education. Among the various new models, two of the prevailing approaches are video-assisted health education and interactional peer education [5,6].

In professional medical education, video-assisted lectures are useful tools for students to acquire basic clinical skills. When delivered in bundle with in-person lectures, video-based materials are often preferred by students [7]. In addition, video-assisted health education has been shown to be more effective than oral education in facilitating postoperative recovery of patients [8].

The effects of health education are not necessarily limited to the immediate recipients of the program themselves. Students may also help to shape the opinions and behaviors of their classmates by becoming peer educators of health and hygiene. Peer education is defined as “sharing experiences and learning among people with something in common,” such as a similar age, living environment, and culture [9]. There is substantial evidence that peer education is highly effective in specific areas of medical and health education, including professional medical training, chronic disease prevention, and sexual health behaviors [6,10,11]. Incorporating peer effects into the design of health education programs could, therefore, strengthen the programs' impacts on behavioral change.

However, evidence on the effects of video-assisted lectures and peer education on health literacy among school children is still lacking. Given its substantial potential for public health practice, we designed a health education package that combined video-assisted classroom teaching and peer education and tested the effectiveness of this program. This program, which we anecdotally refer to as the “Little Health Supervisors” project, was anticipated to improve the health literacy of students over an array of infectious diseases.

**Methods**

**Trial Design**

The “Little Health Supervisors” project is jointly enacted by the Longgang District Bureau of Health and the Longgang District Bureau of Education as an administrative task. Third-grade classes from 11 pilot schools in Longgang district of Shenzhen, China, were randomized to the intervention and control groups. Our aim was to allocate equal numbers of third-grade classrooms to the intervention and control groups within each school. However, schools with an odd total number of classes inevitably resulted in uneven groups; hence, one group might outnumber another eventually. This project enclosed 2 health education sessions 4 months apart in Dec 2021 and Apr 2022 in Longgang District, Shenzhen City in the Guangdong Province of China, which is a district with approximately 4 million residents and 0.4 million school students.

**Ethical Considerations**

The “Little Health Supervisors” project was launched by the district government as a public service project. The study protocol was approved by the Biomedical Research Ethics Review Committee, School of Public Health (Shenzhen), Sun Yat-sen University [2021(056)] and was registered with Longgang District Bureau of Health (Figure S1 in Multimedia Appendix 1). Informed consent was obtained from all students and their parents who met the inclusion criteria and were willing to participate. Confidentiality of information was maintained.

**Recruitment**

In the first step of sample enrollment, considering the feasibility of the project’s implementation, the Longgang District Bureau...
of Health and the Longgang District Bureau of Education recommended 1 primary school based on the willingness to participate for each of the 11 subdistricts of this district. Second, all third-grade students in the 11 schools were eligible for participation if they met the following requirements: (1) they were not taking a leave of absence from school at the time of enrollment; (2) they agreed (or their guardians agreed) to spend time on attending lectures; (3) they had access to a computer, tablet, or smartphone with an internet connection; (4) they had sufficient knowledge to use mobile devices or computers (assistance allowed); and (5) they were able to read and interpret Chinese characters. Next, as decided by the researchers, the eligible students were assigned to the intervention and control groups using the class number as the randomizer. Specifically, odd-numbered classes were assigned to the intervention group; even-numbered classes, the control group. The flowchart of participant enrollment is illustrated in Figure 1.

Figure 1. Flowchart of the “Little Health Supervisors” project (a cluster randomized controlled trial) from December 2021 to April 2022.

Data Collection

To standardize students’ knowledge of infectious diseases both before and after the education program, a questionnaire containing a quiz on COVID-19 and selected infectious diseases with relatively high local incidences was curated, which included influenza, chicken pox, norovirus diarrhea, and mumps. The questionnaire also collected demographic characteristics (school, class, student number, sex, and date of birth) and COVID-19 vaccination status. In addition, we delivered a separate questionnaire to a parent of each student who collected parental assent to vaccinate their children against COVID-19. Moreover, family socioeconomic information was also collected in the parent questionnaire, which included monthly household income and the parents’ education level [12].

Questionnaires were distributed via a web-based survey platform (Wenjuanxing, Changsha Ranxing Information Technology Co, Ltd). In the baseline survey, students completed the questionnaires in a computer laboratory with the instructions of either the computer teachers or the class advisors. To collect
parents’ responses, the teachers arranged a meeting with each family using previously connected social media to select a representative for questionnaire responses. Due to COVID-19 outbreaks during the planned time period of the second session, the postintervention survey was distributed on the web. Simultaneously, the researchers also collected the questionnaire from the control group.

Interventions
The intervention was developed by both researchers and the local health department. Details of the development process and the content of the intervention are provided in Table S1 and Figures S2-S3 in Multimedia Appendix 1 [13,14]. Students randomized to the intervention group had access to 2 free sessions of health education during the study. Each session included a 40-minute lecture on the transmission and prevention of different infectious diseases, followed by a 5-minute science video. To incentivize learning, students were informed that there would be interactional question-and-answer sections during the lecture, for which the participating students were eligible for prizes.

In December 2021, the baseline survey and the first health education session were conducted, with the former preceding the latter. The in-person lecture and the videos of the first session pertained to influenza, norovirus diarrhea, and hand hygiene. At the end of this session, 5 little health supervisors were elected by the teachers from each class. They were naturally assumed as opinion leaders, showcasing their ability to effectively convey knowledge and could supervise the learning of health knowledge and the development of good hygiene habits of their classmates. The teachers also handed out brochures, armbands, and stickers to the 5 little health supervisors. In addition, the teachers encouraged all students to take health knowledge home and improve the family’s health literacy by way of “small hands holding big hands,” which aimed to exploit the power of two-step flow theory of communication for information transmission. Originating from political science, the two-step flow theory asserts that information can be conveyed through the chain of media-opinion leaders-audience. Students may also help to shape the opinions and behaviors of their family members by becoming an opinion leader of health and hygiene [15,16].

In April 2022, the second health education session and the postintervention survey were carried out. However, the order of education and survey was reversed in relation to the first session. The lecture and the videos of the second session pertained to chicken pox, mumps, and COVID-19 symptoms. Affected by a local COVID-19 outbreak, students had to take the web-based classes at home, so the health education sessions had to be conducted in the form of recorded course videos. In the intervention group, students were required to watch the video, and the teachers also encouraged all students to distribute health knowledge to the people around them.

As for the control group, the students only received routine health education at school, which included health tips on influenza from school doctors and 1 or 2 public welfare courses conducted by the local health department or hospitals every semester. These routine health education sessions were balanced between the 2 groups.

Outcomes
The primary outcomes of this trial were the score in the original scale (hereafter referred to as “crude score”) and item response theory (IRT) score of questions related to infectious diseases, the correct answer rates of questions related to infectious diseases, and the pre-post changes in the correct answer rates after the intervention. The secondary outcomes were the COVID-19 vaccination rates. For those who did not receive COVID-19 vaccines at baseline or at the end of the program, we also exploratively asked about their willingness to get vaccinated and the reasons for not being vaccinated.

Statistical Analysis
To gain an overview of students’ characteristics, their families’ demographic data were collected. Monthly household income (in ¥) was categorized into 4 levels (<¥5000 [US $702.97], ¥5000–¥10,000 [US $1405.94], ¥10,000–¥20,000 [US $2811.88], and ≥¥20,000 [US $2811.88]). Parent’s education was grouped into 3 levels (junior high or below, secondary school [including technical secondary school], and college and above). For the questions related to infectious diseases, multiple answers were regarded as correct only if all the correct answers were selected. Correctly answered questions contributed 1 point, and incorrectly answered questions contributed 0 points. The crude score of questions related to infectious diseases ranged from 0 to 7, with a higher score indicating higher knowledge of infectious diseases. For the item of willingness to be vaccinated against COVID-19, we assigned 1, 2, 3, 4, and 5 points respectively to the 5 options of very reluctant, reluctant, neutral, willing, and very willing. To comprehensively evaluate the students’ knowledge of infectious diseases, IRT was used to fit the model of 7 items of the questionnaire. Frequently used in studies on education examinations, IRT is a set of psychometric models used to measure unobservable characteristics of the respondents and the development of scoring scales [17-19]. IRT can be used to explain the relationship between a latent trait (eg, the health literacy of school children related to infectious diseases) and observable characteristics and items (eg, questionnaire answers). IRT has at least 3 model specifications. The one parameter logistic model takes item difficulty into account when evaluating individual ability, whereas the two parameter logistic model additionally considers differential discrimination of items [19,20]. In addition to these 2 models, the three-parameter model (TPM) allows the possibility of guessing [19,20]. In this study, a TPM was selected to calculate the IRT score (Table S2 in Multimedia Appendix 1). To score the students’ latent health literacy, we fitted TPM using the R package “ltm: Birnbaum’s three parameter model” to the 7 questions related to the knowledge of infectious diseases [20]. A higher score meant higher health literacy. We plotted the estimated IRT score of questions related to infectious diseases to visualize the students’ performance (Figures S4-S7 in Multimedia Appendix 1).

Although not directly related to our main analyses, we also plotted the item characteristic curves, item information curves, and the test information curve to provide some information...
regarding the difficulty of the test (Figures S8-S10 in Multimedia Appendix 1).

Finally, to summarize categorical sociodemographic characteristics, the correct answer rates of answering the questions, the pre-post changes in the correct answer rates after the intervention, the COVID-19 vaccination rate, the reasons for nonvaccination, and the percentages of the corresponding variables were calculated. We used mean and SD to describe the crude score, the IRT score of questions related to infectious diseases, and the willingness to vaccinate against COVID-19. We used t tests to compare the crude score, the IRT score, and the willingness to be vaccinated against COVID-19 across groups. Regarding the willingness to be vaccinated between 2 groups, we also conducted a stratified analysis based on the parents’ sex. Chi-square tests were carried out on the basis of the correct answer rate, the COVID-19 vaccination rate, and the reason for nonvaccination to investigate differences between the 2 groups. The pre-post changes in the correct answer rates after the intervention were compared between study groups, using the z test. Furthermore, since we used class as our intervention unit, we also conducted an additional analysis using class as the primary unit of analysis. This was undertaken to ensure that our class-based examination would yield coherent findings as well (Tables S3-S5 in Multimedia Appendix 1). A P value less than .05 was considered significant. All data were analyzed using SPSS (version 26; IBM Corp) and R (version 4.2.0; The R Foundation).

Power
We calculated the power of this study on the basis of the sample size of the intervention and on the primary outcome. To calculate power, we used the sample size of 1862 (intervention group: n=1187; control group: n=675), an acceptable probability for type I error of .05, a pooled SD of 0.767, and a minimal difference in the infectious disease knowledge scores between the 2 groups of 0.166 (ie, μ1−μ2). The power of this study was 99.43%.

Table 1. Sociodemographic characteristics of third-grade students from 11 pilot schools in Longgang District of Shenzhen, China.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Intervention group, n/n (%)</th>
<th>Control group, n/n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>691/1306 (52.9)</td>
<td>459/871 (52.7)</td>
</tr>
<tr>
<td>Female</td>
<td>615/1306 (47.1)</td>
<td>412/871 (47.3)</td>
</tr>
<tr>
<td><strong>Monthly household income (¥)²</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;5000</td>
<td>140/1430 (9.8)</td>
<td>95/1066 (8.9)</td>
</tr>
<tr>
<td>5000–10,000</td>
<td>359/1430 (25.1)</td>
<td>267/1066 (25.0)</td>
</tr>
<tr>
<td>10,000–20,000</td>
<td>403/1430 (28.2)</td>
<td>311/1066 (29.2)</td>
</tr>
<tr>
<td>≥20,000</td>
<td>528/1430 (36.9)</td>
<td>393/1066 (36.9)</td>
</tr>
<tr>
<td><strong>Parent’s educational level</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Junior high or below</td>
<td>116/1430 (8.1)</td>
<td>82/1066 (7.7)</td>
</tr>
<tr>
<td>High school or technical secondary school</td>
<td>275/1430 (19.2)</td>
<td>209/1066 (19.6)</td>
</tr>
<tr>
<td>College and above</td>
<td>1039/1430 (72.7)</td>
<td>775/1066 (72.7)</td>
</tr>
</tbody>
</table>

²¥1=US $0.1445.
Correct Answer Rates of Questions Related to Infectious Diseases

At baseline, the correct answer rates for questions related to influenza symptoms, influenza preventive measures, and norovirus diarrhea symptoms were different between the intervention and control groups. Specifically, the correct answer rate was higher in the control group (Table 2). In terms of the correct answer rates for questions regarding influenza transmission, chicken pox symptoms, mumps transmission, and COVID-19 symptoms, there were no significant differences between the 2 groups (Table 2). After the intervention, the differences between the 2 groups in the correct answer rates for questions regarding influenza symptoms, influenza preventive measures, and norovirus diarrhea symptoms were no longer observed (Table 2). By contrast, the differences in the correct answer rates for questions regarding chicken pox symptoms, mumps transmission, and COVID-19 symptoms between the 2 groups at the end point were significant, such that intervention group outperformed the control group (Table 2).

Table 2. The correct answer rates for questions related to infectious diseases in the intervention and control groups.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Total, %</th>
<th>Intervention group, %</th>
<th>Control group, %</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Influenza symptoms</td>
<td>67.57</td>
<td>65.39</td>
<td>70.84</td>
<td>.008</td>
</tr>
<tr>
<td>Influenza transmission</td>
<td>81.86</td>
<td>81.47</td>
<td>82.43</td>
<td>.57</td>
</tr>
<tr>
<td>Influenza preventive measures</td>
<td>84.66</td>
<td>83.08</td>
<td>87.03</td>
<td>.01</td>
</tr>
<tr>
<td>Norovirus diarrhea symptoms</td>
<td>56.41</td>
<td>54.21</td>
<td>59.70</td>
<td>.01</td>
</tr>
<tr>
<td>Chicken pox symptoms</td>
<td>28.34</td>
<td>28.79</td>
<td>27.67</td>
<td>.57</td>
</tr>
<tr>
<td>Mumps transmission</td>
<td>6.89</td>
<td>7.27</td>
<td>6.31</td>
<td>.39</td>
</tr>
<tr>
<td>COVID-19 symptoms</td>
<td>29.54</td>
<td>28.33</td>
<td>31.34</td>
<td>.13</td>
</tr>
<tr>
<td><strong>End point</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Influenza symptoms</td>
<td>86.09</td>
<td>86.77</td>
<td>84.89</td>
<td>.26</td>
</tr>
<tr>
<td>Influenza transmission</td>
<td>78.30</td>
<td>78.69</td>
<td>77.63</td>
<td>.60</td>
</tr>
<tr>
<td>Influenza preventive measures</td>
<td>92.91</td>
<td>93.09</td>
<td>92.59</td>
<td>.69</td>
</tr>
<tr>
<td>Norovirus diarrhea symptoms</td>
<td>72.93</td>
<td>73.80</td>
<td>71.41</td>
<td>.26</td>
</tr>
<tr>
<td>Chicken pox symptoms</td>
<td>43.18</td>
<td>47.01</td>
<td>36.44</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Mumps transmission</td>
<td>10.15</td>
<td>13.23</td>
<td>4.74</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>COVID-19 symptoms</td>
<td>49.14</td>
<td>52.40</td>
<td>43.41</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Regarding the pre-post changes in the correct answer rates after the intervention, the differences between the 2 groups were significant for all items (Table 3). Specifically, the correct answer rates for questions regarding influenza symptoms, influenza preventive measures, norovirus diarrhea symptoms, chicken pox symptoms, and COVID-19 symptoms increased in both groups (for all, P<.001). However, the correct answer rates of the intervention group increased more than those of the control group. In the intervention group, the correct answer rate for questions regarding mumps transmission increased in the intervention group but decreased slightly in the control group. Compared with that before the intervention, the correct answer rate for questions regarding influenza transmission decreased slightly after the intervention (Table 3).

Table 3. Pre-post changes in the correct answer rates after the intervention in the intervention and control groups.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Total, %</th>
<th>Intervention group, %</th>
<th>Control group, %</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Influenza symptoms</td>
<td>18.52</td>
<td>21.38</td>
<td>14.05</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Influenza transmission</td>
<td>−3.56</td>
<td>−2.78</td>
<td>−4.80</td>
<td>.02</td>
</tr>
<tr>
<td>Influenza preventive measures</td>
<td>8.25</td>
<td>10.01</td>
<td>5.56</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Norovirus diarrhea symptoms</td>
<td>16.52</td>
<td>19.59</td>
<td>11.71</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Chicken pox symptoms</td>
<td>14.84</td>
<td>18.22</td>
<td>8.77</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Mumps transmission</td>
<td>3.26</td>
<td>5.96</td>
<td>−1.57</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>COVID-19 symptoms</td>
<td>19.60</td>
<td>24.07</td>
<td>12.07</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>
Crude and IRT Scores for Questions Related to Infectious Diseases

Before the intervention, there was a significant difference in the mean scores for questions regarding infectious disease knowledge between the 2 groups. The mean IRT score of the intervention group (–0.0375, SD 0.7784) was significantly lower (P=0.01) than that of the control group (0.0477, SD 0.7481). After the intervention, the mean IRT score of the intervention group (0.0543, SD 0.7569) surpassed that of the control group (–0.1115, SD 0.7307). Notably, the postintervention mean score of the intervention group increased from that at baseline, whereas the control group displayed an opposite trend (Table 4). The situation is similar for the crude score (Table 4).

<table>
<thead>
<tr>
<th></th>
<th>Intervention group</th>
<th>Control group</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crude score, mean (SD)</td>
<td>3.49 (1.628)</td>
<td>3.65 (1.552)</td>
<td>.02</td>
</tr>
<tr>
<td>IRT-based score, mean (SD)</td>
<td>–0.0375 (0.7784)</td>
<td>0.0477 (0.7481)</td>
<td>.01</td>
</tr>
</tbody>
</table>

COVID-19 Vaccination Rates

The COVID-19 vaccination rates of the intervention and the control groups at baseline were 94.8% and 93.2%, respectively; by the end of the program, they increased slightly to 97.6% and 96.6%, respectively. The differences, however, were not significant (Table 5).

<table>
<thead>
<tr>
<th></th>
<th>Intervention group, n/n (%)</th>
<th>Control group, n/n (%)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1238/1306 (94.8)</td>
<td>812/871 (93.2)</td>
<td>.13</td>
</tr>
<tr>
<td>End point</td>
<td>1158/1187 (97.6)</td>
<td>652/675 (96.6)</td>
<td>.23</td>
</tr>
</tbody>
</table>

Willingness to Get Vaccinated and the Reasons for Not Being Vaccinated

Among the study participants who have not been vaccinated against COVID-19, the differences between students’ and parents’ willingness to receive the vaccine in the 2 groups were not significant (Table S10 in Multimedia Appendix 1). After stratifying by parents’ sex, the differences between the 2 groups were still not significant (Table S11 in Multimedia Appendix 1). For students who had not been vaccinated against COVID-19 after the intervention, the students and their parents were worried about side effects among many other reasons (Table S12 in Multimedia Appendix 1).

Discussion

Principal Results

Using a quasi-randomized controlled design, this study assessed the effectiveness of a video-assisted health education program sequenced by peer education on infectious disease health literacy among school students. The results suggest that the proposed multicomponent model of health education improved the knowledge of infectious diseases among students, and are consistent with those of previous randomized controlled trials in health education among primary school students [12,21,22]. Moreover, this study not only showcases an innovative approach to raise awareness of disease prevention by incorporating technology and behavioral elements, but also represents a preliminary effort to test the effectiveness of an infectious disease health education program using IRT-based scores.

Our results encapsulate important implications for the practice of health education and healthy behavior promotion. First, the inexpensive and convenient innovative health education approach proposed in this study represents a viable approach to improve student health literacy during pandemics and should be considered in future programs of healthy behavior promotion among school students. The fact that the program was effective among third-grade students does not restrict the potential of this approach since senior students are likely to capture the contents of the program better than third-grade students. Second, the results from the second session of this study partially indicate that web-based teaching may also be an effective tool to promote student engagement in health education, which has been highlighted in previous studies but not confirmed [7].

The possible long-term effects of the first session from our findings should not be ignored. The postintervention survey was carried out immediately after the second education session (including chicken pox, mumps, and COVID-19) and 4 months after the first education session (including influenza and norovirus diarrhea). Despite the time elapsed, the correct answer rates of questions related to infectious diseases that were of focus in the first session were still higher in the intervention group than in the control group. Therefore, third-grade primary school students may endure the impact of health education for at least 4 months. Given the low likelihood of frequently setting up health education sessions in schools, the slow waning of the program’s effects is a desirable feature. However, the cross-over effect from the second session could not be ruled out. For example, the learning of COVID-19 may strengthen the students’ previous understanding of influenza and increase the effect of intervention in influenza. In addition, the second session may sensitizes the students in the intervention group. They may review the knowledge of the first session to prepare for the postintervention quiz, which may also enhance the effect of the first session.
It is noteworthy that there was some difference in response rates between the interventional and control groups. The difference in response rates might be attributable to an absence of treatment blinding. In fact, the intervention in this study could not be blinded due to its physical nature, in which case, the intervention group students might be motivated by the education sessions to meet the expectation of the educators to respond to the surveys.

In addition, there was no significant difference in the correct answer rate for questions related to flu transmission routes before or after the intervention, but the pre-post changes in the correct answer rates was different between the 2 groups, and the intervention group performed better than the control group. Owing to countrywide vaccination campaigns, the COVID-19 vaccination rates between the intervention and control groups were not significantly different. The results of the 2 questionnaire surveys showed that the vaccination rates of the 2 groups increased, which was related to the local epidemic and the country's policy encouragement for vaccination.

Limitations
Several limitations of the study should be noted when interpreting the results. First, we did not collect data on the incidence of related infectious diseases before and after the intervention. A previous study reported that in areas with a high incidence of infectious disease, the health education package had no overall effect in preventing infections. However, the intervention was effective in preventing infections in areas where the baseline prevalence was relatively low [21]. Further studies are needed to explore the impact of our composite intervention on preventing infections. Second, this study was limited in its ability to evaluate component-specific versus composite effects of the educational video, the didactic lessons, the cooperative learning exercises, and peer engagement. The 2-arm trial design could not parse out the influence of each element. Future work should incorporate multiple comparison arms to better isolate the impacts of intervention components. Third, we regret that we did not measure changes in attitudes and behaviors after the intervention, as the health education package is hypothesized to influence these aspects. This is a gap that exists in our study, which future research could explore. Fourth, we used a self-rating questionnaire to collect data. Although self-reporting is a common and accepted method, we could not completely rule out the possibility of measurement error. However, the reliability and validity of self-reporting among children aged >8 years have been shown to be good in health-related questionnaires [23,24]. Fifth, the contamination in this study may underestimate the effect of our intervention. We adopted a clustered quasi-randomized controlled trial design to mitigate within-class person-to-person contamination, although interclass contamination caused by students and teachers could not be eliminated. However, the contamination, if any, happened more likely to the first session rather than the second session since students were physically isolated during the latter. Sixth, as we did not receive the questionnaire from the students lost to follow-up, the primary analysis was not intent-to-treat. Seventh, the second session of health education originally scheduled to enter the campus was changed to web-based classes owing to the serious local epidemic. Therefore, the students were required to fill in the web-based questionnaire at home, which affected the independence of the participants in answering questions; hence, the correct answer rates of the 2 groups were generally higher than those at baseline. Besides, the recovery of the questionnaire was decreased probably due to the lack of the teachers’ supervision outside the schools. However, the missing rates were balanced between the 2 groups, thereby reducing the chances of influencing our conclusions. Moreover, the effect of the health education provided herein may be underestimated because this missing group of students and parents might have lower health literacy, in which case, the intervention would have incremental value.

Comparison With Prior Work
Despite these limitations, the primary strengths of our study are that it is the first quasi-randomized controlled trial to evaluate the effect of a video-assisted health education program sequenced by peer education on the health literacy of COVID-19 and other infectious diseases among school children, and it is also the first to report IRT scores for questions related to the infectious diseases. Additionally, while our study is quasi-randomized, the allocation process likely achieved reasonable randomization, effectively balancing confounding factors across study arms as evidenced by the systematic allocation of students to intervention or control groups based on their odd or even class numbers, as outlined in Table 1. Importantly, the allocation of students to odd or even classes was not based on systematically different characteristics, as the Ministry of Education of the People’s Republic of China does not permit students to be segregated into different classes based on specific attributes. Therefore, the grouping of students based on class number parity can be considered to approximate the effects of randomization. Moreover, the sample size in this study allowed minimal chances of underpowered analyses. Previous studies might have engaged nonrandomized designs such that mixed results were reported [12,21,22,25-30]. Although most studies demonstrated that the health intervention is effective in improving health knowledge and health literacy, a quasi-randomized controlled trial in China found that the intervention’s effect was not significant among primary school students [25]. Moreover, a number of studies adopted self-control, or observational designs, based on which solid conclusions are difficult to derive [3-5,26-30].

Conclusions
Our study confirmed that the combination of video-assisted and peer education in a health education program had significant effects on school children. In addition, the effect of the first health education session may endure after 4 months. As such, the proposed program was effective in improving health literacy related to infectious diseases among school children and should be considered for en masse health promotion campaigns during pandemics.
Acknowledgments
This study was carried out by the Longgang District Bureau of Health as an administrative task. This sponsor was involved in developing the intervention and study design, and was kept informed during data collection and data analysis. The corresponding authors had full access to all of the data in the study and had the final responsibility for the decision to submit the paper for publication. We also appreciate Jing Chen and Wanjing Zhang from the Longgang District Health Inspection Institute for their contribution and support to the early development of the intervention package.

Data Availability
The data presented in this study are not publicly available because of ethical requirements but are accessible upon reasonable request to the corresponding author.

Authors’ Contributions
All authors conceptualized and designed the study and provided administrative, technical, and material support; they also supervised the study. XZ and YJW acquired the data. XZ analyzed and interpreted the data and drafted the manuscript. YJ critically revised the manuscript for important intellectual content. XZ carried out the statistical analyses. NH is a co-correspondent (email: 207baby@163.com).

Conflicts of Interest
None declared.

Multimedia Appendix 1
Supplementary figures and tables.
[DOCX File, 922 KB - humanfactors_v11i1e43943_app1.docx]

Multimedia Appendix 2
CONSORT eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1240 KB - humanfactors_v11i1e43943_app2.pdf]

References


**Abbreviations**

| IRT | item response theory |
| TPM | three-parameter model |
Characteristics, Opportunities, and Challenges of Osteopathy Based on the Perceptions of Osteopaths in Austria: Qualitative Interview Study

Jonas Manschel¹, MSc; Jan Porthun², MMMSc; Ulrike Winkler³, Dr phil; Jean Marie A T Beuckels¹,⁴, Dr rer medic; David Martin¹,⁵, Dr med

¹Institute of Integrative Medicine, Department of Medicine, Health Faculty, University Witten/Herdecke, Herdecke, Germany
²Norwegian University of Science and Technology, Campus Gjøvik, Gjøvik, Norway
³Bundeswehr University Munich, Neubiberg, Germany
⁴Department of Osteopathy, Faculty of Health and Social Sciences, Hochschule Fresenius, Munich, Germany
⁵Tübingen University Children’s Hospital, Tübingen, Germany

Corresponding Author:
Jan Porthun, MMMSc
Norwegian University of Science and Technology, Campus Gjøvik
Teknologivik 22
Gjøvik, 2815
Norway
Phone: 47 95007916
Email: jan.porthun@ntnu.no

Abstract

Background: There are no uniform regulations for the osteopathic profession in Europe. It is subject to country-specific regulations defining who shall be allowed to practice osteopathy and which qualification shall be required. In recent years, legal regulations have been established in several European countries for the profession of osteopathy; however, these are also still pending for Austria. Currently, physiotherapists and physicians with osteopathic training are practicing osteopathy in Austria.

Objective: This study aims to examine the characteristics, challenges, and opportunities of osteopaths in Austria.

Methods: Guideline-based interviews with osteopaths (N=10) were conducted. The different research questions were examined using a qualitative content analysis.

Results: The study provided a differentiated insight into the professional situation of osteopaths in Austria. The most important result was that all interviewees unanimously supported a legal regulation of their profession. However, owing to their different professional self-image—on the one hand, individuals working on a structural basis, and, on the other hand, individuals working on a cranial or biodynamic basis—they were able to imagine a uniform professional regulation only to a limited extent. Additional topics for the interviewed osteopaths in Austria were the quality assurance of training and the urgent need for scientific research. Furthermore, the study also dealt with the influence of the COVID-19 pandemic on daily practice and on education and training in osteopathy.

Conclusions: This study is a pioneering study with regard to systematic basic research on osteopathy in Austria. The obtained results and the newly acquired research questions not only have the potential to serve as a basis for further studies but also provide insight into the working and professional situation of osteopaths in Austria for universities, schools, professional associations, politics, and—last but not least—all interested parties.

International Registered Report Identifier (IRRID): RR2-10.2196/15399

(JMIR Hum Factors 2024;11:e45302) doi:10.2196/45302

KEYWORDS
osteopathy; osteopath; osteopaths; osteopathic profession; health care system; Austria
**Introduction**

Osteopathy is a manual treatment method, the principles of which are based on its own philosophy and the consideration and treatment of special structure-function relationships in the human body [1]. Since osteopathy was established and as long as it has been applied, both its methods and the professional competence of osteopaths have been the subject of controversy among medical and therapeutic specialists. The European Committee for Standardization has defined osteopathy as a holistic, patient-centered, manual treatment method based on the interactions between the structure and function of the body and the body’s self-healing ability [2].

There is no uniform European or international regulation regarding who is allowed to practice osteopathy and which qualifications are required. However, an increasing number of European countries are developing occupational laws for osteopaths. So far, 12 European countries including Cyprus, Denmark, Finland, France, Iceland, Liechtenstein, Luxembourg, Malta, Norway, Portugal, Switzerland, and the United Kingdom have adopted legal regulations regarding the practice of osteopaths [3].

In the German-speaking countries, there is no uniform picture of the profession. In contrast to Switzerland, a legal basis for the profession of osteopathy does not exist in Germany or Austria. In Austria, physicians and physiotherapists trained in osteopathy practice as osteopaths. Physicians are allowed to practice osteopathy without any restrictions, whereas physiotherapists are only allowed to practice osteopathy upon medical assignment [4]. According to the Austrian Society for Osteopathy (Österreichische Gesellschaft für Osteopathie; OEGO), approximately 2000 osteopaths practiced in Austria in 2022.

Studies about osteopathic identity are progressing internationally. However, the various legal regulations and intraprofessional conflicts make it difficult to perceive a collective identity [5]. Especially in countries where osteopathy is not regulated by law, the data about osteopathic practitioners are considered to be weak. However, quantitative studies that have surveyed the population of osteopaths with regard to work status, training, professional identity, or characteristics of clinical practice such as the typical patient profile and the use of diagnostic and treatment modalities exist already. In Austria, 2 surveys of osteopaths have been conducted in the past as part of final theses [6,7]. In 2022, the results of the Osteopathic Practitioners, Estimates, and Rates survey were also published for Austria, thus creating a solid data basis about osteopathic practitioners in Austria for the first time. The typical osteopath was defined as female, aged between 40 and 49 years, self-employed, worked before as a physiotherapist, trained in osteopathy part time, and successfully completed a master’s degree [8]. However, there is a lack of studies with qualitative designs to capture and examine the work of osteopaths in German-speaking countries in more detail.

The overall aim of this study was to make a substantial contribution to the largely unexplored profession of osteopathy in the German-speaking countries. Structured, basic research was necessary to implement this project. The first steps were taken in the framework of the study, “Characteristics, Opportunities, and Challenges of Osteopathy (COCO) in the Perceptions of Osteopaths in Germany, Austria, and Switzerland: Protocol for a Comprehensive Mixed Methods Study.” The study protocol was published in *JMIR Research Protocols* in 2019. The Characteristics, Opportunities, and Challenges of Osteopathy (COCO) project investigates how osteopaths in Germany, Austria, and Switzerland distinguish themselves from other medical professions and the characteristics of their work.

This study is a partial study of the COCO project, with a focus on the situation of osteopaths in Austria. Osteopaths practicing in Austria were asked about their professional profile and their professional practice. The following questions were of particular interest: (1) How do osteopaths from Austria describe osteopathy? (2) What are the challenges faced by osteopaths in Austria? and (3) What opportunities do the interviewees see for osteopathy in Austria?

**Methods**

**Design**

This qualitative study included the planning and implementation of guideline-based interviews with osteopaths practicing in Austria. Subsequently, a qualitative content analysis was performed according to Mayring [9]. A qualitative research design was selected to obtain questions relevant to the project that had not been considered before and views about the topic that had not yet been taken into consideration. The target of this qualitative partial study was the development of hypotheses. Accordingly, a relatively small sample of 8 to 10 participants could be used, because the results obtained shall be examined in subsequent studies with respect to their general validity using a quantitative study design [10].

To ensure the reporting quality regarding the research methodology of this qualitative study, COREQ (Consolidated Criteria for Reporting Qualitative Research) was used [11]. A checklist including the COREQ items taken into consideration has been attached to the paper (Multimedia Appendix 1). The registration identifier of the study is the International Registered Report Identifier: PRR2-10.2196/15399.

**Ethical Considerations**

This study (corresponding to partial study 1.2 in Figure 1), led by DM, has received ethics approval (S-287/2020) from the ethics committee of the University of Witten/Herdecke, Germany. Participants were not compensated for their participation.
Setting and Sampling

Related to the research topic, osteopaths in Austria were questioned through guideline-based interviews. All the interviewed osteopaths (10/10, 100%) had completed at least 4 years of training as osteopaths and practiced as osteopaths in Austria. Instead of asking individual osteopaths to participate, OEGO was contacted with the study project itself, thus avoiding cold-calling. In this way, the criterion of comprehensive training in osteopathy was fulfilled, because otherwise, the participants could not be members of the professional association. This procedure ensured that the participants had provided evidence of their competence. To obtain the widest range of views, the sample was intended to show a high degree of diversity among the participants. Therefore, a further criterion for the whole group of participants was sex distribution according to the population of osteopathic practitioners in Austria. According to the OEGO’s membership register, two-thirds of practicing osteopaths in Austria are female and one-third are male. Moreover, care was taken to ensure that the residences and workplaces of the participants were subject to as wide a geographical distribution as possible, so that district-specific phenomena could be excluded. As osteopathy is not an independent profession in Austria, the participants should include the different occupational groups that practice osteopathy. Inclusion and exclusion criteria were subsequently formulated (Textbox 1).

OEGO forwarded the contact details of 11 osteopaths. Appointments for an interview were made with 8 (73%) of the 11 osteopaths. No appointment could be made with 1 osteopath during the study period. From 1 other osteopath, no response to the request was received. Another 1 osteopath did not want to participate in the study; 2 new osteopaths were suggested by the osteopaths themselves. A total of 10 interviews were thus conducted.

All the participants were contacted via email and received a letter containing information about the course of the study, declaration of consent to participate in the study, and data protection declaration. The entire participation in the research project and the answering of individual questions was on a voluntary basis; nonparticipation did not lead to any disadvantages for the participants. The participants always had the option to end the survey (e.g., in the case of unexpected, stressful questions). Through the format of web-based survey, any increased risk of infection for the participants owing to the COVID-19 pandemic could be excluded. Explicit cancellation criteria for the project were not set.
Textbox 1. Inclusion and exclusion criteria.

<table>
<thead>
<tr>
<th>Inclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>- The individual has completed 4 years of osteopathic training.</td>
</tr>
<tr>
<td>- The individual is currently practicing osteopathy.</td>
</tr>
<tr>
<td>- The individual has provided consent to participate in the study.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>- The individual cannot be interviewed within the examination period.</td>
</tr>
<tr>
<td>- The individual does not have the technical equipment required for participation in a web-based survey.</td>
</tr>
<tr>
<td>- The individual demands compensation.</td>
</tr>
</tbody>
</table>

Development of the Interview Guidelines and Data Collection

The guideline-based interview was chosen as a suitable research tool, because the aim of the data collection was to obtain concrete statements about the practice of osteopathy in Austria. In addition, the use of a guideline increased the comparability of the individual data sets [12]. Furthermore, this procedure avoided the possibility that essential aspects of the research question might be overlooked in the interviews [13]. The interview guideline was developed by JM on the basis of the research problem and 2 previous qualitative, partial studies of the COCO project [14,15]. The interview guideline was developed by JM on the basis of the research problem and 2 previous qualitative, partial studies of the COCO project [14,15]. The questions were formulated as open questions (Textbox 2) and arranged according to the groups of topics (Textbox 3). In addition, sociodemographic data about the participants were collected. Alternative questions were prepared to be able to react flexibly to the course of the interview and to respond to the potential needs of the participants. To maintain the flow of the conversation, additional questions were developed in advance. Before beginning the interviews, a test interview was conducted with a German osteopath to test the interview guideline in practice and to improve the interview technique. Finally, the questionnaire was discussed and adapted together with JP, an osteopath with experience in qualitative research.

Textbox 2. Example interview questions.

- Where did you first hear about the osteopathic profession?
- How would you define osteopathy?
- In your opinion, what differentiates osteopathy from other professions?
- What does a typical osteopathic treatment look like for you?
- Should osteopath be its own profession?
- Are there any difficulties or problems that you face in your daily work life as an osteopath?
- Has the corona pandemic changed anything in your daily practice?
Textbox 3. Contents of the interview guideline.

<table>
<thead>
<tr>
<th>Training and work in osteopathy</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Educational background</td>
</tr>
<tr>
<td>• Motivation</td>
</tr>
<tr>
<td>• Training structure</td>
</tr>
<tr>
<td>• Work experience</td>
</tr>
<tr>
<td>• Acquisition structure</td>
</tr>
<tr>
<td>• Fields of activity</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Characteristics of osteopathy</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Definition</td>
</tr>
<tr>
<td>• Properties</td>
</tr>
<tr>
<td>• Differentiation of professional profile</td>
</tr>
<tr>
<td>• Competences</td>
</tr>
<tr>
<td>• Features</td>
</tr>
<tr>
<td>• Limits</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Challenges of osteopathy</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Health value</td>
</tr>
<tr>
<td>• Employment policy</td>
</tr>
<tr>
<td>• Obligations</td>
</tr>
<tr>
<td>• Restrictions</td>
</tr>
<tr>
<td>• Conflicts</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chances and opportunities in osteopathy</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Perspective</td>
</tr>
<tr>
<td>• Research</td>
</tr>
<tr>
<td>• Desires</td>
</tr>
</tbody>
</table>

Data Collection

Data were collected from November 29, 2021, to January 26, 2022. In total, 10 interviews were conducted. The interviews were conducted using the Zoom software (Zoom Video Communications; audio and video were recorded). Apart from research economy and temporal and local flexibility, interviews were primarily conducted on the web to protect the participants from infection during the COVID-19 pandemic. No other individuals were present during the interviews. No interview was repeated. The interviews were recorded as a video file for transcription. The participants agreed in writing to the archiving of the files until the end of the publication activities or up to a maximum of 5 years after data collection. The files were protected against unauthorized access and stored and evaluated on local data carriers of a password-protected computer. Only encrypted files were transferred among the study colleagues. Upon completion of the study project, the recordings of the interviews shall be deleted irrevocably. All the interviews were conducted by JM in German. The interviewer is male, holds a master of science degree, and has already published in 2019 within the COCO project. At the time of the survey, he worked independently as a physiotherapist in his own practice in Germany and was a doctoral student at the University of Witten/Herdecke. For this research project, JM was trained within a 3-part seminar at the Freie Universität Berlin regarding the collection and evaluation of qualitative data and the conduction of interviews. There was no previous personal relationship with any of the interviewees. The recordings were transcribed by JM. Transcription was performed according to pre-established rules, which were consistently observed, as there are no generally accepted transcription rules [16]. The rules were based on the transcription rules of Kuckartz and Rädiker [17] for computer-assisted evaluation. The participants were sent the transcript to gather their comments, if any, and to receive their final approval.

Data Analysis

On the basis of the results of the interview studies already conducted within the COCO project, deductive (ie, indirectly theory-driven) categories were formed first. It is indirect because the categories are descriptive and their definition is not the basis of a theory-driven description. As a first step, classical deductive codes were derived from the interview guideline (deductive category application). The transcripts were analyzed in the original language by means of content structuring. With the
help of the MAXQDA 2022 software (VERBI Software) [18], quotations that were relevant for the abovementioned research questions were categorized. Owing to an extensive interest in further knowledge, the preselected segments were expanded or specified by means of inductive categorization to deduce further important aspects (deductive-inductive categorization) [17]. The quotations were summarized in categories, which were subsequently grouped into high-level categories. The main strategies of data analysis are described in Textbox 4.

**Textbox 4. Data analysis strategies.**

<table>
<thead>
<tr>
<th>Coding</th>
<th>Discussion</th>
</tr>
</thead>
<tbody>
<tr>
<td>After developing an initial category system together (first, deductive; second, inductive), 2 complete material iterations were performed by 2 separate evaluators</td>
<td>Comparison and discussion of the results, with the main focus on the integration of different perspectives and the elimination of ambiguities</td>
</tr>
</tbody>
</table>

- **Quality control**
  - To check the quality of the category system (intercoder reliability) with its coding rules by means of the Cohen κ coefficient

- **Final iteration**
  - After the quality control step, a final, complete material iteration was performed by an evaluator using the final category system

---

**Shared Coding**

A first interview was coded by 2 evaluators together (JM and UW), and the category system was inductively expanded. An initial category system was developed, and coding rules were determined on the basis of 2 partial studies (eg, Figure 1—partial studies 1.1 and 1.3) [14,15] and the interview questions and guidelines developed specifically for this partial study. The result was an initial, deductive category system, and the first coding rules were defined. The deductively determined categories were based on the research questions and increased based on inductive subcategories during the evaluation. Then, the first iteration was run with the entire material, followed by further inductive categorization (UW). A second evaluator (JM) ran the second iteration of the entire material based on the category system resulting from the first material iteration. The results were then compared and discussed, with the main focus being on the integration of different perspectives and the elimination of ambiguities. The code definitions and anchor examples were also revised in this step. Finally, the category system was standardized. This shared coding from the beginning of the study was intended to increase the intersubjectivity of the statement.

**Quality Control**

Subsequently, of the 10 interviews, 3 (30%) were selected via lot procedure as a subsample to check the quality of the category system with its coding rules by means of the Cohen κ coefficient. The determination of the Cohen κ coefficient is a method for checking the intercoder reliability [19]. UW received the 3 allotted coded interviews from JM. The segment boundaries of the encodings were retained during this iteration. In this way, the evaluator was able to recognize which text parts were encoded and arrange these according to the categories. Without random adjustment, a match of 62.1% between the 2 encodings of this subsample was reached. The randomly adjusted coefficient was 0.61. Pursuant to Altman [20], accordance is considered as “good” in the case of a κ coefficient of 61-80. After this quality control, as there was good accordance; a final complete material iteration was performed by an evaluator (JM) using the final category system.

**Results**

**Overview**

The result of this study was a system of categories in which the statements of the participants were classified and subcategorized according to the research questions regarding the groups of topics, such as characteristics, challenges, and opportunities of the osteopaths practicing in Austria. Sociodemographic data and more general statements about the practice of osteopathy were also classified in the main category, “training and work in osteopathy.” Out of current concern, the osteopaths were also asked about the influence of the COVID-19 pandemic on their work. On the basis of the interesting statements, we decided to dedicate a special category to this topic.

The final category system consists of 71 categories with a total of 783 encoded text passages. A definition was created for each individual category, and a quotation was recorded as an anchor example.

Only a part of these results could therefore be described in this paper. The printed quotations have been translated into English and serve for illustration purposes only. The question whether individual statements of the participants (O1 to O10) represent the entire population of osteopaths practicing in Austria shall be subject to further investigation.

**Training and Work in Osteopathy**

A total of 10 osteopaths practicing in Austria were interviewed—6 (60%) women and 4 (40%) men. Of the 10 osteopaths, 3 (30%) were physicians and 7 (70%) were physiotherapists. Of the 10 interviewees, 9 (90%) worked independently in their own practices and 1 (10%) had an employment relationship at the time of the interview. Only 10% (1/10) had other employees. The longest interview lasted 61
minutes, and the shortest interview lasted 29 minutes. The average duration of the interviews was 46 (SD 9.2) minutes. The participants’ characteristics are described in Table 1.

Table 1. Participants’ characteristics (N=10).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Participants, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>4 (40)</td>
</tr>
<tr>
<td>Female</td>
<td>6 (60)</td>
</tr>
<tr>
<td>Profession</td>
<td></td>
</tr>
<tr>
<td>Physician</td>
<td>3 (30)</td>
</tr>
<tr>
<td>Physiotherapist</td>
<td>7 (70)</td>
</tr>
<tr>
<td>Degree</td>
<td></td>
</tr>
<tr>
<td>Diploma in osteopathy</td>
<td>3 (30)</td>
</tr>
<tr>
<td>Master of science</td>
<td>6 (60)</td>
</tr>
<tr>
<td>Nondegree</td>
<td>1 (10)</td>
</tr>
<tr>
<td>Training facility</td>
<td></td>
</tr>
<tr>
<td>Vienna School of Osteopathy, Vienna, Austria</td>
<td>8 (80)</td>
</tr>
<tr>
<td>European College of Osteopathy, Munich, Germany</td>
<td>1 (10)</td>
</tr>
<tr>
<td>The International Academy of Osteopathy, Darmstadt, Germany</td>
<td>1 (10)</td>
</tr>
<tr>
<td>Clinical experience (y)</td>
<td></td>
</tr>
<tr>
<td>1-5</td>
<td>2 (20)</td>
</tr>
<tr>
<td>6-15</td>
<td>5 (50)</td>
</tr>
<tr>
<td>&gt;15</td>
<td>3 (30)</td>
</tr>
</tbody>
</table>

Characteristics of Osteopathy

On the basis of the descriptions provided by the participants regarding the properties of osteopathy, the following categories were developed: “definition of osteopathy,” “patient profile,” “anchor personalities and literature,” and “limits of the treatment technique.”

Definition of Osteopathy

When asked about the definition of osteopathy, several participants had difficulties in explaining the concept:

Yes, it’s really a difficult question. [O1; item 35]

The explanation of the term was mostly based on the manual work, origin of the word, differentiation or overlapping with other professional groups, or citation of definitions of third parties. Often, reference was made to the philosophy of osteopathy, holism of the treatment method, and activation of self-regulating forces. There was no uniform definition of osteopathy among the answers of the interviewees.

Patient Profile

Most of the interviewed osteopaths treat patients of all ages:

Oh, everything actually, there are patients of all ages. From...three-month-old babies to over 90-year-old men, women, so I couldn’t paint a typical picture. [O2; item 54]

Common indications for child treatments mentioned by the interviewed osteopaths are sleep disorders, torticollis, scoliosis, asthma, abdominal colic, and plagiocephaly. The treatment of adults was mainly based on the diagnosis or leading symptoms from the orthopedic area: back and neck pain and joint pain. Neurological diseases such as Parkinson disease were also mentioned. Moreover, patients were regularly treated for headache, migraine, tinnitus, chronic pain, craniofacial dysfunction, abdominal pain, and hormonal imbalances or if wishing to become pregnant. However, internal diseases such as sinusitis, bronchitis, and cystitis were also treated by the physician O2 on the basis of osteopathic methods.

The indication of the treatment was usually given by the patients’ treating physician:

Patients are often assigned by the doctors, meaning that the doctor writes a prescription with a recommendation to contact a certain therapist. [O8; item 51]

A participant working as a general practitioner in addition to his osteopathic activity also acquired patients during his regular consultation as a physician:

And actually, many of those who go to the general practitioner’s clinic in the village, they come to me, too. [O2; item 54]

It appears that, in general, a broad medical field was covered by osteopathy. The selection of the appropriate therapist seemed to depend on personal recommendations of others, on the therapeutical possibilities in the patients’ vicinity, and on the training or specializations of the osteopaths.
Because the patients who come to me come by word of mouth, yes. [O5; item 47]

Everything else is, I think, very average, that is, all the people who come to me do so because I am in their vicinity. [O3; item 38]

Patients who travel a long way mainly come because of endocrinological, metabolic problems.... gynaecological problems.... that is to say, where...the focus of my...training has mainly been during recent years. [O3; item 38]

Anchor Personalities and Literature

It was noticeable that many participants referred to other individuals when answering questions about osteopathy in theory and practice. These “anchor personalities” seem to have a great influence on the self-image of osteopaths in Austria and have therefore been included in a separate category. Both historical personalities from the history of osteopathy and currently active osteopaths were repeatedly mentioned:

But there will always be people who really care about this innermost quality of osteopathy.... And just as osteopathy has developed from Still to Sutherland, Becker, Viola Frymann.... or Mitchell and Jim Jealous now..., so it will continue to develop. [O5; item 84]

With regard to the self-study of osteopaths, primarily only the German-language journals of osteopathy were mentioned as reading material. Of the 10 participants, only 1 (10%) indicated that they regularly read an English journal:

Yes, I regularly read the two journals, DO and Osteopathische Medizin. [O5; item 88]

Limits of Osteopathy

There was agreement regarding the limits of osteopathy. The primary treatment of structural injuries or the care of patients with cancer without medical supervision were clearly mentioned by the interviewees as limits of osteopathic activity. However, patients were given osteopathic treatment nevertheless:

Yes, of course I see the limits in the pathologies that are there. If there is actually...an osteoarthritis that is simply there and will not vanish, osteopathy shall certainly have its limits; one can perhaps relieve the pain, but the osteoarthritis cannot be cured by osteopathy, now can it. I also see limits in some diseases. [O7; item 56]

O9 differentiated upon request that it is not the diagnosis that is decisive for the objective of treatment when it comes to whether osteopathic treatment is indicated or contraindicated:

Well, that depends on the objective.... It all depends totally on the objective. If I say that I want to treat coxarthrosis curatively, I think that we shall soon reach our limits with osteopathy; if we do a control X-ray after six months, we shall see that it is still coxarthrosis.... But if I say that I want to improve the quality of life, I would treat them nevertheless. The question is what the objective is. [O9; item 41]

O4 brought another aspect to mind:

We must not exceed the limits...of our own competence. That is very important. Unfortunately, many colleagues do this by suddenly giving dietary recommendations, by suddenly recommending medicines...or by talking patients out of taking medicines. Especially now when it comes to vaccination. [O4; item 75]

Challenges of Osteopathy

The challenges mentioned for the field of osteopathy can be classified mainly into the categories, “identity problem,” “disagreements within the osteopathic community,” “research,” “training quality management,” and “conflicts and difficulties.”

Identity Problem

One of the great challenges of osteopathy is its unclear definition and lack of differentiation from other professional groups. As O6 clearly pointed out, osteopathy has an identity problem:

I would say that the identity problem is the most important issue....The definition is the most difficult question and no one can answer that. And if we can’t answer that and don’t deal with it, how can we argue what we are if we ourselves don’t know exactly what we are. [O6; item 49]

Disagreements Within the Osteopathic Community

The identity problem or the problem of missing a uniform professional self-image might be based on disagreements within the osteopathic community described by several participants. Overall, 2 groups can be identified among practicing osteopaths: on the one hand, the structurally working osteopaths and, on the other hand, the cranially or biodynamically working osteopaths:

Yes, there is really a gap between biodynamic osteopathy and structural osteopathy. [O7; item 104]

This conflict might be decisive not only in terms of a common definition but also with respect to a possible recognition of osteopathy in terms of professional policy. O9, who was involved in professional policy, feared that these disputes might even prevent recognition:

The problem concerning regulation is - and that’s simply the case now and that’s also the elephant in the room about which no one is talking -- the problem with regulation has always been cranio.... You cannot say it openly, but it was always the problem of craniosacral therapy, no matter who I talked to. [O9; item 93]

Those participants who worked biodynamically were more critical toward regulation:

If one tries now to take this out of this mental...source,.... I see the risk that it is practically shifted into evidence-based, as important as that is, well, but only into evidence-based, visible and perceivable dimensions, then osteopathy shall lose its...
Research

In this context, O1 pointed out that evidence-based research can only substantiate a certain part of osteopathy scientifically, whereas other aspects might be lost:

A good scientific basis in order to argue how...many benefits osteopathy has... in the end... academisation probably cannot be avoided and will certainly be necessary. Even if all these developments are not entirely without risks. That is, the risk of losing sight of the holistic aspects of osteopathy. [O1; item 81]

Training Quality Management

Many osteopaths considered existing weaknesses in the training courses and their structures as a further obstacle. According to O4, a central aspect is the inadequate teaching of the skills for scientific work at osteopathy institutes and, thus, the lack of evidence-based research in osteopathy:

[Oh my] the training...We should learn from the beginning, not only during the last year when we have to write a Master Thesis, we should learn from the beginning what it means to work in an evidence-based way, to do research...It works in physiotherapy and is continuously getting better there, but in osteopathy...At the beginning, we never learn to deal with available studies, it is a matter of training. From the beginning, not only during the fifth year shortly before the Master Thesis, we should have the first lessons in statistics. [O4; item 113]

Conflicts and Difficulties

The lack of clarity regarding the profession is evident in the differentiation with respect to other professional groups. With regard to the settlement for osteopathic treatment with health insurance providers, several participants also reported potential for conflicts. Osteopathic treatment is often provided on the insurance providers, several participants also reported potential differentiation with respect to other professional groups. With the lack of clarity regarding the profession is evident in the differentiation with respect to other professional groups.

Many colleagues work as physiotherapists, they also charge for osteopathy as physiotherapy, and yes, they are refunded in this way. And as a result, osteopathy is also a little...less in the focus than it should be. I’ve been working for 20 years now, I’m only writing osteopathic invoices.... But...of course...I understand the problem. If somebody has fewer patients...and...has to charge for...physiotherapy, I absolutely understand the situation.... But...these problems are of course...long-burning issues. [O5; item 78]

No, [the bill] of course says physiotherapy and remedial massage, because otherwise the patient doesn’t get his/her money from the insurance company. For the insurance company, well, this is ok or it is tolerated. I have already received the feedback from many patients that they told the company that they went to an osteopath, and the health insurance said that of course that can’t be billed, [but] we shall write physiotherapy and remedial massage and then that’s it. [O8; item 55]

When asked about the challenges for osteopaths in general, the physicians working as osteopaths did not report any difficulties related to their practice. A physician and osteopath, in contrast, was aware of the potential for conflict:

Yes, of course I know that. I have a bonus, because I’m simply a doctor. And of course, osteopaths that aren’t doctors have greater difficulties and are often rejected...well,...because they are no medical doctors in a manner of speaking and...there are obviously difficulties. [O1; item 101]

Another participant stated it even more clearly:

No, I’m a doctor, I have...no restrictions. [O9; item 85]

Opportunities in Osteopathy

Regarding the questions about the opportunities and chances in osteopathy, most of the statements could be classified into the categories “professional profile” and “position in the health care system.” A central opportunity in osteopathy is the installation of an independent profession. Almost all the osteopaths explicitly formulated the desire for their own professional profile. However, there was disagreement about the questions regarding where and how this profession should be integrated into the health care system or which competences it should include:

In the midst of the other health professions..., well, I don’t see us as special consultants, as it is now in America, for example. But I see us as a health profession next to physiotherapists, occupational therapists... [O4; item 81]

In this context, many possible applications were mentioned for the field of osteopathy. An osteopath saw a great opportunity in the prevention of diseases:

Concerning also prevention..., I believe that osteopathy has an enormous potential for people’s health by simply doing something really good and also really preventing things...follow-up problems or operations or God knows what...I see a huge opportunity there. [O10; item 107]

O2 also attributed the potential for cost reduction to preventive osteopathy. From his point of view, examinations and medical consultations might be reduced:

I see a huge and very central importance of osteopathy in primary care...and I am convinced...from my daily experience that an incredible number...of diagnostic measures or specialist care...might be avoided if people were primarily also treated by osteopaths. [O2; item 68]

Whether osteopathy actually contributes to disease prevention and can thus also lead to cost reduction or relief for the health care system is to be investigated using clinical study designs.
on the effectiveness of the treatment method itself. The position of osteopathy in the health care system and its differentiation from other professional groups have also not been uniformly described by practitioners in other countries.

The COVID-19 Pandemic

For current reasons, the participants were questioned about the effects of the COVID-19 pandemic, existing since March 2020, on their professional activities. Similar to a magnifying glass, crises very often reveal the weaknesses and failures of structures and concepts; however, they can also show their viability and strengths. Most respondents described the time of the COVID-19 pandemic with the lockdowns and the associated measures as a turning point in their practice. However, none of the participants described economic losses or existential fear:

Well, the time during Covid-19 wasn’t easy at all. [O5; item 75]

At the time of the survey, everybody had to wear a face mask, patients and osteopaths alike. The interviewees not only described the difficult communication with the patients because of the mask but also mentioned limitations during examination and treatment. Certain treatments, for example, techniques relating to the mandibular joint, could not be performed for patients wearing a face mask:

I believe that a lot of communication is lost through the mask, because you don’t see the whole face of the patient. Of course, you have communication through the eyes, but there is still a barrier; a lot is lost...It already starts with and continues during inspection: you only see half of the face and in the case of jaw problems, I have to take down the mask first. [O8; item 107]

Almost all the interviewees described a change in the clientele of their patients. Stress, sleep disorders, headaches, and dysfunctions of the mandibular joint were increasingly mentioned:

Psychosocial stress is increasing immensely...This in turn results...in sleep disorders...Mandibular joint problems due to stress, but also - and this is my own observation - because you constantly want to push around this mask if you have to wear it all day...I think that this has a huge influence...[O6; item 79]

Teaching also seemed to be affected by the protection measures for pandemic control. An osteopath reported that teaching on inpatients at the hospital ceased. The question of whether the osteopathic treatment of inpatients in institutions was disturbed by a lack of external osteopaths remained unanswered:

Prior to the lockdown, our osteopathic child centre also paid visits to the neonatal ward...where we treated premature babies. Unfortunately, this is not possible at the moment. [O1; item 47]

Owing to the cancellation of congresses and courses or their transfer to the digital world, interviewees experienced a gap in their personal training plans:

Of course, I...repeatedly attended courses. However, I scarcely did so in the last two years, actually...[O5; item 88]

This can only be a small insight into the impact of the pandemic in the field of osteopathy. The effects of the COVID-19 pandemic on osteopathic care should be investigated systematically in the next few years.

Discussion

Principal Findings

This study identified numerous aspects, possibilities, and opportunities in osteopathy in Austria from the point of view of the osteopaths practicing in Austria.

In our survey, the typical osteopath presents as female and has previously worked as a physiotherapist, as previous studies have found [8]. This is consistent with other surveys from Europe regarding osteopathy. Moreover, in accordance with a study from Italy, the typical osteopath practicing in Austria works independently in their own practice and without employees [21].

The osteopaths interviewed usually found it difficult to define osteopathy. The respondents were not able to provide a uniform definition of osteopathy. Many respondents even expressed difficulties in precisely describing their profession. Nevertheless, recurring patterns can be recognized in the explanations given by the respondents.

The participants attempted to define osteopathy by drawing a distinction or differentiation from other professions and using third-party definitions. Furthermore, the philosophy of osteopathy, various osteopathic concepts or models of thought, the holistic nature of the treatment method, and activation of the patient’s self-healing powers are often referred to. A possible reason for the heterogeneous attempts at explanation may lie in the difference in training and previous education. A recent study showed that only 17% of osteopaths surveyed in Austria identified themselves “exclusively” as osteopaths [8]. Therefore, there is a suspicion that, as our study also showed, the basic profession and nonregulation have a major influence on self-image. We observed a fundamental distinction between therapeutic and medical osteopaths.

From our point of view, the clear statements regarding the disagreements within the osteopath community were surprising. The conflicts do not remain in the specialist circles of osteopathy, but they even extend to the level of professional policy. The question is whether this is a country-specific observation for Austria. In their study in Australia in 2018, for example, Blaic et al [22] found disagreement about the specialization of osteopaths; however, it did not result in the splitting of osteopaths into 2 separate groups. The belief patterns and paradigms of individual treatment techniques that influence professional identity are not new in osteopathy [23]. The fact that, according to the osteopaths interviewed, these intraprofessional conflicts exist even on the political level or are the reason for nonregulation is remarkable. An increasing number of European countries regulate the professional practice of osteopathy. Therefore, it remains to be investigated whether this dispute itself has an influence on the nonregulation of
osteoopathy in Austria. However, conflicts and different opinions within a professional group are not inherent in osteopathy; these also exist in other medical professions such as chiropractic [24,25].

The general development of a profession is not only subject to cultural, historical, and social influences but also to the question of gender [26]. In this context, this study indicates a large influence of anchor personalities on the self-image of the interviewed osteopaths. It is remarkable that the anchor personalities mentioned are almost exclusively men. The historical context is worth noticing here, because Andrew Taylor Still, the founder of osteopathy, explicitly promoted equality between men and women already in the 19th century, in contrast to many other universities or teaching institutes during that time. He expressly included women in his courses [27]. In this context, it should be noted that in other health professions, although the practitioners are predominantly women, the leadership positions are often mainly occupied by men [28]. It is therefore not surprising that most users and practitioners of alternative medicine are women if their health needs are not being met by scientific medicine [29]. This becomes problematic when these professions are or become patriarchally dominated to match scientific standards [30].

To answer the questions about the origin of these conflicts and to deal with these in the future, we believe that a systematic and country-specific scientific analysis will be required. Conflicts in the health care system not only have the potential to weaken a profession but can also have a stimulating influence if understood as an opportunity [31].

Most of the osteopaths surveyed were in favor of a legal regulation of the profession. Under certain circumstances, osteopathically trained physiotherapists could benefit more from this, as they currently still need a physician’s order to be able to practice with legal certainty.

However, nonregulation also has also some advantages—no applications for licenses, no obligation for regular further training, and unregulated pricing for treatment. With integration into the health care system, some participants fear deterioration owing to possible low or lower payment by health insurance companies.

Training quality management and studies in the subarea of osteopathy were also mentioned as challenges in this context. In Italy, Sweden, and Australia, the transfer of scientific results to the practical work of osteopaths has already been systematically investigated in a country-specific manner [32-34]. The openness to evidence-based practice (EBP) appears to exist among practicing osteopaths on a transnational basis, but the skills in dealing with the former vary from country to country.

A study of EBP from Spain characterized the skills of the participants to deal with EBP as being rather low. This might be related to the lack of legal regulations and the inadequate transfer of knowledge in the training institutions [35]. The situation regarding osteopathy is similar in Austria. Additional country-specific studies are required to identify conclusions and connections.

The different situation in everyday practice owing to the COVID-19 pandemic and the respective infection protection measures also had an influence on the daily work of osteopaths. Several interviewees realized an evident change in the patients’ profile. Although economic damage or fear for their professional existence were not explicitly described, most osteopaths working independently were themselves responsible for the implementation of the legal measures in their practices. The impact of the pandemic on the daily work in practice seems to have been less considerable than the impact on the field of training in osteopathy. As a large part of practical teaching occurs with patients under supervision, it is difficult to implement in a web-based format. The impact of the pandemic on clinical research at universities or universities of applied sciences remains to be examined.

In the case of further investigations in this area, we recommend a specific distinction of the participants between physicians and physiotherapists practicing as osteopaths. As there is no uniform training or legal regulation of osteopathy in Austria, only physicians and physiotherapists trained in osteopathy exclusively practice osteopathy. The results of this study suggest that there are evident differences between these 2 professional groups regarding, for example, patient acquisition, conflict management, and cooperation with other professional groups.

The extent to which the individual statements made by the interviewees represent the entirety of osteopaths practicing in Austria will be further investigated. The protocol of the COCO project describes the further procedures. The results of the qualitative partial studies (studies 1.1, 1.2, and 1.3 in Figure 1) will be combined in a following study to verify the results of the qualitative partial studies in relation to the population [36]. We will develop a standardized questionnaire as a measuring instrument.

An important feature of this study is the methodology, including 2 evaluators who completed the entire evaluation process. Through this approach, intersubjectivity increased and new, inductively formed categories were created. During this phase, many aspects of the research problems could be identified and categorized. The intercoder reliability was tested and found to be viable within this study. With another material iteration, the category system can be further refined, and the intercoder reliability can be further increased by optimized code definitions.

Limitations
First, it should be noted that the results of this study do not necessarily allow conclusions to be drawn about the entirety of osteopaths in Austria, as this is not an evaluation of representative surveys with large numbers of participants. Nevertheless, certain tendencies seem to emerge when statements by osteopaths appear to be congruent, that is, confirm each other or complement each other in a meaningful way. The sample represents the entirety of osteopaths in Austria well. Most respondents were women and physiotherapists [8]. Nevertheless, bias cannot be dismissed with such a specific sample. However, they give an idea about how osteopaths in Austria think, and the results obtained can serve as a hypothesis for large quantitative studies to test.
Conclusions

It is difficult to characterize the community of osteopaths in Austria conclusively. On the one hand, there is a great deal of agreement about the urgency regarding regulatory legislation for their profession, a necessary revision of training structures, and the specific promotion of scientific studies of osteopathy. However, when it comes to the concrete practice of osteopathy, deep trenches and even strong disputes have occurred among osteopaths.

The following question remains to be answered: what is “correct” or “true” osteopathy? If we consider that osteopathy has derived from various sources; that its founder did not give a final answer to the question about what he understood by osteopathy; and that each discipline is constantly developing, solely through the different osteopaths practicing, it appears that this question cannot be answered completely.

Apart from this issue, there is another and equally sensitive question, that is, whether and how the different parties can or even must be brought together for the regulation of their profession, which is desired by most of them. The different professional origins of osteopaths should also be considered. With regard to binding legal regulations, which would not least strengthen the professional image, mutual understanding seems to be imperative. Perhaps such an understanding might also lead to greater political weight for osteopathy, which it urgently needs, not only in terms of legal regulations but also to be able to promote important research projects.

The question arises as to whether the conflicts within osteopathy, in particular, with their possible professional-political consequences and the immense influence of the basic profession in the practice of osteopathy, are a country-specific phenomenon for Austria. However, there is a lack of studies in German-speaking countries with comparable qualitative designs to assess the work of osteopaths in more detail. We are therefore planning a meta-synthesis of qualitative studies with the aim of generating new theoretical insights from the accumulation of study results. Both the studies from the COCO project itself and other relevant literature can be used for the meta-synthesis.

To the best of our knowledge, the COCO project is the largest mixed methods study project on the osteopathic profession in German-speaking countries. The category system with its reliability check can be used as a basis for a repetition of the study. Such a research project would also be interesting if the profession was regulated formally and substantially in the near future. The results presented in this paper are not only intended to serve as a basis for further studies but also to provide universities, schools, professional associations, and politicians with an insight into the situation of osteopaths in Austria.

Conflicts of Interest

None declared.

Multimedia Appendix 1

COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist. [PDF File (Adobe PDF File), 870 KB - humanfactors_v11i1e45302_app1.pdf ]

References

2. Osteopathic Education – Osteopathic International Alliance. URL: https://oialliance.org/resources/osteopathic-education/ [accessed 2023-12-24]


18. VERBI software. MAXQDA. 2022. URL: https://www.maxqda.com/about [accessed 2023-12-12]


28. Feldman DE, Bellavance C, Fr...
Abbreviations

COCO: Characteristics, Opportunities, and Challenges of Osteopathy
COREQ: Consolidated Criteria for Reporting Qualitative Research
EBP: evidence-based practice
OEGO: Austrian Society for Osteopathy (Österreichische Gesellschaft für Osteopathie)

© Jonas Manschel, Jan Porthun, Ulrike Winkler, Jean Marie A T Beuckels, David Martin. Originally published in JMIR Human Factors (https://humanfactors.jmir.org), 17.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Human Factors, is properly cited. The complete bibliographic information, a link to the original publication on https://humanfactors.jmir.org, as well as this copyright and license information must be included.
Trust in and Acceptance of Artificial Intelligence Applications in Medicine: Mixed Methods Study

Daria Shevtsova, Anam Ahmed, Iris A Boot, Carmen Sanges, Michael Hudecek, John J L Jacobs, Simon Hort, Hubertus J M Vrijhoef

1Panaxea bv, Den Bosch, Netherlands
2Vrije Universiteit Amsterdam, Amsterdam, Netherlands
3Universitätsklinikum Würzburg, Würzburg, Germany
4Ortec bv, Zoetermeer, Netherlands
5Fraunhofer Institute for Production Technology, Aachen, Germany

Corresponding Author:
Anam Ahmed, MSc
Panaxea bv
Pettelaarpark 84
Den Bosch, 5216 PP
Netherlands
Phone: 31 639421854
Email: anam.ahmed@panaxea.eu

Abstract

Background: Artificial intelligence (AI)–powered technologies are being increasingly used in almost all fields, including medicine. However, to successfully implement medical AI applications, ensuring trust and acceptance toward such technologies is crucial for their successful spread and timely adoption worldwide. Although AI applications in medicine provide advantages to the current health care system, there are also various associated challenges regarding, for instance, data privacy, accountability, and equity and fairness, which could hinder medical AI application implementation.

Objective: The aim of this study was to identify factors related to trust in and acceptance of novel AI-powered medical technologies and to assess the relevance of those factors among relevant stakeholders.

Methods: This study used a mixed methods design. First, a rapid review of the existing literature was conducted, aiming to identify various factors related to trust in and acceptance of novel AI applications in medicine. Next, an electronic survey including the rapid review–derived factors was disseminated among key stakeholder groups. Participants (N=22) were asked to assess on a 5-point Likert scale (1=irrelevant to 5=relevant) to what extent they thought the various factors (N=19) were relevant to trust in and acceptance of novel AI applications in medicine.

Results: The rapid review (N=32 papers) yielded 110 factors related to trust and 77 factors related to acceptance toward AI technology in medicine. Closely related factors were assigned to 1 of the 19 overarching umbrella factors, which were further grouped into 4 categories: human-related (ie, the type of institution AI professionals originate from), technology-related (ie, the explainability and transparency of AI application processes and outcomes), ethical and legal (ie, data use transparency), and additional factors (ie, AI applications being environment friendly). The categorized 19 umbrella factors were presented as survey statements, which were evaluated by relevant stakeholders. Survey participants (N=22) represented researchers (n=18, 82%), hospital staff (n=3, 14%), and policy makers (n=3, 14%). Of the 19 factors, 16 (84%) human-related, technology-related, ethical and legal, and additional factors were considered to be of high relevance to trust in and acceptance of novel AI applications in medicine. The patient’s gender, age, and education level were found to be of low relevance (3/19, 16%).

Conclusions: The results of this study could help the implementers of medical AI applications to understand what drives trust and acceptance toward AI-powered technologies among key stakeholders in medicine. Consequently, this would allow the implementers to identify strategies that facilitate trust in and acceptance of medical AI applications among key stakeholders and potential users.

(JMIR Hum Factors 2024;11:e47031) doi:10.2196/47031
Trust and Acceptance

Artificial Intelligence

Artificial intelligence (AI) is commonly defined as a computer system that uses statistical models, diverse algorithms, and self-modifying systems to make predictions and decisions based on its own aggregated experience. It can therefore perform tasks that usually require or even surpass the human level of intelligence [1,2]. AI has been increasingly integrated in the health care sector, where it helps with administrative workflows, diagnostic image analysis, robotic surgery, and clinical decision-making. Consequently, medical AI applications allow, amongst other things, earlier disease detection, patient-tailored treatments, and more efficient follow-ups, which should drive the health care costs down upon implementation [3]. Although medical AI applications provide various advantages to the current health care system, such as increased efficiency and improved workflows [4], there are also various challenges associated with AI implementation. For instance, a large share of potential users has concerns over privacy issues [5,6]. Equity and fairness are other important concerns, since there is a risk of perpetuating bias within data sets being adopted by AI technology [5,7,8]. Further, the implementation of AI into medical practice raises the question of accountability, since it is currently unclear whether technology developers, hospitals, or regulators should be responsible for mistakes or undesirable outcomes from the use of an AI application.

Trust and Acceptance

To ensure successful implementation of medical AI applications, it is essential to build trust in and acceptance of AI technology among its users [9,10]. In this study, a model of key drivers of trust in and acceptance of AI systems was used [11]. According to this model, trust is influenced by 4 drivers: current safeguards, job impact of AI, familiarity of AI, and AI uncertainty. Current safeguards indicate the belief that current regulations and laws are adequate for ensuring the safety of AI and the protection of people who use it. Job impact refers to the belief that there will be more jobs generated than eliminated due to AI implementation. Familiarity with AI is the level of understanding of how AI technology works and how AI applications are used. These 3 drivers have a positive influence on trust, with current safeguards being its strongest driver. The fourth driver, AI uncertainty, impacts trust in a negative way. It implies the belief that the impact of AI on society is unpredictable and the technology is still not fully explored. Overall, these drivers influence the extent to which people trust the AI system and believe it to be trustworthy. Trust, then, is a large contributor to the level of acceptance, which is the extent to which people accept or approve of AI and are willing to use it without resistance [11]. In the scientific literature, trust can be defined in different ways. In this study, we used literature-derived definitions of trust and acceptance in the context of AI implementation, namely:

- Trust is the belief of an individual that an AI application will do what it promises [12,13].
- Acceptance is the willingness of an individual to use the AI application in medicine [14].

Therefore, it can be argued that acceptance of an AI application depends on trust people have toward this technology [11,15,16]. At the same time, people can often accept their usage of technologies without necessarily trusting them [17]. Therefore, it is important to consider the 2 concepts separately as well as together.

Overall, widespread trust in and acceptance of an AI application is crucial for successful introduction and implementation of the technology. Failure to ensure trust in and acceptance of AI technology would pose the risk of “stifling innovation” and causing unnecessary “opportunity costs” [18]. The lack of trust in AI applications in medicine impedes their adoption in health care, compounded by inadequate public assurance and attention to concerns, thereby exacerbating these challenges. In addition, the anticipated benefits of AI-based innovations can coexist with significant acceptance barriers [15,18-21].

Investigating what factors contribute to trust in and acceptance of AI technology in medicine would help us understand how to make the implementation and regulatory approval of AI-powered advanced therapy manufacturing systems as efficient as possible. This can be achieved by collecting insights into stakeholders’ perspectives with regard to trust and acceptance toward medical AI applications [2,22]. Factors contributing to trust and acceptance toward medical AI applications can be attributed a different weight by various groups of stakeholders with distinct roles in AI.

Study Objectives

Since AI applications are still relatively new, users and providers are hesitant to trust and accept this new technology without restrictions. As for the future implementation of AI applications in treatment centers, it is essential that stakeholders (eg, clinicians, researchers, hospital staff) accept and trust the innovative AI-based manufacturing platform. Therefore, the aim of this study was first to identify the factors related to trust in and acceptance of AI technology in medicine and second to assess the relevance of those factors among relevant stakeholders in medicine.

Methods

Study Setting

This study is part of the European Union’s (EU) Horizon 2020 project AIDPATH (AI-driven Decentralized Production for Advanced Therapies in the Hospital; grant agreement number 101016909) [22,23]. It is an upcoming state-of-the-art AI application in hospitals, which aims to develop an AI-driven, automated chimeric antigen receptor T cell (CAR-T) manufacturing platform at the point of care as a treatment for acute leukemia and lymphoma. In CAR-T therapy, the patient’s
own T cells are removed, genetically modified, and reinfused into the patient in order to find and eliminate tumour cells. Current production is characterized by laborious manual process steps, complex logistics, and a lack of process understanding. This results in long delivery times (up to 21 days) and high costs (approx €320,000, or US $347,890, per treatment) [24,25]. For this reason, AIDPATH is developing a system to fully automate the manufacturing process, from the provision of patient cells to the injection directly in the hospital. An important building block for effective and equitable manufacturing is AI. AI can provide essential process insights into the cell’s characteristics and behavior. This offers a significant benefit for adaptive control of the whole process and the design of personalized process protocols. Furthermore, AI can assist cost-effective platform operation in a smart manufacturing hospital by improving manufacturing schedules and resource management [26]. In general, successful implementation of AIDPATH would serve as an example of an effective AI technology that automates the production and delivery of advanced therapy medicinal products (ATMPs). Furthermore, AI-powered technology can form the basis for a deployable platform for further pilot trials in multiple hospitals and would create a model innovation system for smart manufacturing hospitals [2,22].

In this study, to meet the study objectives, a rapid literature review was conducted, followed by a survey.

Rapid Literature Review

A rapid literature review of peer- and non-peer-reviewed publications was conducted to identify factors related to trust in and acceptance of AI applications used in medicine. As an alternative method to systematic reviews, a rapid review allows for accelerated synthesis of up-to-date evidence, while efficiently informing latest findings in recent health care research [27]. The peer- and non-peer-reviewed literature needed to be published between 2012 and 2022 in English. Data on attitudes toward AI in relation to prognosis, diagnosis, treatment, and care were included. The search was performed in PubMed/MEDLINE with the following search syntax: ((trust) OR (acceptance) OR (attitude) OR (perspective) OR (perception)) AND ((AI) OR (artificial intelligence) OR (machine learning) OR (deep learning)) AND (((prognosis) OR (diagnosis) OR (treatment) OR (care)) OR ((medic*) OR (clinic*) OR (hospital) OR (smart hospital) OR (health care)) AND ((survey) OR (questionnaire) OR (interview))). The reason for inclusion of only survey-, questionnaire-, or interview-based research in the search terms was due to their direct relevance to our research objectives.

In the non-peer-reviewed literature search, similar terms and time frame of publication were used and the first 10 pages on the Google Search engine were examined to identify other relevant papers and reports by (non)governmental and research organizations. This allowed the study findings to be applicable to a broad range of medical AI applications. Papers were screened, and data were extracted by 2 authors (DS and AA). The selected literature was analyzed to identify key trends and explanatory factors related to trust and acceptance toward medical AI applications. The factors were then grouped into 4 categories: human-related, technology-related, legal and ethical, and additional factors. These factor groups formed the basis of the survey designed to investigate factor relevance. This was performed independently by 2 authors (DS and AA).

Survey

The survey was reported in accordance with the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) guidelines [28]. The survey in English assessed the relevance of the factors related to trust in and acceptance of novel AI applications in medicine. The survey started with an introduction to AI applications in medicine and AIDPATH, followed by 7 general questions on each participant’s background, including gender, age, the country they worked in, years of experience, the stakeholder group they belonged to, their familiarity with AI applications in medicine, and their general view on AI. In the last question, the following distinction was made between the answer options: “I embrace AI” meant welcoming and using AI as a constituent part of their work or life, “I approve of AI” implied that the participant agreed with the use of AI in their work or life but did not use it themselves, and “I accept AI” referred to acknowledging the use of AI in work or life but not being ready to fully approve it.

In the core section of the survey, the definitions of trust and acceptance were provided as a reference for participants. The core part also consisted of 2 identical lists of 19 factors related to trust in and acceptance toward AI applications in medicine. Each factor was categorized into human-related, technology-related, legal and ethical, or additional factors. Human-related factors were linked to AI professionals assessed the relevance of the type of organization the AI professionals were affiliated to and the purpose to innovate with a specific AI application. With respect to health care professionals, the factors were related to the knowledge of AI applications and the attitude toward AI application usage in medicine. In relation to patients, the relevance of the following factors was assessed: general knowledge of AI applications in medicine, the attitude toward AI application usage in medicine, and the patient’s age, gender, and level of education. Furthermore, participants were asked to evaluate the relevance of transparency between all parties involved in AI application use. Technology-related factors related to the performance of AI applications in medicine, the possibility of their integration into existing clinical workflows, a clear balance of risks and benefits of the AI applications, and the explainability and transparency of processes and outcomes. The legal and ethical factors were related to the adequacy of regulations and governance of AI applications in medicine, data use transparency, and clear accountability and responsibility for an AI application. The additional factors were concerned with the environmental sustainability of AI applications and AI’s impact on job availability. For each factor, participants could indicate each factor’s relevance to trust in and acceptance of AI applications from their stakeholder perspective using a Likert scale of 1-5, where 1 stood for “not relevant,” 3 for “not irrelevant, nor relevant,” and 5 for “relevant.” Throughout the survey, “relevant” meant being highly significant for ensuring trust in or acceptance of AI applications, while “irrelevant” meant no significance. The N/A (not applicable) option was available as well for each factor. Open questions at the end of both sections...
allowed participants to suggest other relevant factors related to trust in or acceptance of AI applications that were not mentioned in the survey. Furthermore, the participants were invited to suggest any other factors, different from trust, deemed important for acceptance of AI applications in medicine.

**Sampling**

Using the convenience sampling method [29], AIDPATH Consortium members were requested to invite stakeholders in their network but outside the AIDPATH Consortium to fill in the survey on the SurveyMonkey platform. The survey was distributed by email to members of relevant stakeholder groups to capture their professional perspectives (eg, clinicians, scientists, and policy makers). Data were collected from April to May 2022 and analyzed using Microsoft Excel.

**Data Collection and Analysis**

After participants were asked to rate the relevance of each factor from 1 (irrelevant) to 5 (relevant), the mean score of each factor was determined by assigning each response a weight from 1 to 5. Next, means scores were calculated by finding an average of the sum of response values for each question. To visualize the survey responses and compare the scores for each factor included in the survey, a spider diagram was charted. This provided an overview of the factors’ relevance and their relative importance in influencing both trust and acceptance toward AI applications in medicine. In addition, a scatter plot was created to obtain an overview of the interrelationship between the relevance to trust (x axis) in and acceptance (y axis) of AI applications in medicine. The plot allowed us to identify the degree of relevance of each factor in relation to both trust and acceptance. To classify the factors based on their relevance, score ranges were established. Factors with mean scores from 1 to 3 were considered to be of low relevance, while factors from 4 to 5 were deemed of high relevance. The open-question responses were considered when interpreting numerical data.

**Ethical Considerations**

Under Dutch law, no ethical approval was required according to Article 1b of the Dutch Medical Research in Human Subjects Act [30]. However, all participants were informed about the study objectives, their verbal consent was obtained, and all data were processed anonymously. All responses were recorded anonymously. Participants were informed of their right to withdraw from the study at any time without any consequences. They were not financially compensated.

**Results**

**Rapid Literature Review**

The literature search (Figure 1) yielded 301 hits in the PubMed database and 105 hits through gray literature search and snowballing. After screening titles and abstracts, 284 (70%) records were excluded. After full-text screening, 90 (73.8%) records were excluded primarily due to the absence of concepts of trust or acceptance and a lack of factors related to trust or acceptance in the main text or data-containing figures. As a result, 32 (26.2%) papers and reports [7,9-12,15,16,19,21,31-53] were included in the data analysis.

Overall, the rapid review identified a total of 110 factors related to trust and 77 factors related to acceptance toward medical AI technology. The full list of factors identified through the rapid review with corresponding studies can be found in Multimedia Appendix 1. Tables 1-4 show all factors from the rapid review, each with the frequency of its appearance in the literature and the corresponding overarching umbrella factors. Some factors from a single study are repeated in the same category in Tables 1 and 2 on trust and Tables 3 and 4 on acceptance or within the same category (eg, health care professionals and patients subsections of the human-related factors section). The most frequently reported human-related factors related to trust (Tables 1 and 2) in medical AI applications were knowledge and understanding of AI by health care professionals and knowledge and education of AI among patients. In terms of technology-related factors, accuracy, transparency, reliability, safety, and explainability of medical AI applications and their functioning appeared most often in the literature. Regarding legal and ethical factors, the most frequently occurring factors included fairness and equity of medical AI technology and the privacy and security of personal data handled by the AI systems. The most frequently presented human-related factors related to acceptance (Tables 3 and 4) of medical AI technology were the perceived usefulness and provision of better medical services by the AI technology. Regarding technology-related factors linked to acceptance, performance expectancy, design and output quality, and transparency were stated in the literature most often. A wide range of legal and ethical factors were mentioned in the literature, including adequate regulations of medical AI technology, protection and security of patients’ data, and the allocation of accountability and responsibility for the (mal)functioning of an AI application. There were additional factors related to trust in and acceptance of medical AI technology (Tables 1-4). These included replacement of doctors by machines that lack a human touch and moral support, labor market implications, and environmental sustainability. Three studies also highlighted that acceptance of a medical AI application is directly related to trust in the AI application. Overall, there were fewer factors related to acceptance than those related to trust, whereas most of the overarching umbrella factors were fully represented in both tables. Therefore, an identical list of umbrella factors allocated within the 4 categories (human-related, technology-related, legal and ethical, and additional factors) was used in the survey for investigating the relevance of factors for both trust in and acceptance of AI applications in medicine.
Figure 1. PRISMA flow diagram of the rapid review literature screening. AI: artificial intelligence.

Table 1. Human-related factors related to trust (N=110) in medical AI\textsuperscript{a} applications (22/32, 68.8%, studies).

<table>
<thead>
<tr>
<th>Factor category and factors from the rapid review</th>
<th>Umbrella factors used in the survey</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AI professionals</strong></td>
<td></td>
</tr>
<tr>
<td>AI company/provider (n=2, 9.1%)</td>
<td>Type of institution/organization of AI professionals (eg, university, technology company, commercial organization)</td>
</tr>
<tr>
<td>AI role (n=1, 4.5%); perceived helpfulness (n=1, 4.5%)</td>
<td>The purpose to innovate with a specific AI application in medicine (eg, financial vs societal)</td>
</tr>
<tr>
<td><strong>Health care professionals</strong></td>
<td></td>
</tr>
<tr>
<td>Knowledge and understanding of AI (n=6, 27.3%); education (n=3, 13.6%)</td>
<td>Knowledge of AI applications in medicine (eg, by means of training and education)</td>
</tr>
<tr>
<td>Expectation of AI (n=1, 4.5%); perceived actionability (ie, clear recommendation for action; n=1, 4.5%); user’s social network (n=1, 4.5%); user’s media consumption (n=1, 4.5%)</td>
<td>Attitude toward AI application usage in medicine (eg, agreeableness, openness, conscientiousness, engagement)</td>
</tr>
<tr>
<td><strong>Patients informed about AI application usage in the hospital</strong></td>
<td></td>
</tr>
<tr>
<td>Knowledge/education about AI (n=5, 22.7%); awareness of AI (n=2, 9.1%)</td>
<td>General knowledge of AI applications in medicine</td>
</tr>
<tr>
<td>Openness (to AI health care technologies and to judgments of potential benefits and harms; n=1, 4.5%); perceived benefit and lower concern (n=1, 4.5%); user’s social network (n=1, 4.5%); user’s media consumption (n=1, 4.5%)</td>
<td>Attitude toward AI application usage in medicine (eg, agreeableness, openness, conscientiousness)</td>
</tr>
<tr>
<td>Gender (n=2, 9.1%); age (n=1, 4.5%); type of user (n=1, 4.5%)</td>
<td>Age, gender, level of education</td>
</tr>
<tr>
<td><strong>All parties</strong></td>
<td></td>
</tr>
<tr>
<td>Clinicians and patients interaction during AI integration (n=1, 4.5%); human agency and oversight (n=1, 4.5%)</td>
<td>Transparency between all involved parties (AI professionals, health care professionals, patients)</td>
</tr>
</tbody>
</table>

\textsuperscript{a}AI: artificial intelligence.
Table 2. Other factors related to trust (N=110) in medical AI applications (22/32, 68.8%, studies).

<table>
<thead>
<tr>
<th>Factor category and factors from the rapid review</th>
<th>Umbrella factors used in the survey</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technology-related factors</strong></td>
<td></td>
</tr>
<tr>
<td>Accuracy (n=7, 31.8%); reliability (n=5, 22.7%); safety (n=4, 18.2%); design and output quality (n=2, 9.1%); performance expectancy (n=2, 9.1%); ability (n=1, 4.5%); perceived functionality (n=1, 4.5%); self-efficacy (n=1, 4.5%); tool itself (n=1, 4.5%)</td>
<td>Performance of AI applications in medicine (reproducibility of outcomes, accuracy)</td>
</tr>
<tr>
<td>Auditability (n=1, 4.5%); customizability (n=1, 4.5%); understandability (n=1, 4.5%); ease of integration into clinical workflows (n=1, 4.5%); convenience of use (n=1, 4.5%); usability (n=1, 4.5%); (over)alerting and excessive false-positive rate (n=1, 4.5%)</td>
<td>Possibility of integration of AI applications into existing clinical workflows</td>
</tr>
<tr>
<td>Risk and impact mitigation (n=1, 4.5%)</td>
<td>Clear balance of risks and benefits of the AI application</td>
</tr>
<tr>
<td>Transparency (n=6, 27.3%); explainability (n=5, 22.7%); evidence strength (n=2, 9.1%); benevolence (n=2, 9.1%); complexity (n=2, 9.1%); interpretability (n=2, 9.1%); integrity (n=1, 4.5%); predictability (n=1, 4.5%); trialability (n=1, 4.5%); trustworthiness (n=1, 4.5%)</td>
<td>Explainability and transparency of the processes and outcomes</td>
</tr>
<tr>
<td><strong>Legal and ethical factors</strong></td>
<td></td>
</tr>
<tr>
<td>Fairness and equity (n=8, 36.4%); adequate regulations, legislation, and governance (n=3, 13.6%); ethical/legal implications (n=1, 4.5%)</td>
<td>Adequacy of the regulations and governance of AI applications in medicine</td>
</tr>
<tr>
<td>Personal data privacy and security (n=8, 36.4%); data used to train AI/cognitive bias (n=2, 9.1%); data sensitivity (n=1, 4.5%); respect and preservation of human dignity (n=1, 4.5%)</td>
<td>Data use transparency</td>
</tr>
<tr>
<td>Accountability (n=3, 13.6%); power-control balance (n=1, 4.5%)</td>
<td>Clear accountability and responsibility of the AI application (machine vs human responsibility)</td>
</tr>
<tr>
<td><strong>Additional factors</strong></td>
<td></td>
</tr>
<tr>
<td>Environmental sustainability (n=1, 4.5%)</td>
<td>Environment-friendly AI application</td>
</tr>
<tr>
<td>Replacement of doctor/lack of human touch and moral support when evaluated by AI alone (n=1, 4.5%); labor market implications (n=1, 4.5%)</td>
<td>Impact on job availability (machines replacing humans)</td>
</tr>
</tbody>
</table>

*AI: artificial intelligence.*

https://humanfactors.jmir.org/2024/11/e47031
Table 3. Human-related factors related to acceptance (N=77) of medical AI\(^a\) applications (14/32, 43.8%, studies).

<table>
<thead>
<tr>
<th>Factor category and factors from the rapid review</th>
<th>Umbrella factors used in the survey</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AI professionals</strong></td>
<td></td>
</tr>
<tr>
<td>AI company/provider (n=1, 7.1%); brand impact (n=1, 7.1%)</td>
<td>Type of institution/organization of AI professionals (eg, university, technology company, commercial organization)</td>
</tr>
<tr>
<td>Perceived usefulness (n=3, 21.4%); better medical services/understanding of disease (n=3, 21.4%); improve the quality of people's lives (n=2, 14.3%); medical costs (n=2, 14.3%); AI role (eg, saving patients' time; n=1, 7.1%); miniaturization of hardware (n=1, 7.1%)</td>
<td>Purpose to innovate with a specific AI application in medicine (eg, financial vs societal)</td>
</tr>
<tr>
<td><strong>Health care professionals</strong></td>
<td></td>
</tr>
<tr>
<td>Knowledge and understanding of AI (n=1, 7.1%)</td>
<td>Knowledge of AI applications in medicine (eg, by means of training and education)</td>
</tr>
<tr>
<td>Behavioral intention to use (n=2, 14.3%); effort expectancy (n=2, 14.3%); perceived ease of use (n=2, 14.3%); perceived usefulness (n=2, 14.3%); intrinsic motivation (n=1, 7.1%); interest in AI (n=1, 7.1%); professional identity (n=1, 7.1%); concerns about benefit to patient care (n=1, 7.1%); general impression of AI (n=1, 7.1%)</td>
<td>Attitude toward AI application usage in medicine (eg, agreeableness, openness, conscientiousness, engagement)</td>
</tr>
<tr>
<td><strong>Patients informed about AI application usage in the hospital</strong></td>
<td></td>
</tr>
<tr>
<td>Knowledge/education about AI (n=1, 7.1%); awareness of AI (n=1, 7.1%)</td>
<td>General knowledge of AI applications in medicine</td>
</tr>
<tr>
<td>Behavioral intention to use (n=2, 14.3%); general impression (n=1, 7.1%); Interest in topic (n=1, 7.1%)</td>
<td>Attitude toward AI application usage in medicine (eg, agreeableness, openness, conscientiousness)</td>
</tr>
<tr>
<td>Age (n=1, 7.1%)</td>
<td>Age</td>
</tr>
<tr>
<td><strong>All parties</strong></td>
<td></td>
</tr>
<tr>
<td>Expectations of others (n=2, 14.3%)</td>
<td>Transparency between all involved parties (AI professionals, healthcare professionals, patients)</td>
</tr>
</tbody>
</table>

\(^a\)AI: artificial intelligence.

Table 4. Other factors related to acceptance (N=77) of medical AI\(^a\) applications (14/32, 43.8%, studies).

<table>
<thead>
<tr>
<th>Factor category and factors from the rapid review</th>
<th>Umbrella factors used in the survey</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technology-related factors</strong></td>
<td></td>
</tr>
<tr>
<td>Performance expectancy (n=4, 28.6%); design and output quality (n=4, 28.6%); accuracy (n=2, 14.3%); efficiency (n=1, 7.1%)</td>
<td>Performance of AI applications in medicine (reproducibility of outcomes, accuracy)</td>
</tr>
<tr>
<td>Perceived ease of use (n=2, 14.3%); user-friendliness (n=2, 14.3%); actual system use (n=1, 7.1%); compatibility (n=1, 7.1%); facilitating conditions (n=1, 7.1%)</td>
<td>Possibility of integration of AI applications into existing clinical workflows</td>
</tr>
<tr>
<td>Perceived risk (n=1, 7.1%)</td>
<td>Clear balance of risks and benefits of the AI application</td>
</tr>
<tr>
<td>Transparency (n=3, 21.4%); explainability (n=2, 14.3%); evidence strength (n=1, 7.1%); trustworthiness (n=1, 7.1%)</td>
<td>Explainability and transparency of the processes and outcomes</td>
</tr>
<tr>
<td><strong>Legal and ethical factors</strong></td>
<td></td>
</tr>
<tr>
<td>Adequate regulations, legislation and governance (n=2, 14.3%); ethical risks (n=1, 7.1%); political support (n=1, 7.1%)</td>
<td>Adequacy of the regulations and governance of AI applications in medicine</td>
</tr>
<tr>
<td>Data protection/security (n=2, 14.3%); patients’ consent to the continuous collection and processing of data (n=1, 7.1%)</td>
<td>Data use transparency</td>
</tr>
<tr>
<td>Accountability and responsibility (n=2, 14.3%); tort liability (n=1, 7.1%)</td>
<td>Clear accountability and responsibility of the AI application (machine vs human responsibility)</td>
</tr>
<tr>
<td><strong>Additional factors</strong></td>
<td></td>
</tr>
<tr>
<td>Replacement of doctor/lack of human touch and moral support when evaluated by AI alone (n=1, 7.1%)</td>
<td>Impact on job availability (machines replacing humans)</td>
</tr>
<tr>
<td>Trust in AI applications (n=3, 21.4%)</td>
<td>Acceptance emerging from trust</td>
</tr>
</tbody>
</table>

\(^a\)AI: artificial intelligence.
Survey

Participants

A total of 22 respondents participated in the survey, of which 18 (82%) completed the questions on trust and 15 (68%) completed the questions on acceptance. No reasons were provided for not completing the survey. Table 5 shows the characteristics of the survey participants, the majority (n=21, 95%) of whom came from European countries, were aged from 40 to 60 years, and had 0-10 or 21-30 years of professional experience.

Participants were mainly slightly (n=7, 32%) or moderately (n=8, 36%) familiar with AI-based devices used for clinical purposes (Figure 2). In thinking about AI, 9 (41%) of the participants indicated that the statement “I accept AI” best represents their view, followed by “I approve of AI” (n=6, 27%) and “I embrace AI” (n=5, 23%); see Figure 3.

Table 5. Characteristics of the participants (N=22).

<table>
<thead>
<tr>
<th>Characteristic and type of participant</th>
<th>Participants, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stakeholder group</strong></td>
<td></td>
</tr>
<tr>
<td>Researchers</td>
<td>18 (82)</td>
</tr>
<tr>
<td>Technology providers</td>
<td>5 (23)</td>
</tr>
<tr>
<td>Hospital staff</td>
<td>3 (14)</td>
</tr>
<tr>
<td>Policy makers</td>
<td>3 (14)</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>8 (36)</td>
</tr>
<tr>
<td>Male</td>
<td>13 (59)</td>
</tr>
<tr>
<td>Prefer not to say</td>
<td>1 (5)</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
</tr>
<tr>
<td>≤30</td>
<td>2 (9)</td>
</tr>
<tr>
<td>31-39</td>
<td>3 (14)</td>
</tr>
<tr>
<td>40-49</td>
<td>6 (27)</td>
</tr>
<tr>
<td>50-59</td>
<td>5 (23)</td>
</tr>
<tr>
<td>≥60</td>
<td>6 (27)</td>
</tr>
<tr>
<td><strong>Country of work</strong></td>
<td></td>
</tr>
<tr>
<td>Netherlands</td>
<td>11 (50)</td>
</tr>
<tr>
<td>Germany</td>
<td>3 (14)</td>
</tr>
<tr>
<td>Ireland</td>
<td>2 (9)</td>
</tr>
<tr>
<td>Spain</td>
<td>2 (9)</td>
</tr>
<tr>
<td>France</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Hungary</td>
<td>1 (5)</td>
</tr>
<tr>
<td>India</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Italy</td>
<td>1 (5)</td>
</tr>
<tr>
<td><strong>Years of professional experience</strong></td>
<td></td>
</tr>
<tr>
<td>0-10</td>
<td>6 (27)</td>
</tr>
<tr>
<td>11-20</td>
<td>4 (18)</td>
</tr>
<tr>
<td>21-30</td>
<td>7 (32)</td>
</tr>
<tr>
<td>31-40</td>
<td>5 (23)</td>
</tr>
</tbody>
</table>

aParticipants sometimes represented more than 1 stakeholder group.
Figure 2. Familiarity with AI-based devices used for clinical purposes (N=22). AI: artificial intelligence.

Figure 3. Statement best representing participants’ view when thinking about AI (N=22). AI: artificial intelligence.

Relevance of Factors for Trust in and Acceptance of AI

In Table 6, the mean scores per factor for its relevance to trust and acceptance are shown. Figure 4 demonstrates a spider diagram with the 19 summarized statements and the corresponding mean scores of relevance to trust in and acceptance of AI applications in medicine. The degrees of relevance of the factors related to trust and to acceptance closely followed each other for all but 1 (5.3%) of the 19 factors. Only the type of AI organization was slightly more relevant to trust than to acceptance toward AI applications in medicine. In Figure 5, a scatter plot displays the combined relevance of the factors related to trust (x axis) and acceptance (y axis) toward medical AI applications. Of the 19 factors included in the survey, 3 (16%) were found to have, on average, low relevance, while the other 16 (84%) had high relevance. There were no factors relevant to acceptance and irrelevant to trust (upper-left section in the plot) and vice versa (bottom-right section in the plot).
Table 6. Mean (SD) factor relevance to trust and acceptance (N=22).

<table>
<thead>
<tr>
<th>Factor</th>
<th>Trust</th>
<th>Acceptance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of AI&lt;sup&gt;a&lt;/sup&gt; organization</td>
<td>4.72 (0.75)</td>
<td>4.27 (0.88)</td>
</tr>
<tr>
<td>Purpose to innovate with AI</td>
<td>4.33 (0.84)</td>
<td>4.47 (0.64)</td>
</tr>
<tr>
<td>Clinicians’ knowledge about AI</td>
<td>4.50 (0.51)</td>
<td>4.73 (0.46)</td>
</tr>
<tr>
<td>Clinicians’ attitude towards AI</td>
<td>4.50 (0.51)</td>
<td>4.47 (0.64)</td>
</tr>
<tr>
<td>Patients’ knowledge of AI</td>
<td>4.17 (0.62)</td>
<td>4.20 (0.68)</td>
</tr>
<tr>
<td>Patients’ attitude toward AI</td>
<td>4.28 (0.57)</td>
<td>4.47 (0.64)</td>
</tr>
<tr>
<td>Patients’ age</td>
<td>3.17 (1.04)</td>
<td>3.47 (1.19)</td>
</tr>
<tr>
<td>Patients’ gender</td>
<td>2.61 (1.14)</td>
<td>2.67 (1.05)</td>
</tr>
<tr>
<td>Patients’ education level</td>
<td>3.50 (0.99)</td>
<td>3.53 (1.19)</td>
</tr>
<tr>
<td>Transparency between all parties</td>
<td>4.61 (0.50)</td>
<td>4.47 (0.64)</td>
</tr>
<tr>
<td>Performance of AI</td>
<td>4.83 (0.38)</td>
<td>4.67 (0.62)</td>
</tr>
<tr>
<td>Possibility of AI integration into existing workflows</td>
<td>4.56 (0.62)</td>
<td>4.53 (0.83)</td>
</tr>
<tr>
<td>Clear balance of AI risks and benefits</td>
<td>4.67 (0.49)</td>
<td>4.60 (0.63)</td>
</tr>
<tr>
<td>Explainability and transparency of AI processes</td>
<td>4.78 (0.43)</td>
<td>4.60 (0.63)</td>
</tr>
<tr>
<td>Adequacy of AI regulations</td>
<td>4.72 (0.57)</td>
<td>4.60 (0.83)</td>
</tr>
<tr>
<td>Data use transparency</td>
<td>4.61 (0.50)</td>
<td>4.67 (0.49)</td>
</tr>
<tr>
<td>Clear accountability and responsibility of AI</td>
<td>4.61 (0.61)</td>
<td>4.80 (0.41)</td>
</tr>
<tr>
<td>Environmental friendliness of AI</td>
<td>3.83 (0.79)</td>
<td>3.87 (0.92)</td>
</tr>
<tr>
<td>Impact on job availability</td>
<td>3.78 (1.11)</td>
<td>4.07 (0.88)</td>
</tr>
</tbody>
</table>

<sup>a</sup>AI: artificial intelligence.

**Figure 4.** Mean scores of factors’ relevance to trust in and acceptance of AI applications in medicine (N=19). Score=1 means irrelevant; score=3 means not irrelevant, nor relevant; and score=5 means relevant. AI: artificial intelligence.
Factors of Low Relevance

With regard to patients informed about AI application usage in the hospital, participants deemed the patient’s gender, age, and educational level to be of low relevance to trust in and acceptance of novel AI applications in medicine.

Factors of High Relevance

The majority of factors were deemed highly relevant to trust in and acceptance of novel AI applications in medicine by participants. Regarding AI professionals, it was observed that the type of institution or organization where AI professionals originated from (e.g., university, technology company, commercial organization) and the purpose to innovate with a specific AI application in medicine (e.g., financial, societal, or clinical purpose) were considered relevant. Participants reported that the involvement of health care professionals having knowledge of the AI application (e.g., by means of training and education) and the purpose to innovate with a specific AI application in medicine (e.g., financial, societal, or clinical purpose) were considered relevant. Participants reported that the involvement of health care professionals having knowledge of the AI application (e.g., by means of training and education) and the purpose to innovate with a specific AI application in medicine (e.g., financial, societal, or clinical purpose) were considered relevant. Participants reported that the involvement of health care professionals having knowledge of the AI application (e.g., by means of training and education) and the purpose to innovate with a specific AI application in medicine (e.g., financial, societal, or clinical purpose) were considered relevant.

Other Factors

Participants were able to share other factors that were not mentioned in the survey questions. Factors related to trust included solidarity and understanding the bias and interdomain knowledge of AI in software development, data science, and medicine. Other factors related to acceptance were the extent to which alternatives to AI applications are available, the length of experience, transparency about limitations, reproducibility, risks evaluation, resources, and the fear to use an AI application (i.e., fear of making the wrong decision or fear of losing control).

Discussion

Principal Findings

This study aimed to identify factors related to trust and acceptance toward medical AI applications by means of a rapid...
review and to assess their relevance by conducting a survey. Through the rapid review, 19 key factors related to trust in and acceptance of AI-powered medical technologies were identified and subsequently grouped into 4 categories. Our survey results highlight that of all examined factors, 84% (16/19) were considered highly relevant to trust in and acceptance of novel AI applications in medicine. Only the patient’s gender, age, and education level (3/19, 16%) were deemed to be of low relevance by participants.

**Comparison With Prior Work**

Previous studies have reported that trust in technology is mainly determined by human characteristics [54], technology-related factors [55], and environment-related factors [56], which is in line with the findings of our survey. According to Tran et al [57], who investigated patients’ perceived benefits and risks of using digital and AI technology in health care, the important factors to consider are the new technologies requiring an overhaul of the current health care system as human care is being replaced by machines and health care professionals becoming sufficiently equipped with increasing knowledge of AI technology. This highlights the importance of several survey factors, including the possibility of AI integration into existing clinical workflows. Therefore, setting features such as understandability, usability, and user-friendliness (factors that frequently appeared in the rapid review) by AI professionals as key goals in the development of novel AI applications would increase the chances of successful integration of AI technology into health care systems. Tran et al [57] also highlighted the increasing importance of data transparency toward patients and the acute need for clear accountability and responsibility (machine vs human responsibility) concerning the new technology, which also goes hand in hand with the findings from the rapid review and the survey [57]. The patient data handling must be organized in accordance with the existing data protection regulations in respective countries, with additional precautionary measures due to the sensitive nature of such medical data [57]. Shin et al [58] demonstrated that explainability of AI plays a big role in user trust and attitude toward AI. Explainability, along with transparency, was also found to be highly relevant in our study, especially in relation to the AI application processes and outcomes. In addition, Vourgidis et al [59] recommended that AI systems be regularly checked for being up to date, since today’s technology is continuously evolving. This again highlights the relevance of the education of health care professionals, since they are the primary users of medical AI technology and hence need to follow the developments in the field. Yang et al [49] found that gender is not relevant to trust in AI technology in medicine. This agrees with our finding that a patient’s gender has low relevance to trust in and acceptance of AI technology in medicine. Contrary to our findings, it has been reported that younger generations in general have more trust and are more likely to accept AI systems compared to older generations [11].

In our survey, the majority of participants were aged 40-60 years and above and they exhibited a solid awareness of and a positive attitude toward AI technology. Gillespie et al [11] also stated that highly educated people (university level) are more likely to trust and accept AI systems compared to those without a university degree. However, our survey showed that a patient’s educational level has low relevance to trust in and acceptance of medical AI applications.

**Strengths and Limitations**

To the best of our knowledge, this is the first study to use a rapid review of the latest literature to identify factors related to trust in and acceptance of AI applications in medicine in order to create a survey to evaluate their relevance and the attitudes of health care stakeholders toward implementation of medical AI applications. However, the study has several limitations. Since a large number of papers and reports in the rapid review did not provide sufficient context for the factors for trust or acceptance, there could have been an increased risk of personal bias during interpretation and categorization of those factors. Furthermore, some studies did not clarify whether the reported factors were related to only trust or only acceptance, which could also lead to possible misinterpretation. To minimize the effect of such bias and misinterpretation, a third reviewer (author HJMV) was consulted in such cases. Another limitation is the relatively small number of papers included in the rapid review, given the breadth of the topic. However, this rapid review was intentionally conducted focusing on the most relevant and recent literature to provide an initial overview and highlight key themes in a time-efficient manner. We aimed to provide a starting point that formed the basis for the survey. In addition, the number of participants included in the survey can be considered relatively low, which was caused by difficulties in recruiting participants and the time-constrained nature of the study. However, sufficient diversity in participant characteristics (ie, gender, age, country of work, and years of professional experience) was achieved, which could be considered more important in terms of validity of the study findings. Even though the survey benefited from a sample with a wide diversity in participant characteristics, one of the limitations to consider is the underrepresentation of certain stakeholder groups, in particular technology providers, policy makers, and hospital staff members other than clinicians. If these groups had been included in the survey, different patterns in factor relevance might have been observed, potentially shedding light on additional concerns or challenges associated with AI applications in medicine. Moreover, when considering the relevance of factors assessed through the survey, which were predominantly highlighted by researchers, it is important to note that these factors might be readily attainable or already well established within this specific stakeholder group. As a result, these factors may not necessarily represent challenges or barriers for this particular group, as they are already well versed in the aspects related to trust and acceptance.

**Recommendations for Future Research**

The results of this study can be valuable for various stakeholders involved in the implementation of novel AI applications, since trust and acceptance building remains a focus point throughout the different stages, including the pilot, implementation, evaluation, and monitoring phases of the process. In the survey, participants shared other factors related to trust in and acceptance of AI applications in medicine that were not included in the survey. However, due to a lack of context, it is not entirely clear what was meant by some of these factors; since these are
open to interpretation, follow-up research is required to better understand this. In addition, further research is needed to gain insight into the reasons participants considered factors to be of low or high relevance. Regarding the currently underrepresented stakeholder groups in the survey, more research is required to gain insight into the perspectives of policy regulators, technology providers, and hospital staff members. Next, once the implementation of a novel AI technology, such as the AIDPATH system, becomes clear from the trust and acceptance point of view, it would be beneficial to conduct a workshop with experts from the AI and biotechnology fields to identify technical challenges of implementation. This is crucial since, according to the survey results, the technical robustness and clarity of AI applications is a prerequisite for trust and acceptance exhibited toward this technology by stakeholders.

Recommendations for Implementation

By considering the factors that are most relevant in the AI technology adoption process, the implementers can facilitate trust in and acceptance of medical AI applications among their users and other stakeholders. Furthermore, the knowledge of the factors with high relevance to stakeholders can predict concerns the potential users might have regarding the new AI technology and act upon these concerns to implement the AI application efficiently and in a timely manner. There are several ways in how the results of the survey could be used by AI implementers, such as smart hospitals, to build trust and acceptance among various stakeholder groups. For instance, the highly relevant factor of knowledge and understanding of AI among health care professionals could be addressed by providing information about medical AI to clinicians in the form of conferences and educational workshops. These initiatives can ensure that health care professionals remain updated on significant changes in AI technology, facilitating its accurate utilization. Similarly, patients could be informed of medical AI technology through patient information initiatives in (smart) hospitals and within patient communities. The highly relevant technology-related factors could be used by technology developers and scientific researchers as guidance in the development of novel AI technology. For regulators and policy makers, it is crucial to know that users and other stakeholders consider data use transparency and fairness and equity to be of utmost importance regarding novel medical AI technology. Indeed, data privacy is a crucial and ever-so-present topic in legislation and regulations, but it needs to be constantly reviewed by policy makers due to the newness of AI in health care and the speed of its development. The legal aspects of software containing AI have been subjected to the Medical Device Regulation (MDR) [60]. For the acceptance of AI, its implementation in MDR-compliant solutions is invaluable. The tasks of policy makers could involve the risk assessment of various data breaches related to AI in medicine with continuous updating of regulations related to data security and privacy within the field of medical AI. Furthermore, both policy makers and AI professionals have to ensure the maintenance of fairness and equity of AI technology usage.

Conclusion

This study identified and assessed the relevance of factors for trust in and acceptance of novel AI applications in medicine. The survey demonstrated that the majority of the identified human-related, technology-related, and legal and ethical factors for trust in and acceptance of novel AI applications in medicine were considered by stakeholders to be of high relevance. Taken together, these findings and subsequent recommendations could be used by any implementers of medical AI, such as (smart) hospitals, AI technology organizations, biotechnology research institutes, and policy makers, to facilitate smooth and timely adoption of novel AI applications in medicine.

Acknowledgments

Our gratitude goes out to the AIDPATH (artificial intelligence [AI]-driven Decentralized Production for Advanced Therapies in the Hospital) Consortium partners, the Fraunhofer Institute for Cell Therapy and Immunology IZI, University College London, the Foundation for Research and Technology (FORTH)-Hellas, SZTAKI, Aglaris Cell SL, Sartorius Cell Genix GmbH, the Fundació Clinic per a la Recerca Biomèdica, IRIS Technology Solutions, Red Alert Labs, and ORTEC b.v. We would also like to thank the advisory board members of AIDPATH for providing feedback on the survey. Lastly, we thank the survey’s responders.

The paper was written within the framework of the European Union’s (EU) Horizon 2020 research and innovation program project AIDPATH (grant agreement number 101016909). All mentioned colleagues/companies are part of AIDPATH and therefore received funding from the EU within the scope of AIDPATH.

Data Availability

The data used and analyzed in this study are available from the corresponding author upon reasonable request.

Authors’ Contributions

DS was responsible for methods design, search strategy design, data acquisition, data extraction, data analysis, interpretation of data, and design and writing of the manuscript; AA, methods design, search strategy design, data acquisition, data extraction, data analysis, interpretation of data, regularly reviewing the work, design and writing of the manuscript, and providing feedback on the manuscript; IWAB, methods design, data acquisition, and providing feedback on the manuscript; CS, methods design, regularly reviewing the work, and providing feedback on the manuscript; MH, providing feedback on the manuscript and funding acquisition; JILJ, reviewing the work and providing feedback on the manuscript; SH, contributing to the Introduction chapter.
providing feedback on the manuscript, project administration, and funding acquisition; and HJM, concept and design of the overall study, quality assessment, interpretation of data, regularly reviewing the work, providing feedback on the manuscript, and manuscript final approval.

Conflicts of Interest

MH reports speaker honoraria from Novartis, Janssen, and Celgene/BMS and has participated in scientific advisory boards for Janssen and Celgene/BMS. MH is also listed as an inventor on patent applications and has been granted patents related to chimeric antigen receptor (CAR) technologies and CAR T cell therapy that have been filed by the Fred Hutchinson Cancer Research Center and the University of Wurzburg and that have been, in part, licensed to industry. In addition, MH is a cofounder and equity owner of T-CURX GmbH, Wurzburg, Germany. The remaining authors declare that they have no competing interests.

Multimedia Appendix 1

Identified factors and corresponding studies included in the rapid review.

References


34. Brady D. Building public trust to successfully deploy AI in EU healthcare. European Policy Centre. 2022 Jan 26. URL: https://www.epc.eu/content/PDF/2022/CHES_PB_v2-2.pdf [accessed 2023-12-04]


**Abbreviations**

- **AI**: artificial intelligence
- **AIDPATH**: AI-powered Decentralized Production for Advanced Therapies in the Hospital
- **CAR-T**: chimeric antigen receptor T cell
- **MDR**: Medical Device Regulation
A Machine Learning Approach with Human-AI Collaboration for Automated Classification of Patient Safety Event Reports: Algorithm Development and Validation Study

Hongbo Chen¹, MSc; Eldan Cohen¹, PhD; Dulaney Wilson², PhD; Myrtede Alfred¹, PhD

¹Department of Mechanical & Industrial Engineering, Faculty of Applied Science & Engineering, University of Toronto, Toronto, ON, Canada
²Department of Public Health Sciences, College of Medicine, Medical University of South Carolina, Charleston, SC, United States

Corresponding Author:
Myrtede Alfred, PhD
Department of Mechanical & Industrial Engineering
Faculty of Applied Science & Engineering
University of Toronto
27 King's College Cir
Toronto, ON, M5S 1A1
Canada
Phone: 1 4372154739
Email: myrtede.alfred@utoronto.ca

Abstract

Background: Adverse events refer to incidents with potential or actual harm to patients in hospitals. These events are typically documented through patient safety event (PSE) reports, which consist of detailed narratives providing contextual information on the occurrences. Accurate classification of PSE reports is crucial for patient safety monitoring. However, this process faces challenges due to inconsistencies in classifications and the sheer volume of reports. Recent advancements in text representation, particularly contextual text representation derived from transformer-based language models, offer a promising solution for more precise PSE report classification. Integrating the machine learning (ML) classifier necessitates a balance between human expertise and artificial intelligence (AI). Central to this integration is the concept of explainability, which is crucial for building trust and ensuring effective human-AI collaboration.

Objective: This study aims to investigate the efficacy of ML classifiers trained using contextual text representation in automatically classifying PSE reports. Furthermore, the study presents an interface that integrates the ML classifier with the explainability technique to facilitate human-AI collaboration for PSE report classification.

Methods: This study used a data set of 861 PSE reports from a large academic hospital’s maternity units in the Southeastern United States. Various ML classifiers were trained with both static and contextual text representations of PSE reports. The trained ML classifiers were evaluated with multiclass classification metrics and the confusion matrix. The local interpretable model-agnostic explanations (LIME) technique was used to provide the rationale for the ML classifier’s predictions. An interface that integrates the ML classifier with the LIME technique was designed for incident reporting systems.

Results: The top-performing classifier using contextual representation was able to obtain an accuracy of 75.4% (95/126) compared to an accuracy of 66.7% (84/126) by the top-performing classifier trained using static text representation. A PSE reporting interface has been designed to facilitate human-AI collaboration in PSE report classification. In this design, the ML classifier recommends the top 2 most probable event types, along with the explanations for the prediction, enabling PSE reporters and patient safety analysts to choose the most suitable one. The LIME technique showed that the classifier occasionally relies on arbitrary words for classification, emphasizing the necessity of human oversight.

Conclusions: This study demonstrates that training ML classifiers with contextual text representations can significantly enhance the accuracy of PSE report classification. The interface designed in this study lays the foundation for human-AI collaboration in the classification of PSE reports. The insights gained from this research enhance the decision-making process in PSE report classification, enabling hospitals to more efficiently identify potential risks and hazards and enabling patient safety analysts to take timely actions to prevent patient harm.

(JMIR Hum Factors 2024;11:e53378) doi:10.2196/53378
accident; accidents; black box; classification; classifier; collaboration; design; document; documentation; documents; explainability; explainable; human-AI collaboration; human-AI; human-computer; human-machine; incident reporting; interface design; interface; interpretable; LIME; machine learning; patient safety; predict; prediction; predictions; predictive; report; reporting; safety; text; texts; textual; artificial intelligence

Introduction

Since the publication of the seminal report on patient safety—Err Is Human [1], the importance of preventing adverse events in health care has been widely recognized. Adverse events refer to unintended or unexpected incidents that occur during hospital care that cause harm to a patient [2]. Common adverse events include complications, falls, and medication errors. These events can lead to prolonged hospital stays, permanent harm to patients, life-saving interventions, or even contributing to patient deaths [2,3]. Unfortunately, adverse events remain one of the top 10 leading causes of death and disability worldwide, resulting in 251,454 deaths annually in the United States alone [4]. In Organization for Economic Cooperation and Development (OECD) countries, 15% of total hospital activity is the direct result of adverse events [5]. The global cost of adverse events has been estimated at 42 billion USD annually [6].

Patient safety event (PSE) reporting systems, also called incident reporting systems, have been widely adopted in hospitals across the world as part of their efforts to mitigate adverse events and improve patient safety [7,8]. Multiple nations, including Canada, Japan, England, and Norway, have made it mandatory for hospitals to establish and maintain a PSE reporting system, either with individual health care systems or through centralized national incident reporting platforms [9]. The primary purpose of the PSE reporting system is to provide health care organizations with a centralized system for tracking and analyzing PSEs, thereby facilitating continuous learning and maintaining a record of PSEs for risk assessment and prevention [7,10]. PSE reporting systems are tools that allow frontline health care personnel to voluntarily report adverse events, near-misses, and unsafe conditions [11]. Each PSE report includes structured data, such as event types, patient harm level, date, and location of the event, as well as unstructured data, including a free-text section that contains the factual description of the event and the patient’s outcome [12]. Following submission, PSE reports are reviewed by relevant hospital staff, such as risk managers, patient safety analysts, nurse managers, physicians, and biomedical engineers, to identify areas for patient safety and quality improvement within the hospital [13].

Accurately classifying PSE reports into their appropriate event type is crucial to ensure that these reports are directed to the relevant patient safety analyst, support organizational learning, identify patterns and trends in adverse events, and ultimately prioritize measures to reduce adverse events [14,15]. An event type refers to a specific class of events that share common characteristics [16]. Examples of event types include falls, medication-related issues, and diagnosis errors [17,18]. PSE reporting systems may have upwards of 20 categories of events. The formulation of these classification taxonomies generally involves systematically grouping PSE reports based on common characteristics [19]. The descriptions of event types are not always readily accessible to PSE reporters and patient safety analysts [15]. Previous studies have found that the classification of PSE reports is inconsistent depending on the reporter’s profession, interpretation of the adverse event, and understanding of the PSE classification taxonomy [15,20]. Furthermore, 25% of PSE reports are labeled with vague or nonspecific categories such as “miscellaneous” and “other” and require time-consuming retrospective analysis for reclassification [21]. These problems are further exacerbated by the growing volume of PSEs reported [18,22]. For instance, hospitals in the state of New South Wales in Australia reported close to 195,000 PSEs in 2020 [23], while there were approximately 2.3 million PSEs reported to the National Reporting and Learning System in England from April 2021 to March 2022 [24].

In light of these challenges, it is imperative to find an efficient solution to ensure the reliable classification of PSE reports. Recent studies have used static text representations and supervised machine learning (ML) techniques to automate the PSE report classification [17,25,26]. However, static text representations ignore the ordering of the words and do not account for the differences in word meaning across different contexts. These limitations may result in suboptimal classification performance. With the emergence of deep learning, contextual text representation produced from transformer-based deep learning models has achieved state-of-the-art performance on a wide range of natural language processing tasks, including text classification [27]. The contextual representation of each word is based on its surrounding context within the text, allowing for a more accurate understanding of its usage across different contexts and facilitating knowledge transfer across languages [28]. Therefore, using contextual text representation in training ML classifiers presents a promising opportunity for achieving a more precise classification of PSE reports.

The integration of ML models into PSE reporting systems has important implications for human–artificial intelligence (AI) collaboration, given the roles of the incident reporter (front end) and patient safety analyst (backend). Various approaches for using ML classifiers can be developed, including at different levels of automation; however, unifying the strengths of both human expertise and AI offers the most promising route for effective implementation [29-31]. A crucial determinant for successfully implementing the human-AI collaboration approach is decision transparency [32,33], which is often referred to as explainability. Explainability is the concept that an ML model’s prediction can be explained in a way that human operators can comprehend and reconstruct the model’s reasoning [33]. Incorporating explainability techniques in human-AI collaboration is paramount as it facilitates a deeper understanding of the factors influencing the predictions, thereby fostering trust and understanding between human experts and AI systems. Therefore, embedding explainability into the...
human-AI collaboration holds significant potential for enhancing PSE report classification.

The main aim of this study is to examine the efficacy of contextual text representation in improving the accuracy of PSE report classification. To accomplish this, we trained, evaluated, and compared various ML classifiers with both static and contextual text representations. Additionally, we developed an interface to illustrate the integration of the ML classifier in an event reporting system to support human-AI collaboration for PSE report classification. Moreover, we enhanced the explainability of the ML classifiers by using an explainable AI technique. Furthermore, we have investigated the ML classifier’s performance under 2 conditions, differentiated by whether the explanation is valid for the predicted event type. Based on this analysis, we offer recommendations for optimizing human-AI collaboration in the context of PSE report classification.

### Table 1. Prevalence of patient safety event reports by event type in this study.

<table>
<thead>
<tr>
<th>Event type</th>
<th>Extracted reports (n=861), n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Care coordination or communication</td>
<td>186 (21.6)</td>
</tr>
<tr>
<td>Laboratory test</td>
<td>122 (14.2)</td>
</tr>
<tr>
<td>Medication related</td>
<td>89 (10.3)</td>
</tr>
<tr>
<td>Omission or errors in assessment, diagnosis, and monitoring</td>
<td>67 (7.8)</td>
</tr>
<tr>
<td>Maternal</td>
<td>58 (6.7)</td>
</tr>
<tr>
<td>Equipment or devices</td>
<td>56 (6.5)</td>
</tr>
<tr>
<td>Supplies</td>
<td>49 (5.7)</td>
</tr>
<tr>
<td>Total</td>
<td>627 (72.8)</td>
</tr>
</tbody>
</table>

### Data Preprocessing

The free-text section of PSE reports was preprocessed before feeding into ML classifiers as input features. The preprocessing procedures include text normalization, feature extraction, data splitting, and data augmentation (Multimedia Appendix 1 [28,34-39]).

### Classifier Training

A range of ML classifiers, including multinomial logistic regression (MLR), support vector machine (SVM), extreme gradient boosting, light gradient boosting, random forest (RF), k-nearest neighbor (KNN), and multilayer perceptron, were used for the classification of PSE reports. While SVM is a binary classifier, it is also capable of performing multiclass classification using the one-versus-one strategy. This involves treating the multiclass classification problem as a series of binary classification problems, creating $n \times (n - 1) / 2$ binary classifiers for each pair of classes, where $n$ represents the total number of classes, and the final classification is based on the majority vote of all binary classifiers. Extreme gradient boosting, light gradient boosting, and RF are tree-based ensemble algorithms that are commonly used in text classification tasks [17,40]. The KNN classifier predicts the class of a data point based on the majority class among its nearest neighbors in the training data set. Multilayer perceptron is a feedforward neural network consisting of multiple layers of interconnected neurons and trained using backpropagation.

To optimize the performance of ML classifiers, we used the 5-fold cross-validation grid search technique to identify the best combination of hyperparameters. During this process, a range of values of important hyperparameters (ie, regularization strength) is assessed with 5-fold cross-validation. For each combination of hyperparameters, the training set is randomly split into 5 distinct folds, and then the ML classifier is trained and evaluated 5 times, picking a different fold for evaluation every time and training on the remaining 4 folds. The optimized combination of hyperparameters is determined based on the average performance of the classifier on the $F_1$-score across the 5-fold cross-validation runs.

### Classifier Evaluation

We evaluated the performance of the trained classifiers on the testing set with standard classification metrics, including accuracy, precision, recall, $F_1$-score, and area under the receiver operating characteristic curve. We also evaluated classifiers on top-2 accuracy, which measures the proportion of predictions where the correct event type is among the top 2 highest probability event types predicted by the classifier. The definitions and mathematical formulas of the evaluation metrics are shown in Multimedia Appendix 2. Each of these metrics provides a distinct perspective on the performance of the
classifier, and collectively, they offer a comprehensive understanding of how well the classifier is functioning. Since we framed PSE report classification as a multiclass text classification problem, the precision, recall, $F_1$-score, and area under the receiver operating characteristic curve are computed for each class and combined using a weighted average where the weights correspond to the number of data points in each class.

**Development and Assessment of Explainability**

As the contextual text representation is generated from transformer-based neural network, which has a black box nature, we used the local interpretable model-agnostic explanations (LIME) technique to analyze the top-performing ML classifier trained with the contextual text representation. LIME is a post hoc, local perturbation technique that provides the explanation for a single prediction. LIME generates perturbed data by randomly removing words from a text document and trains a locally explainable model with perturbed data to simulate the original classifier’s prediction [41]. By measuring how the classifier’s prediction changes under these perturbations, LIME reflects the contributions of each word to the prediction. The importance of each word can then be assessed for a single prediction, revealing whether the ML classifier has learned to use relevant words for classifying PSE reports. We used LIME to generate explanations for the top-performing classifier’s prediction, specifically by highlighting the words that the classifier deems influential for the prediction. We presented 3 distinct cases: one where the classifier effectively leveraged relevant words for accurate prediction, another where it failed to do so, and a final case that illustrated the explanation for a misclassification. In addition, we analyzed the top 5 most prevalent words identified by LIME for each event type.

A total of 2 human factors graduate students were recruited to assess the quality of the LIME explanations. For each PSE report in the test data set, the reviewers were asked to determine independently if any of the highlighted words were relevant to the predicted event type. Based on these evaluations, the reports were then categorized into 2 distinct groups: those in which the highlighted terms were deemed relevant to the predicted event types and those where they were deemed irrelevant. Discrepancies were resolved through discussions. The interrater reliability index (Cohen $\kappa$) was calculated to quantify the level of agreement between the reviewers. The ML classifier’s accuracy and $F_1$-score were evaluated for these 2 groups of PSE reports. A subsequent comparison will explore the influence of explanation quality on prediction reliability.

**Interface Development**

In the typical workflow of PSE report classification, reporters need to provide a narrative description of the event as well as key attributes such as the event type, level of harm, date, and location of the event. Subsequent to this initial classification, the patient safety analyst will review the submitted report and decide if it needs to be recategorized to better reflect the nature of the event [17,42]. To support efficient and reliable categorization, the classifier will need to provide reporters with real-time support during the reporting process. We developed a PSE reporting interface to illustrate the integration of the ML classifier and the LIME explainability technique. In the design, the ML classifier provides multiple high-probability event types along with explanations for its prediction and allows the user to select the most appropriate event type. The interface was developed in Figma [43] and designed using guidance from previous research on incident reporting systems, including question type, mandatory and optional questions, and taxonomy for event type and harm level [44,45].

**Ethical Considerations**

The study was approved by the Medical University of South Carolina Hospital’s institutional review board (Pro00105892). Following data extraction, PSE reports were anonymized in accordance with privacy regulation guidelines.

**Results**

**Performance Comparison**

We evaluated the trained ML classifier’s classification performance on both static and contextual text representations (Multimedia Appendix 3). The performance of the top-performing ML classifier trained with static and contextual text representations is shown in Table 2. Our results showed that for static text representation, the MLR classifier trained with term frequency–inverse document frequency (TF-IDF) achieved the best performance, with an $F_1$-score of 0.631 and an accuracy of 66.7% (84/126). On the other hand, for contextual text representation, the SVM classifier trained with RoBERTa-base outperformed others, with an $F_1$-score of 0.753 and an accuracy of 75.4% (95/126). The SVM classifier trained with RoBERTa-base showed a 19.3% relative improvement in $F_1$-score and a 13% (11/85) relative improvement in accuracy compared to the ML classifier trained with TF-IDF for contextual text representation. In addition, we compared the accuracy (95/126, 75.4%) and top 2 accuracy (107/126, 84.9%) of the SVM classifier trained with RoBERTa-base and observed that 9.5% (12/126) of PSE reports’ true event type was predicted as the second highest probability event type by the classifier, which represents 39% (12/31) of misclassified PSE reports.
Table 2. Performance of top-performing ML classifiers trained with static and contextual text representations.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Top-performing ML model trained with the static text representation</th>
<th>Top-performing ML model trained with the contextual text representation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Performance</td>
<td>ML classifier</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>66.67</td>
<td>MLR(^b)</td>
</tr>
<tr>
<td>Top 2 accuracy (%)</td>
<td>85.71</td>
<td>MLR</td>
</tr>
<tr>
<td>Precision</td>
<td>0.707</td>
<td>KNN(^f)</td>
</tr>
<tr>
<td>Recall</td>
<td>0.667</td>
<td>MLR</td>
</tr>
<tr>
<td>(F_1)-score</td>
<td>0.631</td>
<td>MLR</td>
</tr>
</tbody>
</table>

\(^{a}\)ML: machine learning.  
\(^{b}\)MLR: multinomial logistic regression.  
\(^{c}\)TF-IDF: term frequency–inverse document frequency.  
\(^{d}\)SVM: support vector machine.  
\(^{e}\)MLP: multilayer perceptron.  
\(^{f}\)KNN: \(k\)-nearest neighbor.

Performance on Classifying Individual Event Types

We analyzed the performance of the SVM classifier trained with RoBERTa-base on individual event types (Table 3). The \(F_1\)-score measure for different event types ranged from 0.958 (laboratory test) to 0.400 (omission or errors in assessment, diagnosis, and monitoring).

Table 3. Performance of support vector machine+RoBERTa-base on the individual event type.

<table>
<thead>
<tr>
<th>Event type</th>
<th>Precision</th>
<th>Recall</th>
<th>(F_1)-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Care coordination or communication</td>
<td>0.721</td>
<td>0.838</td>
<td>0.775</td>
</tr>
<tr>
<td>Laboratory test</td>
<td>1.000</td>
<td>0.920</td>
<td>0.958</td>
</tr>
<tr>
<td>Medication related</td>
<td>0.765</td>
<td>0.722</td>
<td>0.743</td>
</tr>
<tr>
<td>Omission or errors in assessment,</td>
<td>0.417</td>
<td>0.385</td>
<td>0.400</td>
</tr>
<tr>
<td>diagnosis, and monitoring</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maternal</td>
<td>0.750</td>
<td>0.750</td>
<td>0.750</td>
</tr>
<tr>
<td>Equipment or devices</td>
<td>0.700</td>
<td>0.636</td>
<td>0.667</td>
</tr>
<tr>
<td>Supplies</td>
<td>0.778</td>
<td>0.700</td>
<td>0.737</td>
</tr>
</tbody>
</table>

Figure 1 shows the confusion matrix for the SVM classifier trained with RoBERTa-base evaluated on the test set. A confusion matrix is a table that visualizes the performance of a classifier. The main diagonal value is the number of PSE reports that have been classified as true event types, whereas off-diagonal values are the number of PSE reports that have been wrongly classified. While the classifier was able to classify the majority of event types of PSE reports correctly, there is a consistent misclassification of the omission or errors in assessment, diagnosis, or monitoring PSE report as the care coordination or communication (coordination) event type.
LIME-Based Explainability Analysis

We used LIME to evaluate whether the SVM classifier trained with RoBERTa-base has leveraged informative words for classification. Figure 2 presents 3 examples of explanations for the classifier’s predictions. At the top of Figure 2, LIME identified “ketorolac,” “ibuprofen,” and “doses” from the PSE report as important words for classifying the report into the medication-related event type, which is reasonable given the report’s association with incorrect medication doses. Conversely, in the middle of Figure 2, LIME highlighted “our,” “handle,” and “or” from the text as important words for classifying the report into the equipment or device event type. Although the predicted event type was correct, the classifier relied on irrelevant words for the classification. At the bottom of Figure 2, a case of misclassification is shown. LIME highlighted “pitocin,” “pump,” “available,” and “use” as influential words for classifying the PSE report into medication-related event type when it belongs to the equipment class. In addition, for each event type, we extract the 5 most prevalent words that were deemed important for the classifier’s prediction across the whole data set (Table 4). This inclusion of stop words (ie, “was,” “not,” and “till”) among influential terms, as shown in Table 4, demonstrated that the classifier does not always rely on relevant words for making classifications.
Figure 2. Local interpretable model-agnostic explanations of support vector machine classifiers trained with RoBERTa-base. MD: medical doctor; PSE: patient safety event; pt: patient.

Table 4. The 5 most prevalent and important words for each event type were derived from the support vector machine classifier trained with RoBERTa-base.

<table>
<thead>
<tr>
<th>Event type</th>
<th>Prevalent influential words highlighted by local interpretable model-agnostic explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Care coordination or communication</td>
<td>requested, delayed, patient, not, follow</td>
</tr>
<tr>
<td>Laboratory test</td>
<td>specimen, lab, labels, collection, results</td>
</tr>
<tr>
<td>Medication related</td>
<td>patches, doses, orders, medication, pitocin</td>
</tr>
<tr>
<td>Omission or errors in assessment, diagnosis, and monitoring</td>
<td>warning, patient, was, till, late</td>
</tr>
<tr>
<td>Maternal</td>
<td>baby, hysterectomy, stable, pumping, hemorrhage</td>
</tr>
<tr>
<td>Equipment or devices</td>
<td>instruments, trays, notified, malfunctioning, faulty</td>
</tr>
<tr>
<td>Supplies</td>
<td>vendor, sterile, available, needed, OR</td>
</tr>
</tbody>
</table>

After reviewing the LIME explanations for each PSE report in the test data set, 73.8% (93/126) of the reports were categorized into a subset where at least 1 highlighted word was deemed relevant to the predicted event type. The remaining reports comprised a second subset where no highlighted words were relevant. The interrater reliability index measured by Cohen \( \kappa \) between the 2 reviewers was 0.83, indicating substantial agreement. Table 5 presents the performance of the top-performing ML classifier for both subsets. For the first subset, the classifier achieved an accuracy of 84% (78/93) and an \( F_1 \)-score of 0.825. In contrast, the second subset showed a classifier accuracy of 52% (17/33) and an \( F_1 \)-score of 0.549.

Table 5. Performance of a top-performing machine learning classifier on reports that have relevant words highlighted and reports with irrelevant words highlighted.

<table>
<thead>
<tr>
<th>Metric</th>
<th>PSE reports with relevant words highlighted</th>
<th>PSE reports with irrelevant words highlighted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of PSE reports, n</td>
<td>93</td>
<td>33</td>
</tr>
<tr>
<td>Percentage of test data set (%)</td>
<td>73.81</td>
<td>26.19</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>83.87</td>
<td>51.51</td>
</tr>
<tr>
<td>( F_1 )-score</td>
<td>0.825</td>
<td>0.549</td>
</tr>
</tbody>
</table>

\( ^a \) PSE: patient safety event.

PSE Reporting System Interface

We designed an event reporting interface that integrates both the ML classifier and the LIME explainability technique. Figure 3 shows the event classification screen, where reporters enter a narrative description of the event after providing the details of the event, including date, time, unit, and information about the patient and reporter. Before describing the event in narrative form, reporters also choose among factors that contributed to the incident and the level of harm experienced by the patient. Once the reporter enters their narrative and selects the “classify” button, the system activates the ML classifier. Subsequently,
the interface displays the top 2 most probable event types, along with their associated probability distributions, in the lower left section. Simultaneously, the LIME technique will identify influential words that significantly contributed to the predicted event type, highlighting these words in green in the upper section of the dashboard. Based on the predicted event types and words highlighted for their influence on the prediction, the reporter may select the most suitable event type from a drop-down menu located in the lower-right section of the dashboard. Following this selection, reporters are queried on whether they agree with the classifier’s prediction, and the collected data can be used to guide subsequent refinement of the ML classifier.

**Figure 3.** Interface visualization of a patient safety event report classifier coupled with the local interpretable model-agnostic explanations technique. MD: medical doctor.

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicated-related</td>
<td>0.97</td>
</tr>
<tr>
<td>Care Coordination</td>
<td>0.03</td>
</tr>
</tbody>
</table>

* Words highlighted are influential for recommended event types.

**Discussion**

**Overview**

PSE event reporting systems are commonly used in health systems and hospitals across the world [46]. Data collected in PSE reporting systems drive quality improvement and patient safety efforts and supports regulatory reporting requirements for hospitals. The erroneous classification of PSE reports can impede the learning capabilities of the PSE reporting system, leading to suboptimal performance in detecting and preventing potential patient safety hazards [20]. It can also result in a substantial time cost for reclassifying PSE reports and compromise the integrity of a PSE database when analysts are investigating trends in events to develop effective solutions [17]. Previous studies have trained ML classifiers with static text representations for automatic PSE classification [12,17,25,26]. This study aimed to investigate whether using contextual text representations can further improve the accuracy of classifying PSE reports. We trained and evaluated a range of ML classifiers using both static and contextual text representations. To the best of our knowledge, this is the first time that contextual text representation has been used for training ML classifiers for PSE report classification. We analyzed the confusion matrix of the top-performing classifier to identify prevalent misclassified event types. Furthermore, aiming for more accurate and reliable PSE report classification, we incorporated an explainability technique to support human-AI collaboration and designed an interface to illustrate the possible integration of the ML classifier in PSE reporting systems.

**Principal Findings**

In this study, we extensively investigated the potential of using contextual representation for improving PSE report classification. The leading classifier trained with the static text representation (MLR trained with TF-IDF) was able to achieve an accuracy of 66.7% (84/126). This accuracy considerably exceeds the baseline accuracy of 29.4% (37/126), which involves classifying all PSE reports into the majority event type. However, using contextual text representation proved more efficacious. The SVM trained with contextual text representation (RoBERTa-base) was able to achieve an accuracy of 75.4% (95/126), reflecting a 13% (11/84) relative improvement in accuracy compared to the best-performing classifier trained with static text representation. While the achieved accuracy of 75.4% may not appear outstanding in isolation, it represents a significant advance compared with static text representations and exceeds the baseline, given the limited size of the data set. The improvement in classifier performance can be attributed to the use of contextual text representations, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context. Therefore, contextual text representation overcomes some limitations of static text representation, which can capture not only the meaning of individual words but also the complex and subtle ways in which words interact with each other in a specific context.
text representations. Hence, when training ML classifiers for PSE reporting systems, contextual text representation should be prioritized over static text representation to ensure the highest level of accuracy in classifying PSE reports.

As part of our investigation, we evaluated the performance of the top-performing classifier trained with contextual text representation on individual event types. While the classifier demonstrated impressive performance in accurately classifying laboratory test PSE reports ($F_1\text{-score}=0.958$), it struggled with classifying omissions or errors in assessment, diagnosis, and monitoring PSE reports, resulting in an unsatisfactory $F_1\text{-score}$ of 0.400. To investigate this discrepancy, we analyzed the confusion matrix for the classifier and discovered that omissions or errors in assessment, diagnosis, and monitoring PSE reports were frequently misclassified as the coordination event type. This misclassification can be attributed to the multiclass nature of PSE reports. For example, a failure to document the removal of a patient’s epidural catheter (omission or errors in assessment, diagnosis, and monitoring) could lead to a medication ordered by a physician (such as Lovenox) being withheld by the pharmacy due to a complication risk (coordination). On the other hand, the laboratory test is a more distinct event type in comparison to the other event types, and the classifier was able to correctly classify the majority of these reports. The observation obtained from the confusion matrix implies that PSE reports can potentially have more than 1 event type. This finding is consistent with previous studies [25,26]. The finding also underscores the need for further refinement in the development of the PSE taxonomy to create more distinctive event types. Another potential solution for addressing the multiclass nature of PSEs is to enable multiple event-type assignments [47]. Alternatively, the ML classifier can provide several probable event types, allowing the user to select the most appropriate one. We evaluated the top 2 accuracy of the top-performing ML classifier trained with contextual text representation and observed that 39% (12/31) of misclassified PSE reports’ true event type was predicted as the second-highest probability event type by the classifier. The finding suggests that there is a greater chance for the ML classifier to provide the correct event type when considering multiple options. As event reporting systems usually encompass over 20 event types, which can be difficult to memorize or access [17], narrowing down the PSE report’s potential event types to a smaller range also reduces the cognitive workload for PSE reporters during the classification process [48] and enhances the efficiency of reclassifying PSE reports for patient safety analysts.

We used LIME to showcase 3 predictions’ explanations and demonstrated cases where the ML classifier used informative words for classifying the PSE report and where it used irrelevant words for classification. These results highlight the importance of not solely relying on the ML classifier’s prediction and underscore the need for explainability and transparency in using the ML classifier for PSE report classification. Additionally, we showed the top 5 most prevalent words the ML classifier deemed important in the PSE reports for each event type. These words are indicative of the prevalent themes and issues within specific event types. Understanding the context and relationships between these prevalent informative words and specific event types can potentially provide valuable insights into the factors contributing to different types of PSEs. Furthermore, we have evaluated the top-performing ML classifier’s performance on 2 subsets of PSE reports, differentiated by whether the highlighted word by LIME is relevant to the predicted event type. Our findings reveal that the majority of PSE reports (93/126) have at least 1 relevant word highlighted, with the classifier achieving an accuracy of 84% (78/93) on these reports. Conversely, accuracy drops to 52% (17/33) when irrelevant words are highlighted. Such a disparity in performance emphasizes the necessity for additional scrutiny from reporters and patient safety analysts, particularly when dealing with PSE reports that have irrelevant words highlighted.

While previous research has focused on the development of ML classifiers, none of these previous works have investigated the potential integration of the classifier within the PSE reporting system in a manner that aligns with the workflow of the front-end reporter. We designed an interface to demonstrate the feasibility of a collaborative human-AI approach for event categorization. The interface provides the PSE reporter with multiple probable event types and associated explanations for the ML classifier’s prediction. This approach aligns with the principles of level 2 automation, where ML classifiers aid human decision-making rather than fully automating it [49]. This collaboration optimally combines human expertise with ML capabilities, potentially reducing cognitive workload and memorization of the taxonomy while also reducing the risks associated with overreliance on automation. Numerous studies have shown that the human-AI collaboration approach can improve the decision-making process [50-52], indicating its potential for enhancing PSE report classification. Furthermore, the interface also integrates the LIME explainability technique, which offers real-time insights into the rationale for the probable event types. Given the role of reporters and patient safety analysts in the incident reporting process, the use of explainability techniques can also increase trust in the recommendation provided by the ML classifier as it provides transparent and interpretable reasoning for the classification decisions [50,51]. Using LIME to highlight top informative words in real time for a PSE report can assist PSE reporters by emphasizing keywords in their narratives that are linked to the proposed classification. Highlighting informative words can also facilitate patient safety analysts working at the back end by providing insights into why a specific event type was chosen for classification. Such transparency not only clarifies current recommendations but also guides analysts in identifying influential terms for future report classifications. Previous research has illustrated the value of automation transparency in supporting appropriate levels of trust in the system, including decision support systems [32]. Additionally, regularly checking the explanations of the ML classifier’s prediction enables continuous monitoring of the classifier’s performance, identification of issues, and refinement [52]. As we have only designed the interface, additional research is needed to test the effectiveness of this approach in PSE report classification. Assessing the interface’s cognitive workload and decision-making accuracy is essential for ensuring its usability and adoption in the event reporting system. We plan to undertake
a usability testing study with health care professionals in a subsequent study.

**Comparison With Previous Work**

Research into the use of ML classifiers for the automation of PSE report classification has been relatively scarce. Wang et al [26] used logistic regression and SVM with the binary count, term frequency, and TF-IDF text representation to classify ten types of PSE reports, reaching an $F_1$-score as high as 0.783. However, they used a considerably larger data set (n=2860). Fong et al [17] achieved an accuracy rate of 92.0% (284/309) when they examined the usage of an ML classifier for classifying miscellaneous PSE reports using SVM, RF, and logistic regression with TF-IDF [17]. They also used a much larger data set (n=70,051). Ong et al [12] investigated the feasibility of using an ML classifier to automatically classify 2 types of PSE reports, including inadequate clinical handover and incorrect patient identification. They used Bag of Words model for text representation and trained both SVM and naive Bayes on classifying PSE reports, reaching accuracy as high as 98% (364/372). However, they framed the problem as a binary classification problem, which inherently has a higher baseline accuracy compared to our investigation. In this study, we’ve performed an in-depth comparative analysis with the available PSE data set and compared the established methods of classifying PSE reports and our novel method of using contextual text representations for classification. Our findings reveal that our proposed method outperforms the traditional models in terms of accuracy (ie, 84/126, 66.7% vs 95/126, 75.4%) and $F_1$-score (ie, 0.631 vs 0.753). This underlines the significance of our approach and its potential to advance the field of using ML classifiers for PSE report classification.

**Limitations**

There are several limitations to this study. First, the PSE reports used to train the ML classifiers were obtained from the maternal care units of a single hospital in the United States; therefore, the classifier might not generalize well to other settings. Second, this research’s scope was constrained by the limited amount of PSE report data, and only 7 prevalent classes were incorporated for training the ML classifiers. The restricted quantity of PSE reports might also result in an underestimation of the ML classifier’s actual capabilities [12]. Third, the quality of the LIME explanations was assessed by 2 graduate students; thus, further investigation is needed for a more robust validation of explanation quality. Furthermore, we have not yet empirically tested the interface for potential decision-making biases it may introduce.

Future research should investigate the performance of ML classifiers trained with contextual text representations on a larger and more diverse data set. Additionally, while we plan to refine the interface and test whether it supports event classification, future research can continue to investigate the appropriate way of incorporating the ML classifier into the reporting and reviewing workflow of PSE report classification and examine various human-AI collaboration approaches. Future studies should explore the potential biases (ie, automation bias) that the interface may introduce into the analysts’ decision-making process.

**Conclusions**

Improving the precision of PSE report classifications is a multifaceted task, involving both the refinement of the event type taxonomy and adequate training of hospital staff on the event reporting system. Despite these challenges, ML classifiers offer substantial potential to support accurate classification throughout the reporting and reviewing process. The findings of this study contribute to the advancement of ML classifiers for PSE report classification by demonstrating the superior performance of contextual text representation over static text representations in achieving more accurate classification outcomes. The integration of explainability techniques in ML classifiers fosters trust in their usage and provides valuable insights for informed decision-making and potential adjustments to the classifier. An event reporting interface that integrates an ML classifier with collaborative decision-making capabilities offers the potential to achieve an efficient and reliable PSE report classification process. These approaches can ultimately help hospitals identify risks and hazards promptly and take timely and informed actions to mitigate adverse events and reduce patient harm.

**Acknowledgments**

This research was funded by the Agency for Healthcare Research and Quality (1R03HS027680).

**Authors’ Contributions**

HC was responsible for the conceptualization, data analysis, and drafting of the manuscript. EC contributed to the conceptualization, methodology design, and review and revision of the document. DW contributed to data acquisition and funding acquisition. MA contributed to data acquisition, conceptualization, funding acquisition, and the review and revision of the manuscript.

**Conflicts of Interest**

None declared.

Multimedia Appendix 1
Data preprocessing procedures.
[DOCX File. 14 KB - humanfactors_v11i1e53378_app1.docx ]
References


4. Makary MA, Daniel M. Medical error—the third leading cause of death in the US. BMJ 2016;353:i2139 [FREE Full text] [doi: 10.1136/bmj.i2139] [Medline: 27143499]


43. The collaborative interface design tool. Figma. URL: https://www.figma.com/ [accessed 2023-09-24]


https://humanfactors.jmir.org/2024/1/e53378

JMIR Hum Factors 2024 | vol. 11 | e53378 | p.372

(page number not for citation purposes)


50. Lai Y, Kankanhalli A, Ong DC. Human-AI collaboration in healthcare: a review and research agenda. 2021 Presented at: 54th Hawaii International Conference on System Sciences (HICSS); January 5-8, 2021; Grand Wailea, Maui, Hawaii URL: https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/495a50c0-16e9-4c3e-a00b-34aee9df6b/content [doi: 10.24251/hicss.2021.046]


Abbreviations

| AI: artificial intelligence |
| KNN: k-nearest neighbor |
| LIME: local interpretable model-agnostic explanations |
| ML: machine learning |
| MLR: multinomial logistic regression |
| OECD: Organization for Economic Cooperation and Development |
| PSE: patient safety event |
| RF: random forest |
| SVM: support vector machine |
| TF-IDF: term frequency–inverse document frequency |

Edited by E Borycki, A Kashyuruk; submitted 06.10.23; peer-reviewed by H Tibble, S Matathil; comments to author 19.11.23; revised version received 30.11.23; accepted 03.12.23; published 25.01.24.

Please cite as:
Chen H, Cohen E, Wilson D, Alfred M
A Machine Learning Approach with Human-AI Collaboration for Automated Classification of Patient Safety Event Reports: Algorithm Development and Validation Study
JMIR Hum Factors 2024;11:e53378
URL: https://humanfactors.jmir.org/2024/1/e53378
doi: 10.2196/53378
PMID: 38271086

©Hongbo Chen, Eldan Cohen, Dulaney Wilson, Myrtede Alfred. Originally published in JMIR Human Factors (https://humanfactors.jmir.org), 25.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Human Factors, is properly cited. The complete bibliographic information, a link to the original publication on https://humanfactors.jmir.org, as well as this copyright and license information must be included.
Testing the Feasibility and Acceptability of Using an Artificial Intelligence Chatbot to Promote HIV Testing and Pre-Exposure Prophylaxis in Malaysia: Mixed Methods Study

Min Hui Cheah¹, MClinPharm; Yan Nee Gan², MSc; Frederick L Altice³,4,5,6, MA, MD; Jeffrey A Wickersham³,5,6, PhD; Roman Shrestha⁵,6,7, MPH, PhD; Nur Afiqah Mohd Salleh², PhD; Kee Seong Ng¹, MD, PhD; Iskandar Azwa¹,6, MD; Vimala Balakrishnan⁸, PhD; Adeeba Kamarulzaman⁹, MBBS; Zhao Ni⁵,10, PhD, BMed

¹Department of Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia
²Department of Social and Preventive Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia
³Section of Infectious Disease, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, United States
⁴Division of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT, United States
⁵Center for Interdisciplinary Research on AIDS (CIRA), Yale University, New Haven, CT, United States
⁶Centre of Excellence for Research in AIDS (CERiA), Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia
⁷Department of Allied Health Sciences, University of Connecticut, Storrs, CT, United States
⁸Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
⁹Monash University Malaysia, Kuala Lumpur, Malaysia
¹⁰School of Nursing, Yale University, Orange, CT, United States

Corresponding Author:
Zhao Ni, PhD, BMed
School of Nursing
Yale University
400 West Campus Drive
Orange, CT, 06477
United States
Phone: 1 203 737 3039
Email: zhao.ni@yale.edu

Abstract

Background: The HIV epidemic continues to grow fastest among men who have sex with men (MSM) in Malaysia in the presence of stigma and discrimination. Engaging MSM on the internet using chatbots supported through artificial intelligence (AI) can potentially help HIV prevention efforts. We previously identified the benefits, limitations, and preferred features of HIV prevention AI chatbots and developed an AI chatbot prototype that is now tested for feasibility and acceptability.

Objective: This study aims to test the feasibility and acceptability of an AI chatbot in promoting the uptake of HIV testing and pre-exposure prophylaxis (PrEP) in MSM.

Methods: We conducted beta testing with 14 MSM from February to April 2022 using Zoom (Zoom Video Communications, Inc). Beta testing involved 3 steps: a 45-minute human-chatbot interaction using the think-aloud method, a 35-minute semistructured interview, and a 10-minute web-based survey. The first 2 steps were recorded, transcribed verbatim, and analyzed using the Unified Theory of Acceptance and Use of Technology. Emerging themes from the qualitative data were mapped on the 4 domains of the Unified Theory of Acceptance and Use of Technology: performance expectancy, effort expectancy, facilitating conditions, and social influence.

Results: Most participants (13/14, 93%) perceived the chatbot to be useful because it provided comprehensive information on HIV testing and PrEP (performance expectancy). All participants indicated that the chatbot was easy to use because of its simple, straightforward design and quick, friendly responses (effort expectancy). Moreover, 93% (13/14) of the participants rated the overall chatbot quality as high, and all participants perceived the chatbot as a helpful tool and would refer it to others. Approximately 79% (11/14) of the participants agreed they would continue using the chatbot. They suggested adding a local language (ie, Bahasa Malaysia) to customize the chatbot to the Malaysian context (facilitating condition) and suggested that the chatbot should also incorporate more information on mental health, HIV risk assessment, and consequences of HIV. In terms of social influence, all
participants perceived the chatbot as helpful in avoiding stigma-inducing interactions and thus could increase the frequency of HIV testing and PrEP uptake among MSM.

Conclusions: The current AI chatbot is feasible and acceptable to promote the uptake of HIV testing and PrEP. To ensure the successful implementation and dissemination of AI chatbots in Malaysia, they should be customized to communicate in Bahasa Malaysia and upgraded to provide other HIV-related information to improve usability, such as mental health support, risk assessment for sexually transmitted infections, AIDS treatment, and the consequences of contracting HIV.

(KEYWORDS: artificial intelligence; acceptability; chatbot; feasibility; HIV prevention; HIV testing; men who have sex with men; MSM; mobile health; mHealth; preexposure prophylaxis; PrEP; mobile phone)

Introduction

Background

HIV continues to be a global health concern causing approximately 630,000 deaths yearly worldwide [1]. In 2019, approximately 62% of the new HIV infections among adults worldwide occurred within key populations and their sexual partners [2]. Men who have sex with men (MSM) accounted for 23% of new infections of HIV, which was much higher than the percentage of new infections in other key populations, such as people who used drugs (10%), sex workers (8%), and transgender people (2%) in 2019 [3]. Malaysia is a Southeast Asian country with a population of 33.5 million, with 1 in 5 MSM living with HIV [4]. Over the past 2 decades, the mode of HIV transmission in Malaysia has shifted from needle sharing and i.v. drug use to sexual transmission, particularly among MSM [5].

HIV testing is a prerequisite to effective HIV prevention and early treatment initiation [6]; people at risk for HIV or seropositive individuals need to be tested for HIV before being linked to health care services [7]. Despite the importance of HIV testing, it is disproportionately lower among MSM in Malaysia [8]. New HIV testing guidelines recommend that MSM at high risk for HIV should be tested every 3 to 6 months, but most MSM in Malaysia do not test optimally. Studies in Malaysia have found that only 9.5% of MSM tested more than once a year. In Malaysia, engaging in same-sex sexual behavior is prohibited by both secular and Sharia laws, leading to significant levels of stigma and discrimination within society [9]. As a result, many MSM may be hesitant or unwilling to engage with health care providers and outreach workers. Therefore, designing new strategies to promote HIV testing among MSM in Malaysia is urgently needed [10].

Using portable electronic devices with software programs to deliver health care services and manage patient information is known as mobile health (mHealth) [11]. mHealth interventions could reduce barriers to HIV testing for MSM by reducing in-person contact and offering internet-based platforms for HIV testing [12,13]. Studies have demonstrated that mHealth interventions using smartphones and apps could increase the uptake of HIV testing while protecting the privacy of MSM [14-16], and MSM in Malaysia have a high acceptance of the use of mHealth for HIV testing and prevention [13,17,18]. Recent breakthroughs in artificial intelligence (AI) and machine learning can potentially automate and scale up these mHealth interventions through chatbots, a computer program that can mimic human conversation [19]. However, leveraging chatbot technology to promote HIV testing and prevention is in its infancy [15,20]. Although chatbot technology holds immense potential to prevent HIV, a lack of research in this field undermines its significance. The creation of ChatGPT has brought attention to the significance of studying chatbot technology for health care.

Our team has conducted formative research to understand HIV prevention chatbots in Malaysia and has identified the perceived benefits, limitations, and preferred features of AI chatbots for HIV testing and prevention among MSM [13]. On the basis of the study findings, we developed an HIV prevention AI chatbot prototype named Haris (a common Malaysian name) and a website called MYHIV365 (MY symbolizes Malaysia, HIV implies health care services aimed at preventing HIV, and 365 indicates the services are available every day of the year). Haris was hosted on MYHIV365 and could provide information on the 3 themes most needed by MSM: HIV testing, mental health, and pre-exposure prophylaxis (PrEP). PrEP is a highly effective HIV prevention method that involves the use of antiretroviral medication by at-risk individuals to prevent getting HIV from sex or injection drug use. Haris imitates human intelligence and can interact with users to provide support, including ordering free HIV self-testing kits, screening for depression, and recommending MSM-friendly clinics where individuals can get tested for HIV and receive PrEP.

Objectives

Despite the meticulous design and alpha testing (internal testing) of Haris among professors, experts, and community advisory board members, its feasibility and acceptability in preventing HIV among MSM is still unknown. Therefore, we conducted beta testing (testing in a real-world environment by actual users) of Haris among 14 MSM in Malaysia to address this knowledge gap. Specifically, we examined the use of the AI chatbot for delivering health information and improving linkage to HIV testing, PrEP, and care. We also investigated key strategies to refine the feasibility and acceptability of the AI chatbot in this study.

(JMIR Hum Factors 2024;11:e52055) doi:10.2196/52055

KEYWORDS: artificial intelligence; acceptability; chatbot; feasibility; HIV prevention; HIV testing; men who have sex with men; MSM; mobile health; mHealth; preexposure prophylaxis; PrEP; mobile phone

https://humanfactors.jmir.org/2024/1/e52055
Methods

Study Design and Participants
Beta testing of the AI chatbot prototype was conducted with 14 MSM by an experienced qualitative interviewer (ZN) with expertise in chatbot development and HIV prevention in Malaysia and 4 trainees in the Malaysian Implementation Science Training program (Fogarty International Center, D43TW011324). Participants were recruited in Malaysia from February to April 2022 via social networking apps commonly used by MSM, including Grindr, Hornet, Blued, and WhatsApp. The procedures for participant recruitment have been published elsewhere [13]. A web-based screener including questions on demographic characteristics and HIV prevention practices was used. The eligibility criteria included (1) self-identification as a cisgender man, (2) age ≥18 years, (3) condomless sex with another man in the past 6 months, and (4) being HIV negative or of unknown status.

Each beta test involved the following three steps: (1) a 45-minute human-chatbot interaction using the think-aloud method [21]; (2) a 35-minute semistructured interview; and (3) a 10-minute web-based survey. The first 2 steps were conducted via Zoom (Zoom Video Communications, Inc), recorded, and transcribed verbatim. Specifically, 2 days before the test, a research assistant sent a calendar invite with Zoom meeting information to the interviewer and participant. One day before beta testing, the research assistant emailed the participant a detailed description of the human-chatbot interaction (Multimedia Appendix 1). During the human-chatbot interaction, participants were asked to share their screen via Zoom and access the chatbot through a URL sent by the research assistant. After the participants obtained access to the chatbot, the research assistant randomly selected 3 to 5 tasks from the list of beta testing tasks (Multimedia Appendix 2) and asked the participants to complete them through the chatbot. Some examples of the tasks include “find a clinic that can provide HIV testing service in Kuala Lumpur” and “find out the common symptoms of depression through the chatbot.” The study procedure is described in Figure 1.
After the human-chatbot interaction, we conducted a semistructured interview (Multimedia Appendix 3 [22]) soliciting participants’ feedback on two themes: (1) experience navigating the chatbot and (2) how the chatbot should be made available to a wider audience. During the interview, participants were asked several questions regarding their experience with the chatbot, such as “How was your experience with the AI chatbot?”, “What feature of the AI chatbot do you like the most?”, and “What information needs to be added to the AI chatbot to increase its popularity among MSM?” After the interviews concluded, the participants were provided with a survey link to assess the feasibility and acceptability of the AI chatbot. The feasibility of the chatbot was measured through 4 outcomes, including participants’ ratings of the chatbot’s quality, satisfaction, intention to continue using the chatbot, and willingness to refer it to others. The outcomes were measured using a 10-point rating scale, with higher scores indicating more favorable outcomes (Multimedia Appendix 4). For example,
participants’ satisfaction with the chatbot was measured by using the question, “How satisfied were you with the experience of interacting with the chatbot?” The score of “0” stands for not satisfactory at all and “10” stands for extremely satisfactory. The acceptability of the chatbot was measured using the standardized System Usability Scale [23] and an adjusted Chatbot Usability Scale [24]. The combination of the 2 scales provided a comprehensive evaluation of the acceptability of our chatbot.

**Ethical Considerations**

The participants provided electronic consent before initiating the beta testing. This study was approved by the institutional review board of Yale University (approval #2000027864) and Medical Research Ethics Committee of the University of Malaya (approval #2021112-10729). This research was conducted in accordance with the ethical standards of the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

**Conceptual Framework for Analysis**

The Unified Theory of Acceptance and Use of Technology (UTAUT) was used as a conceptual framework to guide the analysis of the experience of MSM using the AI chatbot for HIV testing and prevention in Malaysia. UTAUT consists of four domains: (1) performance expectancy, (2) effort expectancy, (3) facilitating conditions, and (4) social influence [25]. The definitions of these 4 domains have been published elsewhere [13]. UTAUT was chosen for the following reasons. First, this AI chatbot was developed based on the findings from a formative research project that was analyzed using UTAUT [13]. Therefore, using the same theory, we can compare the results of the 2 studies on the 4 domains and are more likely to find out the feasibility and acceptability of the AI chatbot. Second, UTAUT emphasizes user-centered perspective, which allows researchers to assess the acceptance of the AI chatbot from the users’ perception. Third, UTAUT has been extensively used to identify users’ acceptance of technology and was reported to be effective and of high validity [26,27].

**Analyses**

All transcripts were cross-checked for accuracy and completeness by 7 researchers (MHC, YNG, NAMS, KSN, ZN, and 2 research assistants). Each of the 7 researchers independently coded 2 transcripts using NVivo 10 software (QSR International), compiled codes, and mapped the emerging themes from the qualitative data on the 4 domains of UTAUT, including performance expectancy, effort expectancy, facilitating conditions, and social influence. Discrepancies in codes and themes were addressed in group discussions where there was discordance in coding. We ceased the qualitative analysis when the results reached saturation, and no new themes emerged. The participants’ quotes are presented throughout the results with additional quotes given in Multimedia Appendix 5. Quantitative data from the survey were analyzed using SAS (version 9.4; SAS Institute) and are presented as descriptive statistics.

**Results**

**Participant Characteristics**

The 14 participants were on average in their mid-20s (mean 25.6, SD 4.2 years), and most of them (13/14, 93%) used smartphones as the primary means to access the internet. Most participants (10/14, 71%) were Malay, followed by Chinese (3/14, 21%) and Indian (1/14, 7%). About one-third of the participants (5/14, 36%) had taken PrEP previously, and only 14% (2/14) of them were currently taking PrEP. The demographic characteristics are summarized in Table 1.
Table 1. Participant demographic details (N=14).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (y), mean (SD)</td>
<td>25.6 (4.2)</td>
</tr>
<tr>
<td>Ethnicity, n (%)</td>
<td></td>
</tr>
<tr>
<td>Malay</td>
<td>10 (71)</td>
</tr>
<tr>
<td>Chinese</td>
<td>3 (21)</td>
</tr>
<tr>
<td>Indian</td>
<td>1 (7)</td>
</tr>
<tr>
<td>Sexual orientation, n (%)</td>
<td></td>
</tr>
<tr>
<td>Bisexual</td>
<td>3 (21)</td>
</tr>
<tr>
<td>Gay</td>
<td>11 (79)</td>
</tr>
<tr>
<td>Employment status, n (%)</td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>6 (43)</td>
</tr>
<tr>
<td>Working full time</td>
<td>8 (57)</td>
</tr>
<tr>
<td>Highest level of education, n (%)</td>
<td></td>
</tr>
<tr>
<td>Diploma or bachelor degree</td>
<td>8 (57)</td>
</tr>
<tr>
<td>Master degree or PhD</td>
<td>3 (21)</td>
</tr>
<tr>
<td>Secondary school</td>
<td>3 (21)</td>
</tr>
<tr>
<td>Average monthly income (MYR;&lt; 1 MYR=US $0.21), n (%)</td>
<td></td>
</tr>
<tr>
<td>&lt;2000</td>
<td>6 (43)</td>
</tr>
<tr>
<td>2000-4000</td>
<td>5 (36)</td>
</tr>
<tr>
<td>&gt;4000</td>
<td>3 (21)</td>
</tr>
<tr>
<td>Daily access to the internet, n (%)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>14 (100)</td>
</tr>
<tr>
<td>Primary device for accessing the internet, n (%)</td>
<td></td>
</tr>
<tr>
<td>Smartphone</td>
<td>13 (93)</td>
</tr>
<tr>
<td>Laptop computer</td>
<td>1 (7)</td>
</tr>
<tr>
<td>Had ever taken PrEP,b, n (%)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>5 (36)</td>
</tr>
<tr>
<td>No</td>
<td>9 (64)</td>
</tr>
<tr>
<td>Currently taking PrEP, n (%)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>2 (14)</td>
</tr>
<tr>
<td>No</td>
<td>12 (86)</td>
</tr>
</tbody>
</table>

aMYR: Malaysian Ringgit.  

Feasibility

The mean scores on the 4 metrics of the feasibility of the chatbot, overall quality, satisfaction, intention to continue using, and willingness to refer to others were 7.86 (SD 1.03), 8.14 (SD 1.23), 8.64 (SD 1.65), and 8.93 (SD 1.07), respectively, on a scale from 0 to 10 (Figure 2).
Acceptability

The participants found the chatbot acceptable, as it was perceived as easy to navigate and capable of providing valuable information (Multimedia Appendix 6). Specifically, all participants (14/14, 100%) expressed confidence in using the chatbot, believing that others could also quickly master its use (Figure 3). The overall mean (SD) score of the System Usability Scale was 76.07 (SD 8.19), which is greater than the recommended acceptable cutoff score of 68 [23].

Moreover, 78% (11/14) of the participants agreed that the chatbot could understand their inputs accurately (Figure 4). However, only 36% (5/14) of the participants agreed that their interaction with the chatbot felt like a natural conversation.
In addition to measuring feasibility and acceptability, we summarized the study findings based on the 4 domains of UTAUT: performance expectancy, effort expectancy, facilitating conditions, and social influence.

**Performance Expectancy**

**Overall Perception**

Participants responded positively about the performance of the AI chatbot. Quantitative data analysis revealed that all participants (14/14, 100%) perceived the chatbot as a helpful tool, they would refer it to others, and 93% (13/14) of them highly rated the overall quality of the chatbot. These results were consistent with the qualitative finding that participants were satisfied with the information provided by the chatbot and deemed it a trustable source.

**Contributors to Positive Performance Expectancy**

The chatbot served as a reliable source of information. Participants expressed their trust in this chatbot as they believed it was developed to help people learn more about HIV. For instance, one participant stated that he trusted the chatbot more than other internet-based platforms as the following:

> I can actually trust, trust this chatbot more than I can trust the Internet. [Interview E]

Most participants expressed that the information provided by the chatbot was comprehensive and satisfactory. The same participant highlighted the following:

> ...everything is there, everything is informative... [Interview E]

Participants described the chatbot’s function of ordering free HIV self-test kits as one of the most useful functions. The simple and straightforward instructions from the chatbot significantly encouraged participants to perform HIV self-testing and helped prevent misuse of the self-test kit. A participant elaborated on the following:

> ...it never crossed my mind that you can do HIV self-test just using it (test stick) over the gums and without blood... [Interview G]

In addition to the positive feedback on the chatbot’s ability to order free HIV self-test kits, participants also expressed appreciation for the information provided by the chatbot on MSM-friendly local clinics where they could test for HIV and receive PrEP consultation. For example, the participant stated the following:

> That was beyond amazing...they give [me] the addresses, contact numbers. So, I would say, if a person really needs to do the [HIV] testing, essential information like that would be very useful, so I think it’s more than helpful. The options [of HIV testing clinics] are not just [limited to] one or two [clinics], you know, so that’s good. [Interview G]

The other participants stressed that the chatbot’s features related to HIV self-testing and venue-based HIV testing were complementary and appreciated having access to both options through the chatbot. Although performing self-testing at home may be convenient, it may lack the human interaction that some MSM need for support during the testing process. By offering both self-testing and venue-based testing options, the chatbot gave MSM the flexibility to choose the option that best suited their individual needs and preferences. For example, a participant emphasized that the chatbot could enable his MSM friends to conduct HIV self-testing and then see a health care professional for advice on complex and sensitivity issues:
Along with the positive feedback on the chatbot’s features relevant to HIV testing, the participants also expressed their favorable feelings toward the PrEP information provided by the chatbot. For example, a participant mentioned that the chatbot could send him introductory information to allow him to comfortably assess his risk level and help him decide if he needed to take PrEP. Another participant reiterated the first participant’s point by emphasizing the difficulties that MSM in determining if they should take PrEP, stating the following:

A lot of people [MSM] are always asking themselves, “Should I get PrEP? Am I at risk? Do I need to take PrEP as a precautionary measure?” ... So these are [questions] that MSM usually a bit too scared to ask the doctors. [Interview C]

Major Concerns

Participants suggested that the chatbot would be improved if it could provide more information and resources relevant to mental health, as mental health issues were the prominent problems that MSM faced in Malaysia. Participants wanted the chatbot to provide information on strategies for managing stress, statistics about depression among MSM, peer consultation for depression, and professional health care services to prevent and treat depression. The participants also highlighted that the MYHIV365 website, where the chatbot was embedded, should provide more resources related to mental health. For example, one participant described this problem as follows:

The website did not have links to any information regarding mental health issues, and that is a glaring issue for it to be left out like that. [Interview A]

The same statement was echoed by another participant, who stressed the following:

I just find that for the mental health, it’s kind of short. [Interview G]

Relevant Features Suggested by Participants That Are Needed to Improve the Chatbot

Although the AI chatbot was developed primarily for HIV prevention and to assist with HIV testing and access to PrEP, the participants pointed out that some participants may test positive for HIV and would benefit from learning more about accessing HIV care and related antiretroviral therapy services. In addition, participants suggested that the chatbot should provide more information about antiretroviral therapy so that users could better manage HIV by knowing potential drug interactions and side effects. The participants also recommended providing more information about high-risk behaviors and sexually transmitted diseases to help increase awareness about HIV and sexually transmitted diseases among MSM. One participant stated the following:

I think [providing more information about] HIV treatment would [be] very helpful because those who might be exposed to HIV would definitely want to know what the treatment is all about. [Interview E]

This participant’s statement was echoed by another participant, who stated the following:

...HIV and STDs...[are] not the same, but...I thought [they were] the same...I thought HIV and STDs were not curable...so I think it will be great if you can add STDs [to the chatbot]. [Interview E]

Effort Expectancy

Overall Perception

All surveyed participants (14/14, 100%) agreed that the chatbot was easy to use, and 86% (12/14) of the participants were satisfied with the chatbot. In the qualitative interviews, participants reported consistent feedback that the chatbot was user friendly and convenient to use, and they were satisfied with the chatbot because of its simple, straightforward design and quick, friendly responses. However, they were concerned about the technical issues, including the address input and text alignments (refer to the Major Concerns section). The participants also felt that tailoring the chatbot to the local context and adding a “human touch” would be helpful.

Positive Contributors to Low Effort Expectancy

Many participants expressed satisfaction with the chatbot because of its prompt response, expert information, and plain interface. Two participants commented the following:

...white and blue colors [are] neutral, and it [the chatbot] takes into account [of] color blindness as well, so that’s great. [Interview C]

...[I] got a quick response [from the chatbot]. [Interview I]

The individualized and user-centered features of the chatbot, which cater to users with different levels of communication skills, were highlighted among the participants. For instance, one participant stated that the chatbot offered an ideal platform for MSM who are less comfortable interacting with others. A participant stated the following:

As we all know, some of us didn’t have the skills to communicate, so I think...[the] chatbot... will definitely help. I think it was great. [Interview E]
Moreover, participants thought the chatbot was useful as it facilitated them to obtain culturally tailored health information. The chatbot met users’ needs by providing a menu of options for users to choose from. Compared with obtaining health information in clinical settings in Malaysia, the chatbot was much simpler. A participant elaborated on the following:

When [the chatbot] come[s] up with three options like that, I can explore myself...I would say that [the chatbot] is more precise; it gets to the point directly. [Interview G]

The health intervention being tailored to the local setting was highly valued by the participants. Responses from the AI chatbot that contained localized features, specifically the use of “Manglish,” a less formal form of Malaysian English, were appreciated by several participants. The feature of “Manglish,” which was not in the standardized form of English, has added a local flavor to the AI chatbot, which some participants found amusing. A participant stated the following:

The impression that this chatbot...probably comes from America. It’s in English, so the moment it puts up a Malaysian style saying “Boleh”... I’m very amused with this [style]. [Interview G]

### Major Concerns

Some participants spoke about the difficulty in filling in their home addresses using the current prompts on the AI chatbot when they needed to order an HIV self-test kit; the chatbot required a step-by-step input of addresses, which was counterintuitive and inefficient. Participants preferred the standard address format in Malaysia over the step-by-step input format, in which incomplete addresses would triage further prompts to ask participants to refill the HIV self-test order. For example, one participant stated the following:

In Malaysia, we don’t use the term “line address” or “street address”. We usually enter the full address with the postcode and then the city and state. The one on the chatbot seems to be how addresses are filled in the United States. That part needs to be tweaked slightly based on Malaysian cities. [Interview C]

In addition, participants expressed that the address of the clinics provided by the chatbot needed to be tailored to Malaysian culture. For example, the district options may only be needed for certain states in Malaysia. A participant stated the following:

I think depends on the size of the state...we don’t have to call out (provide choices for) all the districts because Perlis is already small enough, and I think...people can just go easily from one place to another in Perlis. But if...it is a big state...we need to divide it using district. [Interview I]

### Relevant Features Suggested by Participants That Are Needed to Improve the Chatbot

Although participants were satisfied with the AI chatbot, 2 participants suggested that the chatbot’s interface could be improved by adding more spaces between sentences, and the alignment of sentences should be adjusted to make the chatbot look more professional. Two participants described the following:

...everything is tightly together with very little gap...there should be proper spacing... [Interview L]

The text is not properly centered in some of the boxes, and I feel like it could [be] a better design to make it look more professional. [Interview A]

The use of English as the only language of the AI chatbot was perceived by participants as a barrier to implementing the chatbot in Malaysia. Although all participants were proficient in English, concerns arose for the communities where English was not widely spoken. Participants suggested that the chatbot should be able to communicate in Bahasa Malaysia or Mandarin, given that the 2 languages are widely spoken in Malaysia. A participant stated the following:

...perhaps to have another option of language...I think that would be able to cover more people within the local population. [Interview C]

Adding a “human touch” to the chatbot can create a more engaging and user-friendly experience for the users interacting with the chatbot. The participant described the following:

...ideally, we would want [the chatbot’s response] to be as human as possible, and not so robotic in its responses...a nice touch to make someone feel slightly comfortable. [Interview C]

### Facilitating Conditions

#### Overall Perception

Participants reported 2 major facilitating conditions for the use of the AI chatbot. First, the social distancing policy adopted by the Malaysian government during the COVID-19 pandemic significantly increased the use of internet-based platforms to seek health information and consult about health issues among MSM. The participants expressed that the AI chatbot was a novel tool to promote HIV testing and prevention among MSM in Malaysia. A participant highlighted the convenience of using the chatbot as an alternative to meeting health care workers during the COVID-19 pandemic as follows:

...because now it’s COVID, everyone is doing it in IT (information technology) format. Having an AI chatbot is definitely much more convenient than meeting people... [Interview G]

Second, the AI chatbot’s capability of referring webpages to participants where they could find mental health information, community support, and counselors was a significant facilitator for them to accept the chatbot. Many participants stated that it was much easier to obtain information through the links provided by the chatbot than searching for information via websites or mobile apps. A participant stated the following:

When you interact with it (the chatbot), it throws out links to you. It’s easier to navigate to the particular links from there. [Interview G]
Relevant Features Suggested by Participants That Are Needed to Improve the Chatbot

Participants suggested that the chatbot could be promoted through social media platforms, such as Facebook, Twitter, Instagram, YouTube, TikTok, and Telegram because these platforms were widely used by MSM as sources of information. Among all social networking apps, participants stated that Twitter was the best platform to advertise the AI chatbot because Twitter enabled users to post clickable links in the comments section where other users could access the chatbot. Participants further reported that Telegram was a more suitable platform for hosting the chatbot than the most popular text messaging app in Malaysia, WhatsApp. Telegram offers a more private and secure environment for MSM to ask questions or express concerns about HIV and AIDS. Participants also suggested that building a trustable relationship between the AI chatbot and the MSM community is key to implementing the AI chatbot in Malaysia. Given that there were many scams through pop-up advertisements on social media platforms, a participant described the following imagined scenario:

...if we play our Facebook, Twitter, Instagram, or YouTube, there are always mini advertisements, so who knows, [whether we] can add this [AI chatbot]?...I need to know about this [chatbot], and I hope this [chatbot] is not a scam. [Interview B]

Social Influence

In terms of social influence, the chatbot was perceived as helpful in avoiding HIV stigma and thus could increase the HIV testing rate and PrEP uptake frequency. Quantitative data analysis found that 79% (11/14) of the participants agreed to continue using the chatbot. During the interviews, these participants reported that societal stigma and discrimination related to HIV and AIDS would make them more likely to use the chatbot. They expressed discomfort in asking people questions about HIV and AIDS as they were afraid of encountering stigma and negative attitudes from others. MSM often preferred to seek information through internet-based platforms, and the chatbot was helpful, particularly for people living in small social circles. A participant elaborated on the following:

...this topic [HIV] is quite sensitive to most people, it will create like a negative energy around you...personally I don’t go and ask people what HIV is. I will search myself maybe on the Internet... [Interview E]

The societal stigma and discrimination toward HIV and AIDS also facilitated participants to select HIV self-testing at home rather than testing in a clinical setting. Many participants appreciated that the chatbot offered them an opportunity to receive free HIV self-test kits while protecting their privacy. Two participants who used to be shy about discussing HIV described the following:

Because from the MSM community, some of us are not very comfortable of getting [HIV] test kits on site, because like...fear of the stigma, that the society will judge. [Interview D]

I can directly book the test kit through the chatbot, which is very useful and informative...my identity will remain anonymous, so people don’t know me. [Interview E]

Participants deemed the AI chatbot useful and expressed their willingness to recommend the AI chatbot to others. Some participants suggested that the chatbot should be promoted among MSM who frequently use social networking apps, such as Grindr, Hornet, and Blued, to find sexual partners because those MSM were at higher risk for HIV and had greater need for HIV information. A participant stated the following:

I have the impression that anyone would actually need it [HIV testing]. But if we look at it from another angle, people on hookup apps like Grindr have a high tendency to hook up using those apps compared to those who don’t use them...we need to introduce the chatbot to them because...they...[have] been highly exposed [to HIV]. [Interview G]

Discussion

Principal Findings

The feasibility and acceptability of leveraging AI chatbots to promote HIV testing and PrEP among MSM in Malaysia is high. Discrimination and stigma toward HIV and AIDS are major barriers for MSM to access high-quality HIV testing and prevention services in Malaysia, and they are also primary facilitators for MSM to seek health information via internet-based platforms. Our AI chatbot prototype provides a platform for MSM to order free HIV self-test kits in an MSM-friendly environment and to empower them with resources and instructions. MSM who prefer to interact with health care providers in person can also locate HIV testing clinics or PrEP clinics through the AI chatbot. MSM highlighted these functions of the AI chatbot as very useful.

Similar to other studies, AI chatbots were well received by users [28,29]. An AI chatbot could enhance engagement with the key population [30]. As contemporary social patterns increasingly involve the integration of AI into everyday routines, AI chatbots could contribute to delivering precise details regarding HIV testing to individuals actively seeking such information. A chatbot named Eli, developed by the United Nations Educational, Scientific, and Cultural Organization, received highly favorable user feedback and was widely acclaimed [29]. Eli offers a range of services, including details on HIV prevention, testing, and treatment and assistance in overcoming fears and concerns. Compared with Eli, our AI chatbot did not have information on treatment for AIDS and provided limited mental health support. Integrating these functions into our AI chatbot may support its usability. Nevertheless, our AI chatbot offers free HIV self-test kits and locates local clinics in Malaysia for HIV testing, PrEP consultation, and mental health care.

From our previous formative research, we know that factors facilitating the acceptance of an HIV prevention AI chatbot include providing useful information and having the capacity to solve problems [13]. In this study, participants reported that our AI chatbot was able to provide useful information and help...
solve problems. This was indicated by the results that all participants perceived this chatbot as a helpful tool, and most participants deemed the chatbot a reliable source of information with a high satisfaction score. However, one area that required significant improvement in the chatbot was its conversation flow, as only 36% (5/14) of the participants felt that their interaction with the chatbot resembled a natural conversation. This was similar to another study where the quick response of the chatbot was deemed not humanlike and perceived as a disadvantage [28]. To address this issue and advance the chatbot, improving its algorithm and continuing training it using AI and machine learning techniques based on feedback from a larger sample size is crucial. Considering that the use of AI chatbots in health care is still in its early stages, this finding holds particular significance for designing AI chatbots. To enhance usability and promote the implementation of AI chatbots in health care, the chatbots must possess the ability to initiate natural conversations with humanlike characteristics. In addition, they should be equipped to effectively address users’ questions and concerns while ensuring the security and safety of users’ information.

The chatbot’s plain interface and simple design were popular among MSM. Digital health interventions are useful, but knowing how to navigate a digital system sometimes could be daunting for users. Through this study, we are clear that accurate and simple responses without errors and redundant information were key to the acceptability of AI chatbots among MSM. Our participants reported that the AI chatbot helped them avoid societal stigma and protected their privacy, which increased their acceptability of using the chatbot to test for HIV. This finding is consistent with our previous formative research finding that addressing sociocultural barriers can facilitate the acceptance of an AI chatbot [13]. The chatbot does not require users to provide registration information. Therefore, it can maintain participants’ anonymity. However, it is still necessary for the chatbot to clarify to users that the backend researchers and engineers who have access to users’ conversations and information will not expose users’ information to others. This suggestion is consistent with our study findings and some previous studies showing that mHealth interventions could improve HIV testing rates if users’ anonymity were guaranteed [14,17].

Some technical-related issues negatively affected the participants’ experience of navigating the chatbot. The inconvenient address input process and repeated steps owing to incomplete information contributed to the inconsistency and complexity of the chatbot, prompting participants to seek technical assistance. Many of these resulted from cultural differences, as the address options were designed based on overseas settings. This signifies the importance of tailoring the chatbot to the local context to improve usability. In addition, using localized language could also enhance the participants’ satisfaction with the chatbot. Despite the challenges inherent in adopting novel technology, the advantages of using chatbots to connect with high-risk populations could significantly impact the efforts to address public health emergencies.

In line with other studies conducted in Malaysia, MSM are keen to peruse the information on PrEP and mental health, particularly the information on where the PrEP and mental health clinics are located [14]. Most participants in our study felt that they would like to see more information through the chatbot introducing AIDS, its treatment, mental health issues, and sexually transmitted infections to better understand and manage AIDS, including how to prevent high-risk behaviors and where to seek timely help [12]. In Malaysia, professional and MSM-friendly care for mental health needs to be developed as most MSM reported that culturally sensitive information and resources regarding mental health issues were difficult to obtain. Interestingly, researchers have identified several obstacles to the adoption of AI chatbots for mental health care among users [31]. These include concerns related to privacy risks, restricted conversational engagement, negative user perceptions of personality traits (such as rudeness, lack of empathy, patronization, and being judgmental), and a lack of trust in the app’s creators. Nevertheless, Eli chatbot overcame all these challenges by having a language that merges expertise and respect for the user, ensuring speech that is gender neutral and devoid of stigmatizing elements [29]. Our AI chatbot also had a similar language as Eli, which warrants future support on mental health issues.

To increase the use of the AI chatbot, it needs to be embedded in social media platforms that MSM frequently use. The geosocial networking apps where MSM find sexual partners, such as Grindr, Hornet, and Blued, and websites owned by nongovernmental organizations and MSM-friendly clinics are important venues to advertise the AI chatbot. MSM preferred these platforms because they are trusted and frequently used by MSM. Dissemination of the AI chatbot should be promoted among young MSM who use geosocial networking apps to find sexual partners because they are at a higher risk for HIV. Through this study, we found that to embed an AI chatbot into an internet-based platform for health promotion, researchers and engineers must consider the platform’s characteristics, including its target population, level of privacy, and user-friendliness. Findings from this study will be used to improve the AI chatbot before testing on a larger scale through a national observational study in Malaysia. AI chatbots are a promising tool for promoting HIV testing and prevention. The AI chatbot must be made visible to MSM to increase its usability among MSM. Adopting the right dissemination strategies is key to increasing the visibility of AI chatbots and bringing significant impact to the MSM community. In addition, it is important for researchers to consider the sustainability of AI chatbots for MSM care in a context where sex-same sexual behaviors are criminalized. The policies and laws in Malaysia pose significant challenges on the sustainability of leveraging AI and securing funding for MSM care research. In such a political environment, it is crucial for researchers to collaborate with local nongovernmental organization and MSM-friendly clinics that operate within the existing Malaysian legal framework. Future research should focus on developing innovative and culturally tailored AI interventions to combat HIV among MSM, promote public health in Malaysia, and advocate for changes in discriminatory policies and laws to enhance the testing, implementation, and sustainability of these AI interventions.
Limitations

Testing the AI chatbot among its end users (ie, MSM) was an important step in determining its feasibility and acceptability in Malaysia and collecting feedback to improve the chatbot further. Although this study contributed important scientific knowledge, it had several limitations. One of the limitations is that we only included MSM who can read English, as the AI chatbot is currently only available in this language. Thus, the reach of the AI chatbot may be limited only to those fluent in English, which is not the case for most MSM in Malaysia. Therefore, our findings may not be generalizable to MSM who cannot read English. Considering that Malaysia is a multilingual country with Bahasa Malaysia as the official language, the chatbot must be improved to communicate in Bahasa Malaysia or Mandarin to reach a wider audience and promote greater access to HIV self-testing and PrEP. In addition, our participants were highly educated; this may lead to bias as they might possess a certain level of knowledge and health literacy, thus facilitating their interactions with the chatbot. Therefore, the findings may differ in the less educated or literate group. In addition, our study only included MSM aged ≥18 years; therefore, the study findings do not capture the perceptions of younger MSM who are typically more tech-savvy and susceptible to HIV. Although obtaining consent from younger MSM in Malaysia for HIV-related research is a significant challenge, future studies should consider conducting surveys and interviews with MSM aged <18 years who can provide insights into the experiences and needs of the younger MSM.

Conclusions

The AI chatbot was found to be feasible and acceptable among MSM, highlighting features, such as being informative, being able to respond to users’ questions, and having a simple and user-friendly interface. Adapting the AI chatbot to local cultures, including support for other languages, and providing additional information such as mental health support, risk assessment for sexually transmitted infections, AIDS treatment, and the consequences of contracting HIV would contribute to the successful implementation and dissemination of the AI chatbot in Malaysia.

Acknowledgments

This work was supported by research grants from the National Institutes of Health (R21TW011663 and R33TW011663 for ZN and FLA; R21AI152927 for JAW, FLA, and ZN; and D43TW011324 for FLA) and the Fund for Lesbian and Gay Studies Research Award from Yale University for ZN.

Authors' Contributions

MHC and ZN wrote the first draft of the manuscript. YNG, NAMS, and KSN assisted in both developing the qualitative analysis and writing the qualitative results. All authors contributed to the development of the study protocol, revised the subsequent version of the manuscript, and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
An email that the research assistant sent to participants introducing the human-chatbot interaction.

Multimedia Appendix 2
The list of beta testing tasks.

Multimedia Appendix 3
The guide on chatbot beta testing.

Multimedia Appendix 4
The scales used for measuring outcome variables.

Multimedia Appendix 5
Participants’ insights with illustrative quotes.

Multimedia Appendix 6
References
1. HIV and AIDS. World Health Organization. 2022. URL: https://www.who.int/news-room/fact-sheets/detail/hiv-aids [accessed 2023-12-26]


Abbreviations

AI: artificial intelligence
mHealth: mobile health
MSM: men who have sex with men
PrEP: pre-exposure prophylaxis
UTAUT: Unified Theory of Acceptance and Use of Technology

©Min Hui Cheah, Yan Nee Gan, Frederick L Altice, Jeffrey A Wickersham, Roman Shrestha, Nur Afiqah Mohd Salleh, Kee Seong Ng, Iskandar Azwa, Vimala Balakrishnan, Adeeba Kamarulzaman, Zhao Ni. Originally published in JMIR Human Factors (https://humanfactors.jmir.org), 26.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Human Factors, is properly cited. The complete bibliographic information is available online.
information, a link to the original publication on https://humanfactors.jmir.org, as well as this copyright and license information must be included.
Viewpoint

The Temperature Feature of ChatGPT: Modifying Creativity for Clinical Research

Joshua Davis1,2*, BS; Liesbet Van Bulck3,4*, RN, PhD; Brigitte N Durieux1, BS; Charlotta Lindvall1,5,6, MD, PhD

1Department of Psychosocial Oncology and Palliative Care, Dana-Farber Cancer Institute, Boston, MA, United States
2Albany Medical College, Albany, NY, United States
3KU Leuven Department of Public Health and Primary Care, KU Leuven-University of Leuven, Leuven, Belgium
4Research Foundation Flanders (FWO), Brussels, Belgium
5Department of Medicine, Brigham and Women's Hospital, Boston, MA, United States
6Harvard Medical School, Harvard University, Boston, MA, United States
*these authors contributed equally

Corresponding Author:
Charlotta Lindvall, MD, PhD
Department of Psychosocial Oncology and Palliative Care, Dana-Farber Cancer Institute
450 Brookline Ave
Boston, MA, 02215
United States
Phone: 1 617 632 6464
Email: charlotta_lindvall@dfci.harvard.edu

Abstract

More clinicians and researchers are exploring uses for large language model chatbots, such as ChatGPT, for research, dissemination, and educational purposes. Therefore, it becomes increasingly relevant to consider the full potential of this tool, including the special features that are currently available through the application programming interface. One of these features is a variable called temperature, which changes the degree to which randomness is involved in the model’s generated output. This is of particular interest to clinicians and researchers. By lowering this variable, one can generate more consistent outputs; by increasing it, one can receive more creative responses. For clinicians and researchers who are exploring these tools for a variety of tasks, the ability to tailor outputs to be less creative may be beneficial for work that demands consistency. Additionally, access to more creative text generation may enable scientific authors to describe their research in more general language and potentially connect with a broader public through social media. In this viewpoint, we present the temperature feature, discuss potential uses, and provide some examples.

(JMIR Hum Factors 2024;11:e53559) doi:10.2196/53559

KEYWORDS
artificial intelligence; ChatGPT; clinical communication; creative; creativity; customization; customize; customized; generation; generative; language model; language models; LLM; LLMs; natural language processing; NLP; random; randomness; tailor; tailored; temperature; text; texts; textual

Introduction

ChatGPT [1] is a large language model developed by OpenAI that currently has over 100 million users [2]. As its popularity continues to grow, clinicians and researchers are among many considering its potential applications in health care and academia. In a short time, ChatGPT has been extensively published [3], with clinical researchers exploring its potential utility for a variety of tasks, including answering patient questions [4,5], generating clinical summaries [6], and abstracting data from important documentation (eg, computed tomography reports) [7].

When using ChatGPT, one can interact through the website by providing a single prompt or engaging in a conversation. In addition to this more well-known web-based version of ChatGPT, there is also an application programming interface (API) that allows for more customization and flexibility. With the API, users can programmatically interact with ChatGPT and modify features for their specific use case. Although this approach may currently require more technical expertise for clinicians to use, its features may become available on the web...
interface in future iterations of the tool. Therefore, these features are important to understand and relevant to discuss in terms of their meaning for clinicians and researchers in advance of their more widespread use. Additionally, they have direct implications for introducing greater reproducibility in use cases where this matters.

**The Temperature Feature of ChatGPT**

ChatGPT generates text through a probabilistic language modeling approach, where it writes responses word by word, calculating the most likely next word in the sequence. A key feature that influences this behavior is called temperature [8, 9]. In this context, temperature is a value from 0 to 2 that adjusts how random each subsequent word in the chat output is. A value of 0 will give the most probable word and, thus, the least variability. As the value increases toward and beyond 1, the next word becomes less probable, leading to more randomness and “creativity” in the response. This feature can currently be adjusted in the API, where the default value is 1 [9].

The ability to adjust the “creativity” of ChatGPT output should also be of interest to clinicians and medical researchers using the tool. By accounting for temperature, large language models such as ChatGPT can be tailored for different use cases. Lowering ChatGPT’s creativity level would be preferable for tasks that require more consistent outputs; for clinicians and researchers, tasks of this sort may include summarizing patient data (eg, symptoms and medications) or streamlining administrative tasks (eg, billing inquiries and patient registrations). Alternatively, increasing the creativity level may provide better outputs for less structured tasks and may specifically hold relevance for improving clinicians’ and researchers’ digital communication with other health care workers, patients, and a wider audience.

Currently, many clinicians may find it difficult to engage with the broader public when it comes to disseminating findings. A lack of public accessibility can limit the impact of research. Creativity proves to be a powerful ally in communication, as it simplifies complex concepts, fosters critical thinking, and enhances the accessibility of intricate information. These factors collectively contribute to increased audience engagement and more effective message transmission. This feature has value for developing patient education materials, medical courses, science communication, and public research dissemination [10]. Increasing the creativity level could be useful when trying to report research results on social media [11], at conferences, or in scientific publications. Among other things, researchers could use a more creative ChatGPT to help come up with tags that apply to their work, brainstorm analogies or metaphors for scientific concepts, and check the tone of their writing.

**Illustrating This Feature: Example**

As an illustration of the “creativity” feature in ChatGPT, we have provided ChatGPT with the abstract of a recently published article on the development of a keyword library for capturing symptom talk in oncology [12]. While setting the temperature to low (0.0), medium (0.5), and high (1.0), we asked ChatGPT to create a tweet for a public audience, a title for a scientific journal article, and a title for a keynote address. Example ChatGPT-generated responses can be found in Table 1.

**Table 1.** ChatGPT-generated responses on the prompts for the different temperatures. For each temperature level, the same prompts were run 10 times in a ChatGPT model (text-davinci-003).

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Low (0.0)</th>
<th>Medium (0.5)</th>
<th>High (1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweet for public audience</td>
<td>We developed a keyword library to detect symptom talk in transcribed clinical conversations. Our findings suggest that physical symptoms are more easily detected than psychological symptoms, and that more advanced computational models are needed to detect symptom expressions. #AI #Healthcare #SymptomDetection</td>
<td>Have you heard about the new study that developed and evaluated a natural language processing keyword library to detect symptom talk in oncology conversations? The findings highlight the need for more advanced computational models to detect symptom expressions in transcribed clinical conversations. #Oncology #SymptomTalk #NLP #ComputationalMethods</td>
<td>Check out how computational methods can help us detect patient symptoms in clinical conversations! A recent study developed &amp; evaluated a keyword library for capturing expressed symptoms, highlighting the need for more advanced models for real-time detection of symptoms at scale. #AI @healthinformatics</td>
</tr>
<tr>
<td>Title for scientific journal</td>
<td>Detecting Symptom Talk in Transcribed Outpatient Oncology Conversations: Developing and Evaluating a Natural Language Processing Keyword Library</td>
<td>Keyword-Based Detection of Symptom Talk in Oncology Conversations: Developing a Natural Language Processing Library and Exploring Opportunities for Advanced Computational Models</td>
<td>Unlocking the Ambiguity of Symptom Talk: Developing and Evaluating a Natural Language Processing Library for Detecting Outpatient Oncology Conversations</td>
</tr>
<tr>
<td>Title of a keynote address</td>
<td>Unlocking the Potential of Computational Methods for Detecting Symptoms: Achieving Better Patient Care Through Natural Language Processing</td>
<td>Harnessing the Power of Computational Methods to Detect Patient Suffering: A Study of Symptom Talk in Oncology Conversations</td>
<td>The Ambiguity of Symptom Expressions: Utilizing Computational Methods to Better Attend to Patient Suffering</td>
</tr>
</tbody>
</table>

The examples shown in Table 1 illustrate that when the creativity level in ChatGPT is adjusted, slightly different responses are generated; these different creativity levels may provide more useful output depending on the task at hand. For example, a tweet created with a high level of creativity includes an exclamation mark and directly addresses the audience. Therefore, it may be more engaging compared to tweets with a low or medium creativity level. The title of the article and keynote generated with high “creativity” were more surprising and potentially less useful for these tasks, although this may
depend on the context, setting, and personality of the user. For these tasks, the low- and medium-creativity titles were more straightforward. Importantly, these lower values do translate to more consistent responses. We ran each of these prompts 10 times, and at a temperature level of 0, all responses were identical. Given ChatGPT’s normally variable output, this feature holds exciting implications for scenarios where consistency and reproducibility are preferred.

In addition to the results reported above, we have also experimented with adjustments in temperature level using other ChatGPT models (ie, gpt-3.5-turbo-1106, gpt-3.5-turbo-instruct, and gpt-4-1106-preview). All outputs appear in Multimedia Appendix 1. In contrast to what we found when using the ChatGPT model “text-davinci-003,” some other models showed some variability, even at a temperature level of 0. Regardless, the relative variability of outputs is still modified by temperature, with a higher temperature increasing creativity. Users should consider and test how temperature impacts outputs within the model they are using.

In the examples provided above, we have demonstrated how adjusting the level of creativity can enhance science communication, making it more engaging. However, it is crucial to also acknowledge the potential risks associated with increasing creativity, especially for clinical cases. Using ChatGPT with high creativity settings in clinical contexts, such as for summarizing patient medical data, can be problematic. Excessive creativity might lead to the embellishment or misrepresentation of crucial information, either by omitting vital details or interpreting data too liberally. Such inaccuracies could impact patient treatment and outcomes. Therefore, it is advisable to lower the creativity level of ChatGPT in clinical applications. By doing so, we ensure that the summarized information remains faithful to the original data, thereby prioritizing accuracy and reliability over creative expression.

In summary, the temperature feature of ChatGPT allows users to adjust the level of “creativity.” Although no previous articles have discussed or investigated this feature for its use in clinical research, it shows promising potential for clinicians and researchers. Both high and low creativity levels could have interesting applications for health care and may broaden the ways clinicians and researchers consider using artificial intelligence (AI) tools to close gaps in areas such as digital communication. ChatGPT documentation suggests using a temperature value of 0 to 0.2 for more focused (less creative) tasks and 0.8 to 1 for more random (more creative) tasks [9]. As large language models are variable and use case dependent, we strongly suggest testing and validating the proper temperature level for your specific use case. While this feature is a powerful tool that could be useful for creating easy-to-understand summaries, captivating social media posts, or making complex information more accessible to a wider audience, the parameters need to be carefully tweaked to find a balance between coherence and creativity and to tailor to specific needs. Looking ahead, as AI continues to advance in the health care sector, the temperature feature can play a pivotal role in health care applications in generative AI, unlocking the potential for more accurate, empathetic, or creative interactions between AI and health care stakeholders.

Acknowledgments
This work was supported by the Research Foundation Flanders (FWO; grants 1159522N and K210723N to LVB) and by the National Institutes of Health, National Institute on Aging Medical Student Training in Aging Research Program (grant 5T35AG038027-13 to JD). The funding sources had no involvement in the study design, analyses, or interpretation of the data; writing of the report; or decision to submit the paper for publication. During the preparation of this work, the authors used ChatGPT in order to generate the examples reported in the paper and to assist with the writing process. The use of a large language model has greatly assisted us in rephrasing and ensuring the clarity and effectiveness of our language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data Availability
All data generated or analyzed during this study are included in this published article and Multimedia Appendix 1.

Authors’ Contributions
JD and LVB contributed equally and share first authorship. LVB, JD, and BND contributed to the conception and design of the study and drafted the paper. CL critically revised the paper. JD and LVB both accessed and verified the underlying data reported in the manuscript. All authors approved the final version of the manuscript and had full responsibility for the decision to submit for publication.

Conflicts of Interest
None declared.

Multimedia Appendix 1
All data presented in this article, ChatGPT outputs for tests of 3 prompts across 3 temperature values for 3 different models (gpt-3.5-turbo-1106, gpt-3.5-turbo-instruct, and gpt-4-1106-preview; 100 runs for each test), and a summary document describing the multiple model tests.
References

Abbreviations
AI: artificial intelligence
API: application programming interface

©Joshua Davis, Liesbet Van Bulck, Brigitte N Durieux, Charlotta Lindvall. Originally published in JMIR Human Factors (https://humanfactors.jmir.org), 08.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Human Factors, is properly cited. The complete bibliographic information, a link to the original publication on https://humanfactors.jmir.org, as well as this copyright and license information must be included.