Contents

Original Papers

Mobile Electronic Devices as Means of Facilitating Patient Activation and Health Professional Empowerment Related to Information Seeking on Chronic Conditions and Medications: Qualitative Study (e26300)
Hyllore Imeri, Shane Desselle, Dardan Hetemi, Kreshnik Hoti. ................................................................. 3

Supervised Exercise Therapy Using Mobile Health Technology in Patients With Peripheral Arterial Disease: Pilot Randomized Controlled Trial (e24214)
Katrin Paldan, Martin Steinmetz, Jan Simanovski, Christos Rammos, Greta Ullrich, Rolf Jánosi, Susanne Moebus, Tienush Rassaf, Julia Lortz. 14

Effect of Physician-Pharmacist Participation in the Management of Ambulatory Cancer Pain Through a Digital Health Platform: Randomized Controlled Trial (e24555)
Lu Zhang, Howard McLeod, Ke-Ke Liu, Wen-Hui Liu, Hang-Xing Huang, Ya-Min Huang, Shu-Sen Sun, Xiao-Ping Chen, Yao Chen, Fang-Zhou Liu, Xiao Jian. ......................................................................................... 30

Enabling Wearable Pulse Transit Time-Based Blood Pressure Estimation for Medically Underserved Areas and Health Equity: Comprehensive Evaluation Study (e27466)
Venu Ganti, Andrew Carek, Hewon Jung, Adith Srivatsa, Deborah Cherry, Levather Johnson, Omer Inan. 47

General Practitioners’ Perceptions of the Use of Wearable Electronic Health Monitoring Devices: Qualitative Analysis of Risks and Benefits (e23896)
Lucia Volpato, Maria del Rio Carral, Nicolas Senn, Marie Santiago Delesosse. ................................................... 65

Assessing Electrocardiogram and Respiratory Signal Quality of a Wearable Device (SensEcho): Semisupervised Machine Learning-Based Validation Study (e25415)
Haoran Xu, Wei Yan, Ke Lan, Chenbin Ma, Di Wu, Anshuo Wu, Zhicheng Yang, Jiachen Wang, Yaning Zang, Muyang Yan, Zhengbo Zhang. 76

Comparison of the Validity and Generalizability of Machine Learning Algorithms for the Prediction of Energy Expenditure: Validation Study (e23938)
Ruairi O'Driscoll, Jake Turicchi, Mark Hopkins, Cristina Duarte, Graham Horgan, Graham Finlayson, R Stubbs. 106

Viewpoint

Mobile Apps as Audience-Centered Health Communication Platforms (e25425)
Michael Mackert, Dorothy Mandell, Erin Donovan, Lorraine Walker, Mike Henson-Garcia, Lindsay Bouchacourt. ............... 42
Review

Evaluating the Validity and Utility of Wearable Technology for Continuously Monitoring Patients in a Hospital Setting: Systematic Review (e17411)

Vikas Patel, Ani Orchanian-Cheff, Robert Wu
Mobile Electronic Devices as Means of Facilitating Patient Activation and Health Professional Empowerment Related to Information Seeking on Chronic Conditions and Medications: Qualitative Study

Hyllore Imeri¹,², MPharm; Shane Desselle³, PhD; Dardan Hetemi¹, PhD; Kreshnik Hoti¹, PhD

¹Division of Pharmacy, Faculty of Medicine, University of Prishtina, Pristina, Kosovo
²Department of Pharmacy, University of Mississippi, Oxford, MS, United States
³Department of Pharmacy, Touro University California, Vallejo, CA, United States

Corresponding Author:
Kreshnik Hoti, PhD
Division of Pharmacy
Faculty of Medicine
University of Prishtina
George Bush Street #31
Pristina, 10000
Kosovo
Phone: 383 44945173
Email: kreshnik.hoti@uni-pr.edu

Abstract

Background: Patient activation has an impact on the management of patients’ health, clinical outcomes, and treatment costs. Mobile electronic devices (MEDs) have shown the potential to engage patients in wellness behavior. Furthermore, the potentially positive role of MEDs is evident in supporting health professionals in their practice.

Objective: This study aims to explore the impact of MEDs on patient activation to search for information on chronic conditions and medications and the impact of MEDs on the empowerment of health professionals or future health professionals.

Methods: We conducted 6 focus groups—2 with health sciences students, 2 with health professionals, and 2 with hospitalized patients with chronic conditions. A protocol comprising eight questions was used to guide discussions. Audio-recorded data were transcribed verbatim and analyzed thematically; a ranking system was used to analyze the relevance of identified themes and subthemes, using a coding system depicted by the + symbol, to indicate different relevance levels.

Results: Our results suggest that MEDs can positively affect patient activation to search for chronic conditions and medication information by facilitating patients’ information-seeking behavior. Key drivers leading to patients’ activation to seek information related to chronic conditions and medications through MEDs were the accessibility and abundance of available and detailed information, reduced search time, information updates, and convenience in finding information at any time and place. The lack of accurate information in one’s native language, access to incorrect information, and limited access to the internet were key obstacles to seeking information related to chronic conditions and medications via MEDs. In addition, findings of this study suggest that MEDs in general and mobile apps, in particular, may have a positive impact on the work routine of health care professionals as they enable them to make quicker decisions by accessing the required information faster, thus improving practice efficiency. Furthermore, the appropriate usage of MEDs by patients for seeking information about their chronic conditions and medications may positively impact the physician-patient relationship. All focus groups recognized the questionable reliability of health information on the internet and its potential negative effects on patients. Therefore, our findings suggest the need for an additional role of health professionals in assisting patients in using MEDs to search for health-related information, such as providing reliable websites and mobile apps where patients can safely search for health-related information on the web.

Conclusions: The use of MEDs may help activate patients to seek chronic conditions and medication-related information, potentially leading to better management of their chronic conditions and medications. Our findings also highlight the positive impact MEDs may have on empowering health professionals in their practice and the need for health professionals to help patients through specific education that addresses MEDs utilization for chronic conditions and medication information seeking.
patient activation; mobile electronic devices; health professionals; chronic conditions; medications

Introduction

Background

Patient activation is a behavioral concept defined as “the individual’s knowledge, skills and belief in managing his/her health and healthcare” [1]. Patient activation enables an understanding of why some patients engage and are actively involved in their health whereas others are not. Considering the relevance of patients’ roles, health care settings have moved to a more patient-centered model, where patients are encouraged to be effective managers of their health care [2,3]. A higher patient activation level is associated with a wide range of better health outcomes [4], health-related behaviors, and health care costs [5]. In addition, although nearly half of patients assess information about their medical condition or treatments on the web [6], evidence suggests that activated persons are more likely to use web-based health information [7].

Although there have been rapid developments in mobile health (mHealth) apps, their effects on improving health and health care remain unclear [8,9]. However, the role of mobile electronic devices (MEDs) concerning patient engagement and facilitation of communication between patients and health professionals has been recognized. In this regard, several programs using mobile devices to engage and educate patients have been designed [8,10]. Evidence suggests that mobile phone apps can facilitate medication adherence by using functions such as reminder alerts and providing access to drug information [11]. Furthermore, many mobile apps to support health care professionals during their practice have been developed [12-14], thus promoting personalized and efficient health care. Therefore, the potential of MEDs, including laptops, tablets, and especially mobile phones, related to facilitating activation to navigate health information on the web is expected to be highly advantageous to not only patients but also health professionals.

Patient activation levels may increase with time, making patient activation an important focus of interventions to improve health behaviors and outcomes [6,7]. There are four patient activation levels that patients go through as they become more activated, from being disengaged and overwhelmed to maintaining positive health-related behaviors and pushing further [1]. Providing health information to patients is one of the initial steps of engagement in health care self-management. Patient education through health-related informative letters has been found to drive patient activation in patients with chronic conditions such as hypertension, suggesting a significant role of health information in patient activation [15]. However, considering the increased tendency to seek health information on the web [6] and the high rates of MED ownership [16], a potential role of MEDs in facilitating the process of patient education, and consequently the overall patient activation, could be suggested.

Objectives

Currently, there is limited evidence exploring how MEDs could facilitate the activation of patients seeking information on the web about their health and medications and the acceptance of MEDs on health professionals’ work. This study aims to explore (1) the role of MEDs on patient activation toward seeking information about chronic conditions and medications and (2) the impact of MEDs on health professionals or future health professionals’ empowerment.

Methods

Study Design

This qualitative study used focus groups (FGs) as a data collection method, which allowed us to explore various perspectives. Approvals to conduct the study and access to patients were provided by the ethics committee of the Faculty of Medicine, University of Pristina, and the University Clinical Center of Kosovo (UCCK) office for personal data protection, respectively.

Study Setting and FG Participants

In total, 6 FGs were organized with 4 to 7 participants. These numbers were chosen based on the literature suggesting that themes saturation can be achieved on this basis [17]. In particular, saturation was sought despite seeking input from various groups involved in the medication use process, including patients, health professionals, and student health professionals. To examine the role of MEDs in patient activation toward health information seeking from various perspectives, we conducted FGs with patients, health professionals, and future health professionals. Health sciences students were selected for FG inclusion because of their likelihood of being well-versed and high MEDs users, thus presenting the potential to better understand the acceptance of MEDs from future health professionals. Of the 6 FGs, 2 (33%) were conducted with health science students at the University of Pristina, 2 (33%) with health professionals working at the UCCK, and 2 (33%) with patients with chronic conditions who were hospitalized at the UCCK and receiving more than one medication. FGs were organized on the premises of the UCCK and the Faculty of Medicine, University of Pristina.

Recruitment

A purposive selection strategy was used, targeting health professional staff of the UCCK, patients with chronic conditions receiving health care treatment at the UCCK, and health sciences students. The study’s setting in Kosovo presents an ideal environment in which to conduct the study, given the high use of internet technology and connectivity of Kosovo’s population. However, there is lack of data that would suggest that Kosovo’s residents leveraged this use into active involvement into technologies that potentially assists them with management of their health conditions or seeking information for health

https://mhealth.jmir.org/2021/8/e26300
maintenance purposes. Students were recruited using snowball sampling, whereas health professional staff and patients were approached by the facilitator of FGs and the principal author on the UCCK premises and invited to participate in the study. Before inviting patients and health professionals to participate in the study, researchers informed the respective clinics’ chiefs of staff. Patients were approached during the optimal time of the day after the physicians’ consultations, and FGs were conducted on the same day. Health professionals chose their optimal time and date, which was before starting their shifts or during breaks. Potential subjects who agreed to participate also received an information letter and a letter of consent as an invitation to be a participant in the FG. The design of the FG structure and protocol was based on data from previous studies [18].

FG Discussion Guide
The development of the question protocol for this study considered the concept of patient activation and the utility of MEDs in seeking health information on the web relevant to patients’ health care management. In addition, the researchers who conducted the FGs clarified the definition of MEDs verbally as part of the opening statements by taking examples of mobile phones, tablets, and laptops. The question protocol consisted of an opening question, four transitory questions, and three key questions. The opening question was related to searching for health and medication information and participants’ access to MED use. Transitory questions were related to difficulties in seeking information, factors that would help overcome these obstacles, and patient views on health professionals’ roles. Key questions were related to the potential drivers leading to patient activation to seek information on chronic conditions and medications and the role of MEDs during this process.

The FGs were conducted by a facilitator. The principal author also attended the meetings, took notes, and audio recorded to further facilitate data analysis and mitigate against the potential bias introduced by the facilitator. All FGs were conducted in the native language of the participants, which is Albanian, and were held between May and June 2018. Audio-recorded data from FG meetings were transcribed verbatim into Microsoft Word version 2013 and translated and reviewed by 3 researchers.

Qualitative Analysis
Data were analyzed thematically in Microsoft Word, using the open, axial, and selective coding strategy [19], initially by one of the researchers of this study, and a second analysis and review were conducted by the other researcher of the study. A ranking system was used to analyze the relevance of the identified themes and subthemes. This approach has been previously reported in the literature [20]. If a similar comment was repeated for a given issue, the + symbol was used. If a comment was repeated only within 1 FG, it was marked +, if a comment was repeated in 2 to 3 FGs, it was marked as ++, and when a comment was repeated in all FGs, it was marked with +++; where + indicates low relevance, ++ indicates average relevance, and +++ indicates high relevance.

Results
Overview
A total of 6 FGs involving 31 participants were conducted on the UCCK and Faculty of Medicine premises in Pristina, Kosovo. The 2 FGs with health care professionals were conducted with the health care professional staff of the Clinic of Nephrology and the Infectious Diseases Clinic at the UCCK, 4 and 7 participants, respectively, with ages ranging from 30 to 55 years, of which 73% (8/11) were female. The 2 FGs with patients were conducted with patients admitted to either the Dermatology Clinic (UCCK) or Clinic of Hematology (UCCK), with 4 participants per FG. Patient age ranged from 45 to 75 years, of which 63% (5/8) were women. In addition, a total of 12 students agreed to participate. They were divided into 2 FGs, each composed of 6 participants, with ages ranging from 20 to 25 years, and equal participation of both men and women. The duration of each FG meeting was approximately 50 minutes.

In FG discussions, five main themes were highlighted referring to patient activation via MEDs to seek information on chronic conditions and medications. These themes pertain to motives for seeking information on chronic conditions and medications via MEDs, difficulties and obstacles to seeking information on chronic conditions and medications via MEDs, the overall activation level of patients, the impact of MEDs in activating patients to seek health-related information, and the role of health professionals in facilitating the use of MEDs to enhance patient activation toward seeking health-related information. The themes have been described in more detail below.

Motives for Seeking Information on Chronic Conditions and Medications via MEDs
For most students and health care professionals, the internet was the main source of information on chronic conditions and medication, whereas health care professionals were the main source of information for most patients. Key identified motives for using MEDs to seek health-related information were using time efficiently, reducing information ambiguity, and getting the most up-to-date medication information. The subthemes with corresponding comments are shown in Textbox 1.
Textbox 1. Theme 1—summary of comments corresponding to each of the subthemes.

<table>
<thead>
<tr>
<th>Detailed Medication Information (++)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“I use the applications mostly for medicines, about the effect of a herb, the action mechanism and the contraindications.” [student]</td>
</tr>
<tr>
<td>“We use MED more about drugs than for diseases because it is a problem to find information and read about diseases on Google.” [health professional]</td>
</tr>
<tr>
<td>“Yes, I even use Facebook. If there is any diet to lose weight, I use it. Even when I receive medication information, I get informed, and when I go to the doctor, I ask him about what I have read, without the doctor’s advice I do not use anything.” [patient]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ease and Time Efficiency of Information Access (++)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“I think the reason why we use phones, or the Internet is often the time...you will immediately have thousands of publications or links that send you directly to the requested information. The time of searching for a particular problem is shortened, so it is a way of searching much faster than by searching in books.” [student]</td>
</tr>
<tr>
<td>“The use of MED relates to the comfort offered to find information at any time and in any place.” [health professional]</td>
</tr>
<tr>
<td>“We search drugs (online) because it is faster.” [patient]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reducing Information Ambiguity (++)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Yes, it has become essential to have a mobile with us. In case I encounter ambiguity...” [student]</td>
</tr>
<tr>
<td>“Yes, we use MED, but it is a bit of a problem to open the phone directly with the patient; then the patient perceives us as we do not know things, but we tell them that, eg, a drug could have 50 or 60 commercial names...” [health professional]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>New Drugs on the Market (++)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“…when I hear a new drug name, I search it at least to have the basic information.” [student]</td>
</tr>
<tr>
<td>“We use MED for health and medicines, but mostly to get information about a certain drug that comes out in the market.” [health professional]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Getting Up-to-Date Information (++)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“…information via mobile devices, gets the information fast and uses information that is more updated than books.” [student]</td>
</tr>
<tr>
<td>“…the reasons are to be updated for a certain new medication, to recall and recapture things that could have gone through in the second plan...” [health professional]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wider Range of Information (++)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“The main motive of looking for a drug is the interest in knowing more drugs because the basic literature is not sufficient.” [student]</td>
</tr>
<tr>
<td>“Internet search through MED provides a wider range of information, all areas are there.” [health professional]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Queries From Other Family Members or Society (+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Usually a certain medical condition that I have or someone in the family does, this is the key to pushing me to research, then also for faculty issues if I need something...” [student]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Medical Condition (++)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“On internet you can read something superficially, for illnesses we rather read in books.” [health professional]</td>
</tr>
<tr>
<td>“Our medical condition pushes us to search for information.” [patient]</td>
</tr>
</tbody>
</table>

Difficulties and Obstacles to Seeking Information on Chronic Conditions and Medications via MEDs

Unlike participants from patient-based FGs who expressed confidence in using MEDs for information retrieval, students and health professionals reported difficulties and obstacles while seeking health and medication information. However, it should be noted that many participants in patient-based FGs did not use MEDs for information searches specifically related to health and medications. The following key difficulties were identified: lack of accurate information in the native language, limited access to the internet, and access to incorrect information. The subthemes with the illustrated comments are shown in Textbox 2.
Lack of Accurate Information in the Native Language (++)
- “If the literature was in Albanian, it might have been much easier to understand, but there are data that are not electronically in Albanian, so in other languages it is more difficult to understand.” [student]
- “There is not much information in our [Albanian] language.” [health professional]

Inability to Access Scientific Journals Due to Subscriptions (+)
- “The problem is the inability to access publications. In many cases, if we want to read any publication, we read only the abstract.” [student]

Inability to Access the Internet (+++)
- “Another obstacle I see is access, connection to the Internet. There are quite accurate apps, but you cannot access them offline.” [student]
- “Another difficulty is the Internet access to our clinics, when we spent our mobile data then our internet access ends.” [health professional]
- “I do not have an internet connection.” [patient]

Access to Incorrect Information (++)
- “…if we look for information in electronic books, then I do not see any difficulty, but if I search for information across different web pages, then it could be incorrect.” [student]
- “There is information, but when it comes to different websites, you can read what people who are not competent for that matter have also written.” [health professional]

Overcoming Obstacles in Seeking Information on Health and Medication (++)
- “Initially, it would be good to have English language knowledge as most of the accurate and up-to-date information is in English, and this would help us to research more.” [student]
- “Free internet access should be provided in all clinics of UCCK.” [health professional]
- “We should have consultations with pharmacists of the clinics more often because we hardly see them. In this way, we always have to find the information by ourselves, to research it online or to consult with other doctors.” [health professional]

The Overall Activation Level of Patients
All participants supported the approach that, in general, it is the patients’ responsibility to actively engage in their health management and, therefore, achieve a higher level of activation. This was also the case for the use of MEDs to facilitate patient activation. According to them, the physician’s responsibility is in diagnosing and prescribing, whereas before and after this, it depends on the patient’s behaviors. However, some patients declared that they were not the only ones responsible for actively engaging in their health management. They see themselves and the physician at the same level of responsibility. The subthemes and related comments are listed in Textbox 3.

The Impact of MEDs in Activating Patients to Seek Health-Related Information
The MEDs’ impact on patients was related primarily to their empowerment by facilitating patient activation to seek health-related information. The impact was considered positive if the information was searched adequately and access to accurate and credible information was provided. However, a potential disadvantage to using MEDs from the patients’ perspective, was the possibility of bypassing the physician’s...
Health professionals suggested that when patients are more informed about their health condition, this facilitates health professionals’ work because of better patient understanding and communication level. To ensure the overall positive impact of MEDs, health professionals suggested the option of providing patients with access to credible health information. Furthermore, patients can be given access to websites with health professionals available to respond to their queries. Information uncertainty and the possibility of information misinterpretation were emphasized to be dangerous. Considering this, it was suggested that patients be informed but not decide about their health based solely on the information they find on the internet.

Textbox 4. Theme 4—summary of comments corresponding to each of the subthemes.

<table>
<thead>
<tr>
<th>Positive Impact of Mobile Electronic Devices in Facilitating Activation in Information Seeking (++++)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• “MED certainly facilitate the interest in seeking information on health and medicines...” [student]</td>
</tr>
<tr>
<td>• “It would be definitely helpful to use MED for health-related information.” [health professional]</td>
</tr>
<tr>
<td>• “They [MED] have an impact because everything that interests you can be found there, medicines, food, whatever you want.” [patient]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Impact of Mobile Electronic Devices in the Patient-Physician Relationship (++++)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• “Now for the doctor, it has become a bit harder because patients now are [easily] informed about certain types of illnesses, and doctors need to be more careful not to go directly to the diagnosis and medication.” [student]</td>
</tr>
<tr>
<td>• “When patients have access to this information, their rapport with doctors is changing a lot, and somehow it is thought that the doctor is replaceable because of technology...” [student]</td>
</tr>
<tr>
<td>• “If the patient gets that information from a site or trusted apps, it’s pretty good that when you are in a 24-hour shift, you have 60 to 70 patients within a day, you cannot explain a lot to everyone, and if the patient would be appropriately informed it would be much better and easier.” [health professional]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Potential Negative Impact of Mobile Electronic Devices in Patients (++++)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• “Inaccurate online information in patients who have no health information can lead to a worse condition due to the stress they create...We should not replace the doctor with the information that is on the Internet.” [student]</td>
</tr>
<tr>
<td>• “The patients have never read more, but also, they have never read nonsense things more. It is good to read, but not so that in every portal is the fluid which heals the heart diseases, the fluid that heals the kidney disease...” [health professional]</td>
</tr>
<tr>
<td>• “If I used the internet, I would be schizophrenic. I do not object to the technique, but it is very wide.” [patient]</td>
</tr>
<tr>
<td>• “The use of MED for information is a convenience, but it is not a security, so first we need to consult with the relevant doctor.” [student]</td>
</tr>
</tbody>
</table>

The Role of Health Professionals in Facilitating the Use of MEDs to Enhance Patient Activation Toward Seeking Health-Related Information

Students considered that health professionals’ role in providing information on health and medications is irreplaceable, and according to them, no matter what information can be obtained via the internet, the final source should be health professionals. According to students, there is currently a lack of time management by health professionals during patient contact, causing a lack of health information provision. Health professionals also see their role as a major responsibility but are nevertheless deficient because of the lack of available time. Patients also supported the importance and primary role of health professionals in providing information on health and medications. Regarding the role of health professionals in assisting with the use of MEDs, students suggested that patients should be referred to reliable databases or websites where they could be accurately informed, so the reading of unconfirmed scientific information would be prevented. Health professionals suggested that if they were more active in this role, it could have a significant impact, but this is not the case. Patients considered that if they were instructed to read about their health or medication via MEDs, they would do it and think it would positively affect their health. Subthemes with the comments of participants are shown in Textbox 5.
Textbox 5. Theme 5—summary of comments corresponding to each of the subthemes.

<table>
<thead>
<tr>
<th>Health Professionals Have a Key Role in Providing Health Information (+++)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• “Despite receiving information through the internet, the main and accurate source should be the doctor.” [student]</td>
</tr>
<tr>
<td>• “Informing the client about his/her condition for getting different medication should be done primarily by each health worker.” [health professional]</td>
</tr>
<tr>
<td>• “Health professionals have the main role [ie, in providing health information].” [patient]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Health Professionals' Recommendation on Seeking Health-Related Information Through Mobile Electronic Devices (+++)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• “It would have been good to inform patients in this regard, but I think very few do. I think it would be much better if you visit a doctor and they help you with using an app that will help you in the future.” [student]</td>
</tr>
<tr>
<td>• “I try to give the patient’s family members information on a certain drug, and I recommend them to read on the Internet what we are prescribing.” [health professional]</td>
</tr>
<tr>
<td>• “They accomplish this role very well; they tell us to read.” [patient]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Deficient Role of Health Professionals in Regard to Assisting Patients With Mobile Electronic Devices Used for Health Information (+++)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• “They have a very important role, but I think they do not use it...they do not give the [exact] information, because of time or other engagements. I think it would have been good to inform patients in this regard [ie, using apps, reading health and medications information through MED], but I think very few do so.” [student]</td>
</tr>
<tr>
<td>• “I do not believe it would be functional if a doctor recommends something to the patient to read...” [health professional]</td>
</tr>
<tr>
<td>• “It would have been very good if they would have assisted us in this aspect; I would have read and done so.” [patient]</td>
</tr>
</tbody>
</table>

Discussion

Principal Findings

This study explored the impact of MEDs on activating patients to seek information on chronic conditions and medications. Data were collected from students, health professionals, and patients. The findings of this study suggest that MEDs could help activate patients to seek information on health and medications. MEDs provide quick access to the information required and provide a large amount of information, which could facilitate patient activation to seek health-related information. Furthermore, a previous observational study showed that mHealth apps were most useful for obtaining general health and physical activity and nutrition information, finding no significant difference in mHealth app’ use on patients’ socioeconomic status [21]. Therefore, the potential role of mHealth apps is empowering patients, especially those with lower socioeconomic status, to improve their health outcomes. Considering that there is evidence that supports the effectiveness of mHealth technologies in clinical outcomes and patient-centered care outcomes such as patient satisfaction and patient engagement [22], this study’s findings are especially important as they further delineate the role of MEDs in facilitating patient activation toward seeking health-related information.

The findings of this study suggest a mutual acceptance of MEDs in patient activation to seek health-related information from patients, health professionals, and future health professionals. However, there were different reasons and approaches for all three subsamples when using the MEDs. Students’ motives for seeking health and medication information through MEDs included easy and quick access and the need for constant information updates that assist them in their studies. Technological barriers were identified and were consistent with previous studies [23], suggesting the need for improvement in access to library-licensed mobile resources. Health professionals seek information through MEDs mainly on new drugs while maintaining a preference for books when searching for information on diseases. This is particularly the case with brand drug names, which may be unknown to some health care staff. In addition, the findings of this study suggest that MEDs in general and mobile apps, in particular, may have a positive impact on the work routine of professional health care staff as they enable them to make quicker decisions with a lower degree of error, thus improving practice efficiency and management of clinical cases. These results are consistent with the findings of previous studies [24,25]. In addition to requiring information on their medications, patients were also interested in knowing more about complementary medications, nutritionally advantageous diets, and healthy lifestyles in general. This is because MEDs provide access to vast amounts of information and a source of unlimited information. Therefore, although findings from all FGs suggest a potential positive role of MEDs in patient activation in seeking health-related information, there were different motives and barriers for each of the study subsamples.

Patients using MEDs and receiving information on their health conditions felt more empowered in managing their health and suggested that higher activation levels could be achieved via the use of MEDs. In contrast, patients who did not use MEDs indicated a lower level of activation and difficulty in understanding the actions required to maintain a healthy lifestyle. However, it was interesting to note that none of the patients identified the existence of nonreliable health-related web-based information as an issue when using MEDs. Therefore, it enhances the need for health professionals to discuss the validity of web-based information that patients may consult and apply to their health.

In addition, although patients are getting more involved in their health care decisions, the proportion who are willing to take
serious action and change their health behaviors is still in the minority. It is thus helpful to know which types of patients might become more actively engaged and apt to take such behaviors, and therefore the mechanisms that might be used to aid them and other patients [26]. Although MEDs could facilitate patient activation in seeking health-related information for some patients, the role of MEDs might be vague for some others. Therefore, besides evaluating the patient activation level, assessing the stage of change in the transtheoretical model (TTM) might help determine for which patients MEDs could affect patient activation in seeking health-related information.

Additional research might consider using the TTM, an integrative biopsychosocial model aimed at predicting one’s likelihood of behavior change [27]. Using the TTM can help discern where patients are in their contemplation to engage and use mobile devices and technology for their health, and thus might be used in concert with the results of our study to design educational interventions and communication with patients.

This study suggests that health professionals can influence the activation of patients to seek health-related information using MEDs by assisting them in the use of MEDs to search for information, which would be an additional motive for patients. This guidance and assistance could occur during regular check-ups that patients with chronic conditions have. During this time, health care professionals could advise using certain credible websites to require health-related information and even try to conduct a web search related to the patient’s specific concerns. This assistance would facilitate patients’ interest in using MEDs for seeking information on health and medications. However, the health professionals’ use of MEDs might influence their ability to support and guide patients in their use related to seeking health-related information. Evidence indicates that adequate training of health professionals on this matter would influence health education and improve the population’s overall health on primary and secondary prevention bases [28]. However, it is promising that future health professionals tend to highly use MEDs, positing them in a better empowerment position to assist and advise their patients regarding seeking health-related information on the web than older cohorts of health professionals.

Furthermore, it is interesting to note that even patients who did not use MEDs for information on health and medications suggested that they would do so if health professionals would recommend specific information websites or mobile apps. The reliability of web-based health information remains a concern [29], and the quality of health information accessed by patients remains unevaluated [30]. Thus, as students and health professionals would recognize a valid data source, it would be questionable for patients with different training and backgrounds. In addition, evidence indicates that age differences play a vital role in credibility judgments among patients seeking health information on the web, showing that older adults have a higher tendency to passively accept web-based information compared with younger adults [31], thus suggesting a need for different training approaches to these populations. Finally, research indicates that the only predictor of mHealth use for self-management was patient information technology skills [32].

Therefore, health care professionals should advise all patients about MEDs, regardless of their age.

**Study Limitations**

This study had several limitations. First, the study lacks wider representativeness, as it was conducted in one city of Kosovo, and it did not include a wider range of participants who could be potential users of MEDs. This limitation can be considered minimal in the results obtained because participants originated from various parts of Kosovo, and a saturation point in terms of themes and subthemes was achieved even across diverse groups of individuals. The generalizability of this study’s results is questionable because of the sample size and sample characteristics. However, considering that in 2019 the global internet access rate was 51.4%, and it was estimated that 86.7% of the population in developed countries had internet access [33], data from a study in 2017 showed that 88.5% of households in Kosovo had internet access [34]. Therefore, this study’s findings would be more applicable in countries with similar internet access coverage. However, the qualitative research goal is not the generalizability as much as it is the generation of rich and contextualized understanding of unexplored phenomena [35].

In addition, FGs with health professionals consisted of specialists in various fields of medicine and nurses. The diversity of health professionals in FGs may have facilitated exploring different perspectives, although segmentation of these FG participants could have facilitated comparative data analysis. However, it has been previously reported that homogeneous groups of participants in FG meetings can capitalize on the shared experiences of participants [36]; thus, this approach was used to conduct separate FGs with students, health professionals, and patients. Finally, the FGs were conducted with participants in Kosovo, who were almost entirely of ethnic Albanian descent. As qualitative research, there was no instrumentation to translate directly; however, the concepts and theories from which the FG guide was constructed had their basis in the English language literature. Furthermore, this potential limitation should also be considered in lieu of the fact that regardless of location and language used, access to the internet, MEDs, and mobile apps has increased significantly worldwide; therefore, patients and health professionals are expected to exhibit similar behaviors when adopting technology. In this regard, it may be worth emphasizing that Kosovo is known to have the highest levels of household internet access in the world [37]. Nonetheless, we believe there is a unique internet- and media-related characteristic of our sample that may affect the use of MEDs to navigate health-related information. This uniqueness is derived from the fact that the Kosovo population has strong family and sociocultural ties with its large diaspora living overseas and with whom there is a high reliance on MEDs to exchange information on a regular basis. In this environment, MEDs users in Kosovo would be exposed to cross-cultural experiences derived from various societies that the Kosovo diaspora has influenced globally, which in turn may also have an impact on how they assess and interpret information. This characteristic of our sample as well as of similar populations groups merits further research to better understand the
implications on navigation via MEDs and the use of health-related information.

This study provides data referring to some of the basic motives for using MEDs for information search on health and medications and suggests the need for a change in health professionals’ approach to assist using MEDs to facilitate patient activation in this regard. There is a need for further research into the clinical and economic impact of using MEDs in facilitating patients’ activation to seek information on their chronic conditions and medications.

Conclusions
This study suggests that MEDs might help facilitate patients’ activation to seek information on chronic conditions and medications. The motives for searching for information through MEDs related to the activation of patients have been identified. The findings suggest that health professionals’ roles should be reconsidered to include additional assistance to their patients in using MEDs, specifically in recommending valid and trustworthy websites or mobile apps to search for information on chronic conditions and medications.

Acknowledgments
The authors wish to acknowledge and thank all of the FG participants of this study.

Conflicts of Interest
None declared.

References
5. Hibbard JH, Greene J, Overton V. Patients with lower activation associated with higher costs; delivery systems should know their patients’ scores. Health Aff (Millwood) 2013 Feb;32(2):216-222 [FREE Full text] [doi: 10.1377/hilaff.2012.1064] [Medline: 23381513]


UCCK: University Clinical Center of Kosovo

Edited by L Bais; submitted 06.12.20; peer-reviewed by LY Chen, D Zhang, M Harris; comments to author 11.01.21; revised version received 02.03.21; accepted 07.05.21; published 11.08.21.

Please cite as:
Imeri H, Desselle S, Hetemi D, Hoti K
Mobile Electronic Devices as Means of Facilitating Patient Activation and Health Professional Empowerment Related to Information Seeking on Chronic Conditions and Medications: Qualitative Study
JMIR Mhealth Uhealth 2021;9(8):e26300
URL: https://mhealth.jmir.org/2021/8/e26300
doi:10.2196/26300
PMID:34383686

©Hyllore Imeri, Shane Desselle, Dardan Hetemi, Kreshnik Hoti. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 11.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Supervised Exercise Therapy Using Mobile Health Technology in Patients With Peripheral Arterial Disease: Pilot Randomized Controlled Trial

Katrin Paldán1,2, Dr; Martin Steinmetz3, Dr; Jan Simanovski3, BA; Christos Rammos3, Prof Dr; Greta Ullrich3; Rolf Alexander Jánosi3, PD Dr; Susanne Moebus1, Prof Dr; Tiensh Rassaf3, Prof Dr; Julia Lortz1, PD Dr

1Institute for Urban Public Health, University Clinic of Essen, University of Duisburg-Essen, Germany, Essen, Germany
2Personal Analytics Centre of Competence, Department of Engineering Sciences, University of Duisburg-Essen, Essen, Germany
3Department of Cardiology and Vascular Medicine, West-German Heart and Vascular Center Essen, University of Duisburg-Essen, Essen, Germany

Corresponding Author:
Julia Lortz, PD Dr
Department of Cardiology and Vascular Medicine
West-German Heart and Vascular Center Essen
University of Duisburg-Essen
Hufelandstr 55
Essen, 45147
Germany
Phone: 49 201723 ext 84995
Email: julia.lortz@uk-essen.de

Abstract

Background: Mobile health interventions are intended to support complex health care needs in chronic diseases digitally, but they are mainly targeted at general health improvement and neglect disease-specific requirements. Therefore, we designed TrackPAD, a smartphone app to support supervised exercise training in patients with peripheral arterial disease.

Objective: This pilot study aimed to evaluate changes in the 6-minute walking distance (meters) as a primary outcome measure. The secondary outcome measures included changes in physical activity and assessing the patients’ peripheral arterial disease–related quality of life.

Methods: This was a pilot two-arm, single-blinded, randomized controlled trial. Patients with symptomatic PAD (Fontaine stage IIa/b) and access to smartphones were eligible. Eligible participants were randomly assigned to the study, with the control group stratified by the distance covered in the 6-minute walking test using the TENALEA software. Participants randomized to the intervention group received usual care and the mobile intervention (TrackPAD) for the follow-up period of 3 months, whereas participants randomized to the control group received routine care only. TrackPAD records the frequency and duration of training sessions and pain levels using manual user input. Clinical outcome data were collected at the baseline and after 3 months via validated tools (the 6-minute walk test and self-reported quality of life). The usability and quality of the app were determined using the Mobile Application Rating Scale user version.

Results: The intervention group (n=19) increased their mean 6-minute walking distance (83 meters, SD 72.2), while the control group (n=20) decreased their mean distance after 3 months of follow-up (−38.8 meters, SD 53.7; P=.01). The peripheral arterial disease–related quality of life increased significantly in terms of “symptom perception” and “limitations in physical functioning.” Users’ feedback showed increased motivation and a changed attitude toward performing supervised exercise training.

Conclusions: Besides the rating providing a valuable support tool for the user group, the mobile intervention TrackPAD was linked to a change in prognosis-relevant outcome measures combined with enhanced coping with the disease. The influence of mobile interventions on long-term prognosis must be evaluated in the future.

Trial Registration: ClinicalTrials.gov NCT04947228; https://clinicaltrials.gov/ct2/show/NCT04947228

(JMIR Mhealth Uhealth 2021;9(8):e24214) doi:10.2196/24214
KEYWORDS
peripheral arterial disease; mHealth; supervised exercise therapy; motivation; artery; exercise; mobile applications; lifestyle; well-being

Introduction

Background

The circulatory disorders of peripheral arteries due to atherosclerotic lesions, also known as peripheral arterial diseases (PAD), are the third most frequent manifestations of cardiovascular diseases (CVD) after coronary and cerebrovascular arterial diseases [1].

A primary goal in CVD treatment is to slow down disease progression and avoid major adverse cardiac or limb events. Nonetheless, patients with PAD lag behind those with coronary artery disease (CAD) in terms of optimal treatment patterns [2-4]. Although the survival rates for CAD and PAD have improved worldwide, PAD still comes with a high individual burden regarding the quality of life (QoL) and associated disabilities [2].

The individual restrictions in the daily life of patients with PAD are more important than statistical facts regarding mortality and morbidity. Intermittent claudication causes a progressive reduction of the pain-free walking distance (PWD), and it is an expression of worsening PAD [4]. This decrease in physical capability results in declining mental health and reduces patients’ QoL [5].

Supervised exercise therapy (SET) is a cornerstone in the conservative management of intermittent claudication and it extends the PWD. Even though SET is easy to practice and highly cost-effective, adherence to regular SET performance is relatively low [6,7]. The underuse of exercise can be partly explained by the lack of institutional resources [8] and both patients’ and physicians’ lack of interest in exercise [4,9].

Mobile health (mHealth) technologies increase incentives and provide digital support for patients with PAD on several treatment levels [10-12]. They potentially lead to higher exercise training adherence and widen the scope of patient-centered health care [13], but so far, studies show opposite results [11,14]. While patients with PAD highly desire specific support tools, and app stores are inundated with health and fitness apps, PAD-specific solutions are presently lacking [15].

Objective

We developed a smartphone app named TrackPAD [16] to provide PAD-specific support for SET. This pilot study aims to evaluate the TrackPAD application as to its suitability and feasibility in outcome measures relevant to the prognosis of PAD by assessing the participant’s 6-minute walking distance (meters).

Methods

Study Aims, Research Questions, and Outcomes

The TrackPAD pilot study aimed to answer the following research questions:

1. Is it feasible to implement the app into everyday practice?
2. Is TrackPAD suitable for recording patients’ daily and weekly SET performance?
3. Does the TrackPAD improve the prognosis of PAD and related QoL?

The primary outcome was defined as the change in the 6-minute walking distance using a standardized protocol at baseline and after 3 months of follow-up [17]. The 6-minute walk test was performed under the supervision of a trained exercise technician. Participants were instructed to cover as much distance as possible, walking up and down a 50-meter hallway for up to 6 minutes. Participants were asked to push a measuring wheel during the entire 6 minutes of the test, but they could take breaks if necessary. They were also allowed to use an assistive device during both walking tests if they so desired. The technician stood in the middle of the course and supervised the walking test, but they did not encourage participants. The total distance walked in the test was read off of the measuring wheel. In patients with heart failure and reduced ejection fraction, a decreased 6-minute walking distance was associated with increased mortality, nonfatal cardiovascular events, and heart failure–related hospitalizations [18-20]. A decreased 6-minute walking distance was associated with a predictive value of mortality in patients with chronic obstructive pulmonary disease [21]. Among patients with PAD, the baseline 6-minute walking distance predicts rates for all-cause mortality, CVD mortality, and mobility loss [22,23]. Additionally, an incline of 20 meters was linked to a considerable improvement in total walking ability [24].

Aside from the 6-minute walking distance being objective and well-validated with respect to walking ability predicting mobility loss and mortality in PAD, it has an excellent test-retest reliability [25,26]. The 6-minute walking test offers several advantages over treadmill testing in PAD as it correlates more closely with physical activity levels and is not associated with the learning effect of performing repeated tests [27].

The secondary outcome measures were changes in physical activity and assessing the patient’s PAD-related QoL via PAD-QoL. The PAD-QoL questionnaire is a validated PAD-specific questionnaire [28] containing five factors: (1) social relationships and interactions, (2) self-concept and feelings, (3) symptoms and limitations in physical functioning, (4) fear and uncertainty, and (5) positive adaptation. In addition, individual factors regarding sexual function, intimate relationships, and job function will also be assessed. An evaluation of the use of the TrackPAD app was also performed for the intervention group using the user version of the Mobile Application Rating Scale questionnaire. It provides a 20-item measure including 4 objective quality subscales for engagement, functionality, aesthetics, and information quality [29].
Study Design, Population, and the TrackPAD App

Study Design and Recruitment

This paper reports the results from the pilot study, including TrackPAD app usability tests for the target group (patients with PAD). In preparation for the pilot study, we conducted a recently published questionnaire study [15] evaluating the needs and requirements of designing mobile interventions for patients with PAD.

The TrackPAD pilot study was designed as a 2-armed randomized controlled trial and included patients with diagnosed and symptomatic PAD. It is a closed parallel-group trial (control and intervention groups were assessed simultaneously), with blinded assessors and face-to-face assessment components and a 3-month follow-up. Besides information regarding the pilot study, a call for participation was announced in a local newspaper (Westdeutsche Allgemeine Zeitung, local section for Essen and Duisburg) with the contact information provided, including the phone number and email address (trackPAD@uk-essen.de). In addition, potential participants were actively solicited during their outpatient clinic visits or their inpatient stay at the Department of Cardiology and Vascular Medicine, University Clinic of Essen. Willing patients were asked to register for the pilot trial at the front desk of the outpatient clinic.

Randomization

After screening based on inclusion and exclusion criteria and obtaining written informed consent, participants were randomized into 2 groups by the Center for Clinical Studies in Essen using the TENALEA software. The control group included participants with standard care and no further mobile intervention. The intervention group included participants receiving standard care and additional mHealth-based self-tracking of their physical activity using TrackPAD. The participants were stratified based on their 6-minute walking test (distances less than 362 meters, between 362 and 430 meters, and more than 430 meters) to ensure an even distribution of the walking speed between the two groups. After the randomization process, participants were not replaced, regardless of the reason for exclusion.

Both groups were strongly advised to continue with their SET according to the current standard guidelines [4]. Participants of the intervention group received additional access to the TrackPAD app, which complemented the patients’ current treatment. The TrackPAD app was freely accessible to the intervention group. Besides the support provided during the installation procedure, the app did not require further technical maintenance. The only external contact during the follow-up occurred if participants requested technical support. A nonphysician member of the study team helped participants.

The baseline and follow-up examinations took place at the Department of Cardiology and Vascular Medicine outpatient clinic. They included a 6-minute walking test and a measurement of the ankle-brachial index (ABI). The ABI Measurements were conducted using a Doppler probe on the tibial and anterior artery locations. According to the current European Society of Cardiology (ESC) guideline, the highest value was used for calculation and divided by the highest systolic brachial Doppler pressure [4].

The patients were asked to fill out a questionnaire package at both time points, including self-reported physical activity, demographic characteristics, and the PAD-QoL questionnaire. The PAD-QoL was translated into German by a native speaker and was pretested on 5 PAD patients not included in the study sample. The pretest did not reveal the need for any adjustments.

Inclusion Criteria

Main inclusion criteria were diagnosed and symptomatic PAD of the lower extremities, defined as Fontaine stage IIA or IIB. Fontaine stage IIA indicated intermittent claudication with a walking distance of more than 200 meters, whereas Fontaine stage IIB indicated intermittent claudication with a walking distance of fewer than 200 meters [4]. Additionally, patients must have a personal smartphone suitable for downloading and using the TrackPAD app (IOS version greater than 11.0 or Android version greater than 5.0). A detailed list of the inclusion and exclusion criteria is shown in Textbox 1.
Inclusion and exclusion criteria. ABI: ankle-brachial index; PAD: peripheral arterial disease; NYHA: New York Heart Association; CCS: Canadian Cardiovascular Society.

### Inclusion Criteria
- 18 years of age or older
- Diagnosis of lower extremity PAD based on either an ABI greater or equal to 0.9 in at least one leg, invasive or noninvasive imaging of stenotic lower extremity artery disease, or endovascular or surgical revascularization of lower extremity artery disease
- PAD Fontaine stage 1a/b
- Smartphone with the capacity to use TrackPAD (Android version greater than 5.0 or IOS version greater than 11.0)
- Written informed consent prior to any study procedures, including a specified follow-up evaluation
- Best-medical treatment in the last 2 months per standard guidelines

### Exclusion Criteria
- Wheelchair-bound, use of walking aid, or walking impairment due to another cause than PAD
- Below or above-knee amputation
- Acute or critical limb ischemia
- PAD Fontaine Stage I, III, or IV
- No German knowledge
- Severe cognitive dysfunction
- Congestive heart failure with NYHA III-IV symptoms
- Active congestive heart failure requiring the initiation or up-titration of diuretic therapy
- Angina pectoris with CCS class 3 to 4 symptoms, myocardial infarction, or stroke in the last 3 months
- Active arrhythmia requiring the initiation or up-titration of anti-arrhythmic therapy
- Severe valve disease

**TrackPAD App**

The mobile intervention TrackPAD was designed by Rocket Apes GmbH. There were no associations between the authors and the developer. Moreover, TrackPAD was only designed for study purposes and not commercial use. We did not change any content during the study period, and all content was frozen during the trial. The only dynamic component was the leaderboard, which was adjusted based on the training sessions performed by the participants. The participants set their weekly goal of SET units at the beginning of each week. As recommended by the 2017 ESC guideline [4], each unit included 30 minutes of SET. If participants did not go through an entire unit at once, there were 3 different options: taking breaks, continuing the unit after recovery, or quitting prematurely. After completing each unit (fully completed 30 min or not), user feedback was requested (Figure 1; see Feedback after SET unit).

Figure 1. Main views of the TrackPAD-app.
To account for a PAD-tailored solution, we included the following features (Figure 1):

1. **Weekly goal adaptation**: The app suggested a new weekly goal using an internal algorithm based on the completion rate of a user’s SET units during the previous week.
2. **Feedback after SET units**: The feedback after each SET unit contained PAD-specific information regarding leg pain levels, breathing, and overall exhaustion. Patients had to respond by choosing between 1 (minimum leg pain, no restriction in breathing, or minimum exhaustion) and 10 (maximum leg pain, maximum restriction in breathing, or maximum exhaustion) for each item.
3. **Claudication reminder**: Each SET unit started with a short reminder that the walking pace and incline must be adapted to reach a certain level of claudication to extend the PWD sustainably. The reminder popped up when each new SET unit was initiated and needed to be actively confirmed.
4. **Personal achievements**: The personal progress of each user was recorded to unlock achievement medals (e.g., a notable increase in users’ physical activity, activity performed during public holidays, or successes like an increase of performed SET units per week).
5. **Leaderboard**: The leaderboard contained different categories (e.g., number of steps in single training sessions, number of completed training sessions, total minutes of physical activity, and percent increase of physical activity). The different leaderboards showed individual placements compared to other users using TrackPAD.
6. **Patient events**: Information on upcoming Department of Cardiology and Vascular Medicine patient events focusing on vascular diseases were stored and easily accessible via the main menu.
7. **PAD-FAQ**: An FAQ section was included to address common technical issues, important contact information, and general training advice. Instructions in case of increasing or new pain during the training were also included.

**Ethics Approval and Consent to Participate**
The local ethics committee of the University of Duisburg-Essen (18-8355-BO) approved this study. Written informed consent was collected from each participant before any study procedures, and contact information was delivered to each participant. Any changes will be communicated to the ethics committee. The pilot study started at the beginning of November 2018 and ended in March 2019.

**Data Collection and Security**
Data were stored on an encrypted European server. No personalized data were shared with the developer, and they were only accessible to the study team.

**Sample Size Considerations and Statistical Analysis**
To allow for missing data and loss to follow up, we aimed to recruit 23 to 25 participants per study arm. The results achieved an estimated power of $t=0.46$ (post hoc power analysis; Cohen $d=0.5$; $P=.05$; $F_{1,46}=1.157$). We used a two-tailed $t$-test, and the enrollment goal was 20 participants each for the intervention and control groups. $P<.05$ was estimated as the significance threshold. For sample size consideration and statistical analysis, we used R (version 3.6.0). To account for the heterogeneity of the walking distance to be covered, the analysis was performed separately for Fontaine stage IIa and IIb. The regression model was estimated by ordinary least squares and a differences-in-differences approach.

**Results**

**Study Population and Baseline Characteristics**
After screening and randomization, we included 46 participants in the pilot study, of whom 22 (48%) were randomized to the intervention group, and 24 (52%) were randomized to the control. During the follow-up, 7 (15%) participants dropped out, mainly due to personal reasons. For example, 5 (11%) participants withdrew due to the severe illness of a close relative, and 2 (4%) participants dropped out as a result of either worsening of a nonstudy-related disease or death (Figure 2; see Panel A). Table 1 shows a summary of the remaining participants’ baseline characteristics.
Figure 2. Quantitative development of screened patients including reasons for dropouts and exclusions.
Table 1. Patient characteristics at baseline.

<table>
<thead>
<tr>
<th></th>
<th>Intervention group (n=19)</th>
<th>Control group (n=20)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>64.6 (9.8)</td>
<td>65.6 (7.7)</td>
<td>.72</td>
</tr>
<tr>
<td>Sex (male), n (%)</td>
<td>12 (63)</td>
<td>9 (45)</td>
<td>.34</td>
</tr>
<tr>
<td>Obesity (BMI &gt; 30 kg/m²), n (%)</td>
<td>5 (26)</td>
<td>1 (5)</td>
<td>.16</td>
</tr>
<tr>
<td>Prior MIa, n (%)</td>
<td>2 (11)</td>
<td>3 (15)</td>
<td>.85</td>
</tr>
<tr>
<td>Hypertension, n (%)</td>
<td>12 (63)</td>
<td>16 (80)</td>
<td>.41</td>
</tr>
<tr>
<td>Diabetes, n (%)</td>
<td>4 (21)</td>
<td>6 (30)</td>
<td>.21</td>
</tr>
<tr>
<td>Hyperlipidemia, n (%)</td>
<td>12 (63)</td>
<td>13 (65)</td>
<td>.42</td>
</tr>
<tr>
<td>Previous peripheral intervention, n (%)</td>
<td>8 (42)</td>
<td>5 (25)</td>
<td>.26</td>
</tr>
<tr>
<td>Previous peripheral bypass graft, n (%)</td>
<td>3 (16)</td>
<td>5 (25)</td>
<td>.68</td>
</tr>
<tr>
<td>Previous PCIb, n (%)</td>
<td>4 (21)</td>
<td>6 (30)</td>
<td>.69</td>
</tr>
<tr>
<td>Heart failure, n (%)</td>
<td>2 (11)</td>
<td>3 (15)</td>
<td>.85</td>
</tr>
<tr>
<td>Coronary arterial disease, n (%)</td>
<td>6 (32)</td>
<td>9 (45)</td>
<td>.51</td>
</tr>
<tr>
<td>Active/Former smoker, n (%)</td>
<td>6/11 (32/58)</td>
<td>8/10 (40/50)</td>
<td>.89</td>
</tr>
<tr>
<td>Fontaine stage IIa, n (%)</td>
<td>12 (63)</td>
<td>14 (70)</td>
<td>.44</td>
</tr>
<tr>
<td>Fontaine stage IIb, n (%)</td>
<td>7 (37)</td>
<td>6 (30)</td>
<td>.85</td>
</tr>
<tr>
<td>6-minutes walking distance (meters), mean (SD)</td>
<td>407 (80.8)</td>
<td>390.1 (66)</td>
<td>.35</td>
</tr>
<tr>
<td>ABIc</td>
<td>0.75 (0.21)</td>
<td>0.73 (0.18)</td>
<td>.46</td>
</tr>
<tr>
<td>Reported physical activity (days per week), mean (SD)</td>
<td>2.4 (1.4)</td>
<td>2.3 (1.9)</td>
<td>.35</td>
</tr>
</tbody>
</table>

aMI, myocardial injury.
bPCI, percutaneous coronary intervention.
cABI, ankle-brachial index.

Increase in the 6-minute Walking Distance as a Primary Outcome

Of the 20 participants who increased their 6-minute walking distance at follow-up, 18 (90%) belonged to the intervention group using TrackPAD. The remaining participant in the intervention group did not change his covered distance at follow-up. In contrast, except for 2 (10%) participants, 18 (90%) participants in the control group showed decreased 6-minute walking distance at follow-up.

The mean distance covered in the 6-minute walking test showed a significant increase in the intervention group overall (83.0 meters, SD 72.2), whereas the mean walking distance of the control group decreased on average (–38.8 meters, SD 53.7; P<.001).

Both Fontaine stages showed similar trends, but the mean distance increase for the less progressed Fontaine stage IIa was more pronounced (intervention group: 97.0 meters, SD 78.6 vs. the control group: –35.3 meters, SD 55.9; P<.001). The Fontaine stage IIb showed a slight increase in mean walking distance for the intervention group (59.0 meters, SD 57.0) compared to the control group (–7.0 meters, SD 52.2), but it was still significant (P=.01).

TrackPAD was linked to a mean increase in the 6-minute walking distance of the intervention group, regardless of the Fontaine stage (95% CI 48.2-117.8). In contrast, the control showed either a slight or missing increase (95% CI –63.9-3.6). In total, the difference between both means was 121.8 meters (Fontaine stage IIa: 132.3 meters; IIb: 106.4 meters). Depending on the Fontaine stage, this resulted in a 17% (IIb) to 23% (IIa) increase of the covered distance at follow-up (Table 2).
### Table 2. Differences in the 6-minute walking distance within and between study and control group after 3 months of follow-up.

<table>
<thead>
<tr>
<th></th>
<th>Fontaine IIa (n=26)</th>
<th>Control (n=14)</th>
<th>Fontaine IIb (n=13)</th>
<th>Control (n=6)</th>
<th>Fontaine IIa, IIb (n=39)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference in mean&lt;sup&gt;a&lt;/sup&gt; (meters)</td>
<td>97.0</td>
<td>-35.3</td>
<td>59.4</td>
<td>-47.0</td>
<td>83.0</td>
</tr>
<tr>
<td>Median (meters)</td>
<td>89.9</td>
<td>-22.0</td>
<td>30.0</td>
<td>-22.5</td>
<td>60.0</td>
</tr>
<tr>
<td>SD (meters)</td>
<td>78.6</td>
<td>55.9</td>
<td>57.0</td>
<td>52.2</td>
<td>72.2</td>
</tr>
<tr>
<td>95% CI&lt;sup&gt;b&lt;/sup&gt; (meters)</td>
<td>47.0-147.0</td>
<td>-67.5-3.0</td>
<td>6.3-111.8</td>
<td>-101.8-7.8</td>
<td>48.2-117.8</td>
</tr>
<tr>
<td>Difference in mean between both groups (meters)</td>
<td>132.3</td>
<td>106.4</td>
<td>121.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD (meters)</td>
<td>135.5</td>
<td>71.6</td>
<td>176.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95% - CI&lt;sup&gt;c&lt;/sup&gt; (meters)</td>
<td>75.5-189.0</td>
<td>39.2-172.8</td>
<td>80.2-163.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P value</td>
<td>.01</td>
<td>.01</td>
<td>.01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Positive mean indicates an improvement.

<sup>b</sup> Difference between study and control group of the sub group.

<sup>c</sup> The true difference of the population between both groups.

A difference-in-difference regression with fixed effects for time (accounting for a progression of PAD) and individual participant (accounting for unobserved heterogeneity between the participants) estimating the percentage change in the treatment effect showed that the effect of receiving access to TrackPAD increased the 6-minutes walking distance about 28% (SE 0.04). This effect was significant to a confidence level of 99%.

**PAD-related Quality of Life**

The PAD-related quality of life (PAD-QoL) was assessed by the PAD-QoL questionnaire at baseline and follow-up. No relevant differences were observed at baseline between both groups. However, at follow-up, significant changes were noted in 3 factors of the PAD-QoL, with the most extensive change evident in the “symptoms and limitations in physical functioning.” The intervention group reported reduced limitations in their daily activity: “I have had to greatly reduce my activities because of my PAD” (Q1, intervention group: −1.6 meters, SD 1.4 vs control group: −0.1 meters, SD 1.0; P=.01); “I cannot do many of the things I enjoy because of my PAD” (Q3, intervention group: −1.8 meters, SD 1.5 vs control group: −0.4 meters, SD 1.0; P=.01); and “My legs hurt a lot when I walk because of my PAD” (Q4, intervention group: −1.4 meters, SD 1.3 vs control group: 0 meters, SD 1.1; P=.01). The intervention group also showed a change in a single item of the factor “fear and uncertainty,” reporting a reduced fear of losing life because of PAD: “I am afraid of losing my life as a result of my PAD” (Q8, intervention group: −1.3 meters, SD 1.5 vs control group: −0.3 meters, SD 1.5; P=.048), and a change in the section “positive adaptation”: “I feel very hopeful about the outcome of my PAD” (Q14, intervention group: 1.2 meters, SD 1.1 vs control group: 0.2 meters, SD 1.4; P=.02; Figure 3).
Overall, changes in the PAD-QoL over the 3 months of follow-up showed a less intense subjective symptom perception and fewer limitations in daily life among the intervention group.

Reported Physical Activity

To compare the two groups in terms of physical endurance at baseline, we recorded the reported physical activity. Both groups did not differ in days of physical activity per week (intervention group: 2.9 days per week, SD 2.8 vs control group: 2.4 days per week, SD 1.9; \( P=.44 \)). Both groups had participants who were active for a median of 30 to 60 minutes (intervention group: n=9, 20% vs control group: n=4, 9%). In total, 12 participants (26%) were active for more than 60 minutes (intervention group: n=3, 7% vs control group: n=9, 20%). Among participants who exercised for less than 30 minutes weekly, 5 (11%) participants trained between 10 and 30 minutes weekly (intervention group: n=1, 2% vs control group: n=4, 9%), and 9 (20%) participants exercised less than 10 minutes weekly (intervention group: n=6, 13% vs control group: n=3, 7%).

At follow-up, 37 (80%) participants reported an increase in their weekly physical activity (intervention group: n=15, 33% vs control group: n=16, 35%), resulting in a comparable rise in physically active days per week in both groups (intervention group: plus 0.3 days per week, SD 3.5 vs control group: plus 0.4 days per week, SD 2.6; \( P=.93 \)).

App Evaluation

App Usage

We considered intervention participants as active users if they performed at least 1 weekly training. During week 1, every participant was active. A dip from 19 (100%) to 14 (74%) active users was observed in week 2, increasing to 17 (89%) active users in week 3. During the following weeks, the activity remained stable, with 14 to 15 (70% to 75%, respectively) active users from week 5 to 12 (Table 3). During the 12 weeks of follow-up, the number of training sessions per week stayed roughly the same for the participants that remained active users.
Table 3. TrackPAD-app usage of the intervention group during the 12 weeks of follow-up.

<table>
<thead>
<tr>
<th>Week</th>
<th>Total training sessions (units)</th>
<th>Active user (n)</th>
<th>Intervals per training session (units), mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
<td>19</td>
<td>2.5 (2.8)</td>
</tr>
<tr>
<td>2</td>
<td>48</td>
<td>14</td>
<td>2.8 (2.2)</td>
</tr>
<tr>
<td>3</td>
<td>53</td>
<td>17</td>
<td>2.6 (1.7)</td>
</tr>
<tr>
<td>4</td>
<td>74</td>
<td>16</td>
<td>2.0 (1.2)</td>
</tr>
<tr>
<td>5</td>
<td>60</td>
<td>15</td>
<td>1.8 (1.2)</td>
</tr>
<tr>
<td>6</td>
<td>61</td>
<td>15</td>
<td>2.3 (1.6)</td>
</tr>
<tr>
<td>7</td>
<td>49</td>
<td>15</td>
<td>2.7 (2.8)</td>
</tr>
<tr>
<td>8</td>
<td>57</td>
<td>14</td>
<td>2.4 (2.0)</td>
</tr>
<tr>
<td>9</td>
<td>58</td>
<td>15</td>
<td>2.2 (2.3)</td>
</tr>
<tr>
<td>10</td>
<td>51</td>
<td>14</td>
<td>2.1 (1.9)</td>
</tr>
<tr>
<td>11</td>
<td>55</td>
<td>14</td>
<td>2.5 (2.1)</td>
</tr>
<tr>
<td>12</td>
<td>48</td>
<td>15</td>
<td>2.7 (4.1)</td>
</tr>
</tbody>
</table>

\( ^a \)Total number of recorded training sessions for the respective week as assessed by the TrackPAD-App.

\( ^b \)Active user with at least one training interval in the corresponding week.

\( ^c \)Mean number of intervals during one training session. Each training session could be paused if necessary, resulting in each training session being subdivided into several intervals.

The reasons for discontinued TrackPAD use by the nonactive users (n=5, 26%) from week 5 onward were assessed at follow-up. Reasons for discontinued TrackPAD use were related to personal circumstances (n=3, 16%) and technical issues (n=2, 11%). Of the 2 participants who stopped using TrackPAD for personal reasons, 1 was due to the illness of a close relative, and the other lost interest. One participant stopped using TrackPAD due to reported interference between the TrackPAD app and their Samsung Health app, and another participant stopped the training sessions due to several sequential app crashes (Figure 2; see Panel B).

User Feedback

The vast number of questions regarding functionality, aesthetics, and informational content of TrackPAD were reported as positive to extremely positive (4 or 5 stars out of 5; Figure 4; see Panel A). However, the visual information provided within the app showed potential for improvement (Figure 4; see Panel A, Item 15); for example, the plausibility and correctness of descriptions represented by pictograms or pictures. Participants described this item mainly as “largely unclear.” Only 5 (25%) participants described the visual information as “mostly clear” (n=4) or “absolutely clear” (n=1).

The users’ feedback also included questions regarding the perceived impact of the TrackPAD with respect to their PAD disease (Figure 4; see Panel B). Only 1 (6%) user disagreed, stating the app had not changed their awareness of SET (Q1). The other participants reported that the app had significantly increased their motivation to perform SET (Q4) and their compliance to SET (Q6). They also stated that using the app changed their attitude regarding SET (Q3) and increased their knowledge about SET (Q2).

Most users evaluated the app in all of the 3 categories positively. Only 3 (17%) users would “maybe” or are “unlikely” to recommend the app to people with existing PAD disease. Supporting the positive evaluation illustrate in Panel A (Figure 5), 13 (68%) users rated the app with at least 4 out of 5 stars (Figure 5; see Panel B). Future app use, at least every week, was reported by 10 (53%) users.
Discussion

The implementation of novel technologies, specifically mobile interventions, can substantially change the landscape for the treatment of CVD [12,30]. General benefits of mHealth technologies include the wide reachability and the possibility of continuous access [31]. Although PAD represents a subgroup of CVD, patient characteristics and disease-specific requirements differ substantially from those patients with other CVD. Therefore, disease- and patient-tailored solutions are essential to the development of mobile interventions. One significant difference between the PAD population and patients with other CVD is the older age and the fact that the patient-centered development process needs to be expanded by one additional dimension. Previous studies already explored the use and
acceptability of mobile technologies in health care related to
the users’ age and identified age as an important factor in the
design of mobile interventions, requiring greater technical
support and reporting lower acceptability of using mobile
technologies [32,33]. As such, there are measurable influences
on intermediate outcomes (eg, increased satisfaction with care)
and health outcomes (eg, better metabolic control) [15,34–36].

SET is one of the most relevant interventions in the conservative
treatment of PAD, but barriers to exercise are still high. Besides
low motivational aspects, intermittent claudication limits the
sustainability of regular SET performance. Moreover, the
requirements of primary care for patients with PAD focus on
other priorities other than CVD in general [15,30]. To meet this
specific patient population’s needs and requirements, we
proposed using a PAD-tailored mobile intervention to encourage
the SET performance in patients with PAD.

Mobile technologies are increasingly used for health purposes,
even among older adults who have demonstrated a lower uptake
of technologies compared to younger people [37]. Although
these technologies have the potential to assist in care
coordination activities, like regular SET performance, most
mobile apps are not designed specifically for this population
which has complex health care needs and is older than the
typical app user. The activity recognition mechanism of most
mobile apps cannot accommodate the wide range of human
movement linked to mobility impairment.

In this study, we gathered TrackPAD use input from the patients’
perspective, and we observed a high level of user acceptance.
Overall, we found satisfaction in terms of functionality,
aesthetics, and informational content. Studies combining eHealth
and PAD are rare, but the same trend of mobile technology user
acceptance was observed in patients with noncommunicable
diseases. A review of eHealth interventions for cancer survivors
showed mobile interventions are promising tools [38]. Future
work will need to examine the extent to which personalized
activity recognition can support the diversity of movement.

Improvement was demonstrated through the visual information
within TrackPAD and the clear assignment of pictograms or
pictures. The weakness of the gestural concept resulted from
the advanced age of the user group, which is often inexperienced
in using mHealth and requires an age-adapted presentation [39].
Besides relevant barriers for older adults, lack of desire, costs,
privacy and security considerations, visual acuity, and hand-eye
coordination were important factors with respect to the
acceptance of telehealth interventions [40]. These barriers
will be adapted accordingly to improve the TrackPAD app following
a patient-centered approach.

Since we designed a platform for both iOS and Android, some
technical issues occurred due to the different technical
implementations of the provider. The various mechanisms for
counting steps presented a considerable challenge in designing
a comparable app for both platforms. Depending on the
manufacturer, step counts work either over a physical hardware
mechanism and a software-based solution. This issue might
become less relevant when it comes to personal use [41,42],
but it also limits the analysis within a clinical trial. Because of
the low number of smartphones that use the software-based
solution, this issue did not occur in earlier tests. However, the
disproportionate share of older mobile phones lacking a physical
hardware mechanism within the intervention group revived the
issue. In further trials, the inclusion of newer operating versions
of Android mobile phones should be considered since this issue
was only found in Android-based operating systems.

The disadvantages of simple activity tracking are known and
common limitations in studies. The performance of systems
trained with data in the laboratory setting substantially
deteriorates when tested in real-life conditions [43]. Possible
solutions might be user-calibration processes or the use of
specified study-related devices to gain comparability.

Comparing the 6-minute walking distance between both groups
in our study, we saw a significant increase in the mean walking
distance of 80 meters in the intervention group using TrackPAD.
Remarkably, we did not find any decrease in the walking
distance within the intervention group, whereas 90% (n=18) of
the control group did worse at follow-up compared to baseline.
One reason for the longer walking distance might be because
of the younger age of the study participants. Previous studies
reported a mean age of more than 70 years, whereas the
intervention group using TrackPAD had a mean age of 64. The
higher increase may also be due to comparatively minor
restrictions since two-thirds of the participants were classified
as Fontaine stage Ia (PWD of more than 200 meters). The ease
of initiating exercise among the Fontaine stage Ia patients with
PDA compared to patients with higher Fontaine stages might
be linked to better endurance during exercise and higher
motivation in general. Moreover, the small sample size allows
for substantial individual changes within the intervention group,
leading to an upward deviation.

Although the covered distance in the 6-minute walk test only
increased significantly in the intervention group, the
self-reported physical activity increased in both groups at
follow-up. An accurate assessment of physical activity using
the PDA-QoL questionnaire seems questionable in the entire
study population and has previously been described as a
common issue [44,45]. Digital interventions also increase the
potential to track background activity (ie, receiving objective
statements in terms of physical activity) and will be considered
in future trials. The recording of activity highs and lows
throughout the day might also help identify optimal time points
to send digital motivation notifications. It is important to note
that messages can also decrease productivity if delivered at the
wrong time points. Algorithms based on collected personalized
information in smartphones might reduce the number of
poorly-timed interruptions [46].

We also observed an increase in PAD-QoL regarding
“symptoms and limitations in physical functioning” within the
intervention group. The association between increased physical
activity and an increased PAD-QoL has been reported in other
studies [47–49]. In addition, the increased 6-minute walking
distance was linked to better physical aspects of quality of life
in participants with intermittent claudication, supporting its
value as an outcome measure.

The main limitation of this study was the small sample size of
the intervention group. Since we have analyzed some patient
characteristics (ie, Fontaine stage IIa and IIb) separately, the sample size per group decreased even further. However, the Fontaine stage allowed us to control for the differences in the participants’ physical capability. Although we saw a relevant change in the primary outcome variable after follow-up, recordings of background activity during the follow-up period were available due to privacy restrictions. Based on the study design of this pilot, no blinding of the study participants was feasible, and motivational differences must be considered. Further research is needed to address this issue.

Using the smartphone–based tool TrackPAD, we found a significant increase in the mean 6-minute walking distance at follow-up, indicating a prognostically relevant change in walking ability in patients with moderate PAD. TrackPAD also bolstered a shift in the subjective symptom perception and fewer noticed limitations in terms of PAD-QoL. Thus, the TrackPAD app seems feasible and suitable for the target group of patients with PAD in terms of SET performance. Participants substantially valued the experience of using an app in the management of their care. Still, a further adaption of the visual presentation and the gestural concept that follows a patient-centered approach is needed.

Acknowledgments
This work was supported by the Stiftung Universitätsmedizin (D/106-21637) and the Deutsche Forschungsgesellschaft (DFG; DFG 969RA/12-1).

Conflicts of Interest
None declared.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1). [PDF File (Adobe PDF File), 381 KB - mhealth_v9i8e24214_app1.pdf ]

References


Abbreviations

ABI: ankle-brachial index
CAD: coronary artery disease
ESC: European Society of Cardiology
mHealth: mobile health
PAD: peripheral arterial disease
PAD-QoL: peripheral arterial disease–related quality of life
CVD: cardiovascular disease
PWD: pain-free walking distance
QoL: quality of life
SET: supervised exercise training
Effect of Physician-Pharmacist Participation in the Management of Ambulatory Cancer Pain Through a Digital Health Platform: Randomized Controlled Trial

Lu Zhang\textsuperscript{1,2,}*, MD; Howard L McLeod\textsuperscript{3,4,5,6,}*, Prof Dr; Ke-Ke Liu\textsuperscript{1,2}, MD; Wen-Hui Liu\textsuperscript{7}, Prof Dr; Hang-Xing Huang\textsuperscript{1,2}, MD; Ya-Min Huang\textsuperscript{1,2}, MD; Shu-Sen Sun\textsuperscript{1,2,8}, Prof Dr; Xiao-Ping Chen\textsuperscript{2,5,6}, Prof Dr; Yao Chen\textsuperscript{8}, Prof Dr; Fang-Zhou Liu\textsuperscript{9}, Prof Dr; Xiao Jian\textsuperscript{1,2}, Prof Dr

\textsuperscript{1}Department of Pharmacy, Xiangya Hospital, Central South University, Changsha, China
\textsuperscript{2}Institute for Rational and Safe Medication Practices, National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha, China
\textsuperscript{3}Geriatric Oncology Consortium, Tampa, FL, United States
\textsuperscript{4}Taneja College of Pharmacy, University of South Florida, Tampa, FL, United States
\textsuperscript{5}Department of Clinical Pharmacology, Xiangya Hospital, Central South University, Changsha, China
\textsuperscript{6}Hunan Key Laboratory of Pharmacogenetics, Institute of Clinical Pharmacology, Central South University, Changsha, China
\textsuperscript{7}Department of Pharmacy, The Second Xiangya Hospital, Central South University, Changsha, China
\textsuperscript{8}College of Pharmacy and Health Sciences, Western New England University, Boston, MA, United States
\textsuperscript{9}College of Information Science and Engineering, Hunan Normal University, Changsha, China

*these authors contributed equally

Corresponding Author:
Xiao Jian, Prof Dr
Department of Pharmacy
Xiangya Hospital
Central South University
Room 404
No. 87 Xiangya Road, Kaifu District
Changsha, 410008
China
Phone: 86 13973168895
Fax: 86 073184327460
Email: admanoa@163.com

Abstract

Background: Self-management of ambulatory cancer pain is full of challenges. Motivated by the need for better pain management, we developed a WeChat-supported platform, Medication Housekeeper (MediHK), to enhance communication, optimize outcomes, and promote self-management in the home setting.

Objective: We conducted a randomized controlled trial to assess whether the joint physician-pharmacist team through MediHK would provide better self-management of ambulatory patients with cancer pain.

Methods: Patients were randomly assigned to either an intervention group or control group. During the 4-week study period, the pharmacist would send 24-hour pain diaries daily, adverse drug reaction (ADR) forms every 3 days, and the Brief Pain Inventory form every 15 days to patients in the intervention group via MediHK. If a patient needed a change in drug/dosage or treatment of an ADR after the comprehensive review, the pharmacist would propose pharmacological interventions to the attending physician, who was then responsible for prescribing or adjusting pain medications. If no adjustments were needed, the pharmacist provided appropriate targeted education based on knowledge deficits. Patients in the control group received conventional care and did not receive reminders to fill out the forms. However, if the control group patients filled out a form via MediHK, the pain management team would review and respond in the same way as for the intervention group. The primary outcomes included pain intensity and pain interference in daily life. Secondary outcomes included patient-reported outcome measures, medication adherence, ADRs, and rehospitalization rates.
Results: A total of 100 patients were included, with 51 (51%) in the intervention group and 49 (49%) in the control group. The worst pain scores, least pain scores, and average pain scores in the intervention group and the control group were statistically different, with median values of 4 (IQR 3-7) vs 7 (IQR 6-8; P = .001), 1 (IQR 0-2) vs 2 (IQR 1-3; P = .02), and 2 (IQR 2-4) vs 4 (IQR 3-5; P = .001), respectively, at the end of the study. The pain interference on patients’ general activity, mood, relationships with others, and interests was reduced, but the difference was not statistically significant compared with the control group (P = .10-.76). The medication adherence rate increased from 43% to 63% in the intervention group, compared with an increase of 33% to 51% in the control group (P < .001). The overall number of ADRs increased at 4 weeks, and more ADRs were monitored in the intervention group (P = .003). Rehospitalization rates were similar between the 2 groups.

Conclusions: The joint physician-pharmacist team operating through MediHK improved pain management. This study supports the feasibility of integrating the internet into the self-management of cancer pain.

Trial Registration: Chinese Clinical Trial Registry ChiCTR1900023075; https://www.chictr.org.cn/showproj.aspx?proj=36901

(JMIR MHealth Uhealth 2021;9(8):e24555) doi:10.2196/24555

KEYWORDS

cancer pain; self-management; ambulatory setting; digital health; physician-pharmacist

Introduction

Pain is a common and challenging issue for cancer patients, affecting most at some stage of their disease [1]. A meta-analysis indicated that pain prevalence was 33% in patients after curative treatment, 64% in patients with advanced disease, and 59% in patients on anticancer therapy; approximately 35% of patients reported their pain as moderate to severe [2]. Inadequate pain management continues, with approximately one-third of cancer pain patients undertreated [3]. According to recent surveys, cancer pain management in China remains far from the ideal goal [4]. The barriers are multifactorial, including knowledge deficits, inadequate pain assessment, and misconceptions of pain from both patients and professionals [5]. Managing ambulatory patients is especially tough because of the complex environment, limited communication with health care providers, and difficulty managing their pain-medications regimens [6,7].

Both the World Health Organization and the European Society for Medical Oncology recommended that cancer pain patients should be active in their self-management of their pain. Patient-reported outcomes are increasingly used in routine ambulatory cancer care to guide clinical decisions, enhance communication, and improve symptom management [8]. Electronic patient-reported outcomes, supported by computer-adaptive testing technology, have shown potential in the era of big data [9]. Smartphones and apps such as WeChat (the largest social networking app in China), provide additional value to obtain knowledge and information, as well as making it possible for patients and health care providers to communicate electronically. Most patients are willing to self-report their symptoms via digital health apps. Several studies have reported on applications based on the eHealth model for the self-management of cancer pain [10,11].

This study established a physician-pharmacist collaboration team that participated in the self-management of ambulatory patients with cancer pain through a WeChat-supported platform: Medication Housekeeper (MediHK). We aimed to assess whether the joint physician-pharmacist team operating through MediHK would provide better self-management of ambulatory patients with cancer pain and optimize therapeutic outcomes.

Methods

WeChat-Supported Platform: MediHK Design

Patients were managed by MediHK, a WeChat-supported platform designed by the research team. An engineer from Hunan Normal University’s College of Information Science and Engineering provided technology support for building MediHK. MediHK has been patented by the National Intellectual Property Administration, People’s Republic of China (ZL 2015 1 0648320.2). MediHK contains 2 opening screens: (1) the patient interface (Multimedia Appendix 1) and (2) the medical interface (Multimedia Appendix 2). The former is for patients, and the latter is for the members of the pain management team, which consisted of physicians and pharmacists. The medical interface was designed to manage pain-related problems and provide consultation to patients in a timely fashion. Both interfaces included 3 modules: a user login module for inserting basic user demographics into MediHK; an e-consultation module for communicating between patients and medical providers; and an introductory module for MediHK education, which offers a quick response code for new users (Multimedia Appendix 3).

We conducted 3 rounds of consensus-building using Delphi methods [12] to determine patient-reported outcome measures (PROMs) that could be integrated into MediHK. The pharmacists sent messages to patients, as shown on the far-right side of Multimedia Appendix 2, and the patients would receive a reminder, as shown on the far-right side of Multimedia Appendix 1. Patients could consult on any questions about pain, and the pain management team would receive real-time WeChat messages and respond as soon as possible. The acceptable response time was generally within 2 hours. When patients needed a change in drug/dosage or treatment of an adverse drug reaction (ADR), pharmacists first reviewed the patients’ historical records through MediHK and, then, made recommendations and reminded the physician. In China, pharmacists have no right to prescribe. If the physician had conflicting opinions, an agreement would be reached through offline contact; then, the physician could adjust drug-therapy regimens. All treatment recommendations from the pharmacists and physicians were based on the guidelines of the European Society for Medical Oncology Standard diagnosis and treatment...
of cancer pain, the National Comprehensive Cancer Network, and the European Association for Palliative Care.

When the patient’s expression was not clear, there was no guarantee that the same physician or pharmacist would communicate back. However, since the previously submitted forms and consultation questions were available through MediHK, the new physician or pharmacist would review the patient’s history in all aspects. Patient information was protected by encryption. Note that all screenshots of the app include translations that have been added for clarity for this paper.

PROMs and Forms That Integrated Into the MediHK

The Brief Pain Inventory
This study used the Brief Pain Inventory (BPI; Chinese version) to assess cancer pain, and it has been widely used for its good construct and concurrent validity [13]. It provides 2 main scores, which are “pain intensity” and “pain interference in daily life.” Pain intensity is based on the Numerical Rating Scale (NRS) and includes a 4-item sensory dimension: worst pain, least pain, average pain, and present pain. Each item is rated from 0 (“no pain”) to 10 (“very severe pain”). The “pain interference on daily life” score is a 7-item reactive dimension that includes general activity, mood, walking ability, daily work, relationships, sleep, and enjoyment of life; each item is rated from 0 (“no interference on daily life”) to 10 (“complete interference”).

Medication History and Adherence
We designed a list to record the medication history of ambulatory patients with cancer pain (Multimedia Appendix 4). The Morisky Medication Adherence Measure [14] was used to assess adherence to analgesics because of its excellent reliability and validity in the Chinese cancer pain population [15]. The Morisky Medication Adherence Measure focuses on the following medication-taking behaviors (Multimedia Appendix 5): forgetfulness, carelessness, and cessation of the drug regimen when feeling better or worse. The answers of “yes” or “no” for each item scored 0 and 1, respectively. The scores were divided into 3 categories: complete adherence (4), incomplete adherence (1-3), and nonadherence (0).

ADR Form
Many patients treated with opioids may experience adverse events. To comprehensively capture ADRs of patients outside of the hospital, we designed a form (Multimedia Appendix 6) with World Health Organization terminology for ADRs and classified ADRs into several symptoms.

Pain Diary for 24-Hour Pain
We designed a pain diary to capture patients’ daily pain in the home setting over time. The pain diary included 5 modules (Multimedia Appendix 7): (1) a pain score record, which combined the NRS, Face Pain Scale, and Verbal Rating Scale to assess pain accurately; (2) a form that included the time of medication when feeling better or worse. The answers of “yes” or “no” for each item scored 0 and 1, respectively. The scores were divided into 3 categories: complete adherence (4), incomplete adherence (1-3), and nonadherence (0).

Study Design and Participants

Study Overview
This was a 2-arm, randomized controlled clinical trial, and the study has been registered at ChiCTR.org ChiCTR1900023075; https://www.chictr.org.cn/showproj.aspx?proj=36901. Ambulatory patients with cancer pain in a tertiary hospital were included and assigned to either a control group (ie, joined in the MediHK only, no physician-pharmacist active intervention) or an intervention group (ie, MediHK platform plus physician-pharmacist intervention), with an allocation ratio of 1:1 using a random number table. Our pre-experiment included 72 patients who met the criteria for inclusion. The preliminary results showed that the average NRS of patients in the control group was 5.85 (SD 2.442). The average score of patients in the intervention group was expected to be <4. Assuming a type I error of 5%, a type II error of 20%, and considering the design of similar sample content, the sample size required for each group was calculated to be 37 patients. Allowing for 20% attrition, 100 patients (50 participants per group) were planned to be enrolled. The study was conducted in accordance with the Declaration of Helsinki and was approved by the ethics committee of Xiangya Hospital of Central South University (approval number 2017121139). All participants signed an informed consent form before participation.

Care of Patients in the Intervention Group
The pharmacist would send daily 24-hour pain diaries, ADR forms every 3 days, and the BPI form every 15 days via MediHK. The pharmacist would first review patient demographic information, check the form regarding pain intensity and interference in daily life, conduct medication therapy reviews, and review ADRs and medication adherence. If the patient needed a change in drug/dosage or treatment of an ADR after the comprehensive review, the pharmacist would propose pharmacological interventions to the attending physician. The physician was responsible for prescribing or adjusting pain medications. If no changes were needed, the pharmacist provided appropriate targeted education based on patient knowledge deficits.

Care of Patients in the Control Group
Patients in the control group received conventional care. Before the patient was discharged from the hospital, the pharmacist conducted detailed medication education. However, the control group patients did not receive a reminder to fill out the forms. If they filled out the form via MediHK, the pain management team would also review and respond the same way as for the intervention group.

Inclusion and Exclusion Criteria
The inclusion criteria of participants included: (1) age ≥18 years; (2) diagnosis of malignant tumors by a pathological or cytological method; (3) pain that met the cancer pain diagnostic criteria according to National Comprehensive Cancer Network Guidelines and was moderate to severe (NRS ≥4); (4) ability of patients or their families to read Chinese and use WeChat;
(5) a normal verbal ability and performance status; and (6) agreement to participate in the study and sign the informed consent form.

Exclusion criteria of participants were: (1) nonmalignant pain; (2) hepatic dysfunction (alanine aminotransferase $\geq 2.5 \times$ upper limits of normal [ULN], aspartate aminotransferase $\geq 2.5 \times$ ULN, or total bilirubin $\geq 1.5 \times$ ULN) or renal dysfunction (serum creatinine $\geq 2.5 \times$ ULN); (3) participation in other clinical trials; and (4) hospitalization during the 4-week trial period.

**Patient Enrollment and Intervention**

**Patient Enrollment**

At the patients' first visit to the ambulatory clinic, the physician provided a detailed consultation and, then, determined a treatment plan after discussion with the pharmacist; an account was created for eligible patients. After registration, the pharmacist demonstrated the use of each MediHK module to patients, including what information was collected in each form and how to fill it out and how to switch the interface to send a form or question. Even though the operation of MediHK was simple enough, the training process was approximately 10 minutes. The specific time depended on patient understanding, ability, and proficiency in WeChat. After training, patients were assigned to a pain management team and were required to complete PROMs and forms. Patients at home could contact their pain management team at any time through MediHK if they had any trouble with pain. The pain management team was required to complete standardized clinical pain management training and had at least 10 years of hospital work experience for clinical pharmaceuticals for cancer pain before performing pain-management work.

Patients were observed for 4 weeks. At week 4, the patients were required to complete and submit the PROMs through MediHK or report through phone calls within 1 day. Patients could continue to use MediHK after the completion of the study.

**Outcome Evaluation**

The primary outcomes included pain intensity and pain interference in daily life. Secondary outcomes included PROMs, medication adherence, ADRs, and rehospitalization rates.

**Statistical Analysis**

All data were analyzed using SPSS software (version 22.0; IBM Corp), and all charts were made by the graphing software GraphPad (version 8.0.2 (263); GraphPad Prism). For measurement data, the normality test adopted the Kolmogorov-Smirnov method. If normally distributed, the data were expressed as mean (SD), and the comparison between the 2 groups used 2 independent sample t tests. If not normally distributed, the data were expressed as the median (IQR), and the comparison between groups underwent a Mann-Whitney U test. Counting data were expressed as a frequency and percentage. A chi-square ($c^2$) test or Fisher exact test was used for comparison between groups. We screened for factors affecting the pain intensity of outpatients with cancer pain by multivariate linear regression analysis (backward method, in=0.05, out=0.10). We defined $P<.05$ (test level=.05, two-tailed) as statistically significant.

**Results**

**Principal Results**

A total of 100 patients joined and completed this study, with 51 (51%) in the intervention group and 49 (49%) in the control group. Demographic information (ie, gender, age, height, and weight) and clinical information (ie, diagnosis, pain type, and site of pain) of the 2 groups were not statistically different, nor was the intensity, pain interference, and adherence to pain medications at baseline ($Ps> .05$; Table 1).
<table>
<thead>
<tr>
<th>Variable</th>
<th>Intervention group (n=51)</th>
<th>Control group (n=49)</th>
<th>P value (statistical test)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographic information</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male, n (%)</td>
<td>38 (75)</td>
<td>34 (69)</td>
<td>.57 (χ²=0.325)</td>
</tr>
<tr>
<td>Age (years), mean (SD)</td>
<td>54.6 (14.0)</td>
<td>58.7 (14.8)</td>
<td>.16</td>
</tr>
<tr>
<td>Height (cm), median (IQR)</td>
<td>166.0 (160.0-170.0)</td>
<td>166.0 (160.0-169.5)</td>
<td>.29</td>
</tr>
<tr>
<td>Weight (kg), median (IQR)</td>
<td>55.0 (47.7-65.0)</td>
<td>55.0 (50.0-60.0)</td>
<td>.52</td>
</tr>
<tr>
<td><strong>Diagnosis, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung cancer</td>
<td>16 (31)</td>
<td>24 (49)</td>
<td></td>
</tr>
<tr>
<td>Gastrointestinal cancer</td>
<td>19 (37)</td>
<td>14 (29)</td>
<td></td>
</tr>
<tr>
<td>Head and neck cancer</td>
<td>6 (12)</td>
<td>2 (4)</td>
<td></td>
</tr>
<tr>
<td>Breast cancer</td>
<td>2 (4)</td>
<td>1 (2)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>8 (16)</td>
<td>8 (16)</td>
<td></td>
</tr>
<tr>
<td><strong>Pain site, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥2 sites</td>
<td>26 (51)</td>
<td>30 (61)</td>
<td>.16 (χ²=7.986)</td>
</tr>
<tr>
<td>Chest and abdomen</td>
<td>6 (12)</td>
<td>8 (16)</td>
<td></td>
</tr>
<tr>
<td>Head and neck</td>
<td>6 (12)</td>
<td>5 (10)</td>
<td></td>
</tr>
<tr>
<td>Back</td>
<td>6 (12)</td>
<td>5 (10)</td>
<td></td>
</tr>
<tr>
<td>Limbs</td>
<td>5 (10)</td>
<td>0 (0)</td>
<td></td>
</tr>
<tr>
<td>Other sites</td>
<td>2 (4)</td>
<td>1 (2)</td>
<td></td>
</tr>
<tr>
<td><strong>Pain type, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed pain</td>
<td>20 (39)</td>
<td>21 (43)</td>
<td>.08 (χ²=6.801)</td>
</tr>
<tr>
<td>Visceral pain</td>
<td>27 (53)</td>
<td>16 (21)</td>
<td></td>
</tr>
<tr>
<td>Neuropathic pain</td>
<td>3 (6)</td>
<td>9 (18)</td>
<td></td>
</tr>
<tr>
<td>Body pain</td>
<td>1 (2)</td>
<td>3 (6)</td>
<td></td>
</tr>
<tr>
<td><strong>Pain intensity, median (IQR)a</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Worst pain</td>
<td>7 (5-8)</td>
<td>7 (6-9)</td>
<td>.16</td>
</tr>
<tr>
<td>Least pain</td>
<td>2 (1-3)</td>
<td>2 (1-3)</td>
<td>.26</td>
</tr>
<tr>
<td>Average pain</td>
<td>4 (2-6)</td>
<td>4 (3-6)</td>
<td>.33</td>
</tr>
<tr>
<td>Present pain</td>
<td>2 (1-4)</td>
<td>3 (1-4)</td>
<td>.17</td>
</tr>
<tr>
<td><strong>Pain interference, median (IQR)a</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General activity</td>
<td>7 (4-10)</td>
<td>6 (3-8)</td>
<td>.31</td>
</tr>
<tr>
<td>Mood</td>
<td>5 (2-8)</td>
<td>5 (4-7)</td>
<td>.43</td>
</tr>
<tr>
<td>Walking ability</td>
<td>8 (2-10)</td>
<td>5 (2-9)</td>
<td>.27</td>
</tr>
<tr>
<td>Daily work</td>
<td>9 (4-10)</td>
<td>7 (4-9)</td>
<td>.10</td>
</tr>
<tr>
<td>Relationships</td>
<td>3 (2-6)</td>
<td>3 (2-6)</td>
<td>.94</td>
</tr>
<tr>
<td>Sleep</td>
<td>7 (4-9)</td>
<td>6 (5-9)</td>
<td>.88</td>
</tr>
<tr>
<td>Enjoyment of life</td>
<td>5 (2-7)</td>
<td>6 (2-8)</td>
<td>.61</td>
</tr>
<tr>
<td><strong>Baseline adherence, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonadherence</td>
<td>3 (6)</td>
<td>8 (16)</td>
<td>.10 (χ²=2.784)</td>
</tr>
<tr>
<td>Incomplete adherence</td>
<td>26 (51)</td>
<td>25 (51)</td>
<td></td>
</tr>
<tr>
<td>Complete adherence</td>
<td>22 (43)</td>
<td>16 (33)</td>
<td></td>
</tr>
</tbody>
</table>
These measures represent the baseline characteristics based on the Brief Pain Inventory.

**BPI Outcomes**

Pain intensity in the intervention group was significantly reduced compared with the control group. The worst pain scores, least pain scores, and average pain scores in the 2 groups were statistically different, with median values of 4 (IQR 3-7) vs 7 (IQR 5-8; P=.001), 1 (IQR 0-2) vs 2 (IQR 1-3; P=.02), and 2 (IQR 2-4) vs 4 (IQR 3-5; P=.001), respectively, favoring the intervention group. The difference in the present pain score of the 2 groups was not statistically significant (P=.23). However, the score of the intervention group was numerically lower than that of the control group (Table 2).

**Table 2. Brief Pain Inventory outcomes at week 4.**

<table>
<thead>
<tr>
<th>BPI item</th>
<th>Intervention group (n=51), median (IQR)</th>
<th>Control group (n=49), median (IQR)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pain intensity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Worst pain</td>
<td>4 (3-7)</td>
<td>7 (5-8)</td>
<td>.001</td>
</tr>
<tr>
<td>Least pain</td>
<td>1 (0-2)</td>
<td>2 (1-3)</td>
<td>.02</td>
</tr>
<tr>
<td>Average pain</td>
<td>2 (2-4)</td>
<td>4 (3-5)</td>
<td>.001</td>
</tr>
<tr>
<td>Present pain</td>
<td>1 (0-3)</td>
<td>2 (0-4)</td>
<td>.23</td>
</tr>
<tr>
<td><strong>Pain interference</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General activity</td>
<td>7 (4-8)</td>
<td>6 (3-8)</td>
<td>.76</td>
</tr>
<tr>
<td>Mood</td>
<td>3 (1-6)</td>
<td>4 (2-6)</td>
<td>.58</td>
</tr>
<tr>
<td>Walking ability</td>
<td>7 (4-10)</td>
<td>7 (3-8)</td>
<td>.32</td>
</tr>
<tr>
<td>Daily work</td>
<td>8 (6-10)</td>
<td>8 (6-9)</td>
<td>.15</td>
</tr>
<tr>
<td>Relationships</td>
<td>2 (1-4)</td>
<td>3 (1-5)</td>
<td>.64</td>
</tr>
<tr>
<td>Sleep</td>
<td>4 (1-7)</td>
<td>7 (3-8)</td>
<td>.10</td>
</tr>
<tr>
<td>Enjoyment of life</td>
<td>4 (2-7)</td>
<td>5 (2-8)</td>
<td>.43</td>
</tr>
</tbody>
</table>

**PROM Submission Through MediHK**

The number of forms submitted by the intervention group patients was much higher than that of the control group (710 vs 95), with an average of 4.64 forms per person per day vs 0.06 forms per person per day, respectively. The most common forms submitted by the control group were the BPI (53/95, 56%), pain diary (17/85, 18%), and medication list (15/95, 16%; Table 3). Even though the control group patients did not receive reminders to fill out the forms, they still actively contacted the pain management team through MediHK due to uncontrollable pain intensity, interference in daily life, or severe ADRs.

**Table 3. Number of PROMs submitted by the 2 groups.**

<table>
<thead>
<tr>
<th>Form</th>
<th>Intervention group (n=710), n (%)</th>
<th>Control group (n=95), n (%)</th>
<th>P value (χ²=153.236)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain diary</td>
<td>495 (69.7)</td>
<td>17 (18)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>ADR³ form</td>
<td>87 (12.2)</td>
<td>7 (7)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>BPI⁴</td>
<td>83 (11.7)</td>
<td>53 (56)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Medication list</td>
<td>31 (4.4)</td>
<td>15 (16)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>MMAM⁵</td>
<td>14 (2.0)</td>
<td>3 (3)</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

ADR: adverse drug reactions.  
BPI: Brief Pain Inventory.  
MMAM: Morisky Medication Adherence Measure.

**Medication Adherence**

The complete adherence rate in the intervention group increased from 43% (22/51) to 63% (32/51), while that of the control group increased from 33% (16/49) to 51% (25/49; χ²=12.864; P<.001; Figure 1).
Adverse Drug Reactions

The overall incidence of ADRs was 36% (36/100) across the 2 groups at baseline and increased to 56% (56/100) at week 4. ADR incidence in the intervention group was significantly higher than in the control group ($\chi^2=8.990; P=.003$). In addition, 3 cases of intestinal obstruction and 2 cases of delirium were observed in the intervention group. Table 4 shows the distribution of ADRs between groups.

Table 4. Adverse drug reactions between groups over 4 weeks.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Baseline</th>
<th>Week 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients with ADRa</td>
<td>Intervention group (n=51), n (%)</td>
<td>Control group (n=49), n (%)</td>
</tr>
<tr>
<td>ADR type</td>
<td>$P$ value (statistical test)</td>
<td>$P$ value (statistical test)</td>
</tr>
<tr>
<td>Constipation</td>
<td>24 (47)</td>
<td>12 (25)</td>
</tr>
<tr>
<td>Nausea and vomiting</td>
<td>18 (35)</td>
<td>6 (12)</td>
</tr>
<tr>
<td>Drowsiness</td>
<td>13 (26)</td>
<td>6 (12)</td>
</tr>
<tr>
<td>Dizziness</td>
<td>4 (8)</td>
<td>3 (6)</td>
</tr>
<tr>
<td>Pruritus</td>
<td>6 (12)</td>
<td>1 (2)</td>
</tr>
<tr>
<td>Urinary retention</td>
<td>2 (4)</td>
<td>—</td>
</tr>
<tr>
<td>Ileus</td>
<td>2 (4)</td>
<td>—</td>
</tr>
<tr>
<td>Delirium</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

aADR: adverse drug reaction.
bNot available.

Rehospitalization Rates During the 4 Weeks

The 2 groups had a similar rehospitalization rates within the 4-week trial. There was no significant difference between the 2 groups within 4 weeks ($\chi^2=0.010; P=.92$).

Analysis of Pain Factors

We introduced possible factors that could contribute to pain intensity for each pain item in a multivariate linear regression analysis. The physician-pharmacist intervention through MediHK was an independent influencing factor for the most severe pain ($\beta=-1.413; P=.005$; Multimedia Appendix 8) and...
average pain ($\beta=–1.154; P=.003$; Multimedia Appendix 9). Aside from medication adherence, the intervention was significantly related to the least pain ($\beta=–.701; P=.02$; Multimedia Appendix 10). No factors significantly influenced the present pain (gender $\beta=1.078, P=.16$; age $\beta=.018, P=.26$; height $\beta=.063, P=.18$; weight $\beta=-.017, P=.46$; adherence $\beta=-.282, P=.40$; intervention $\beta=-.598, P=.17$; Multimedia Appendix 11).

### Discussion

#### Principal Results

The self-management of cancer pain is full of challenges, especially, for ambulatory patients. Approximately 70% to 90% of cancer patients can relieve pain adequately when carefully following the treatment guidelines. No more fully developed, digital health helps to achieve good pain management in daily practice for ambulatory patients with cancer pain, particularly, in remote areas of China [6,11]. Patients’ various demands in supporting self-management help encourage the development of a multimodal web application [16,17].

This study included a joint physician-pharmacist team that managed ambulatory patients with cancer pain through a WeChat-supported platform, MediHK, with promising results. Even if the control group did not receive a reminder to fill out the forms, patients in this group actively contacted the pain management team through MediHK to determine whether the medication plan needed to be adjusted due to either uncontrollable pain intensity, interference on daily life, or severe ADRs. The results revealed the patients’ need to contact the professional team via MediHK for better pain management. The patients in the intervention group reported more ADRs compared with control group patients, primarily, because there were more reports obtained from intervention group patients. More ADRs were not in conflict with improving pain. For example, the pain management team added new drugs for pain treatment, which may have caused some ADRs, but most of them were tolerated after a few days and monitored closely by the pharmacist.

#### Comparison With Prior Work

Yang et al [18] developed an app named Pain Guard for better pain management of discharged patients. Its functions were similar to MediHK, such as self-evaluation, real-time medication consultation, and record-keeping. The differences were that, for MediHK, we combined the NRS, Face Pain Scale, and Verbal Rating Scale to assess pain intensity accurately, while Pain Guard had only the traditional scale, NRS. We designed the module to record more medication-related details from patients, including drug name, dose, frequency, initial stop time, pain relief after medication, and adverse reactions. In addition, the physician or pharmacist could send forms embedded in MediHK, such as the ADR or adherence assessment form, to patients according to their status. These functions were unavailable in the Pain Guard. Scriven et al [19] used the BPI to evaluate patients’ pain while participating in the multisite telehealth group model. They found positive changes on the interference scale at the individual level (14% of patients) but no change at the group level. Another study offered standardized education according to their status. These functions were unavailable in such as the ADR or adherence assessment form, to patients who were too old or unable to record, family members or caregivers would help to send the form. In total, we accounted for the universal applicability of MediHK when we developed it to ensure easy operation. The only complicated step was the switch between the interface. However, this was emphasized when training in the outpatient clinic.

Regrettably, we did not record the impact of education status and age in keeping medical records. Patients who had never received education may take longer to keep records. However, since the included patients or their families were all able to use WeChat proficiently, we believed that MediHK was feasible; for patients who were too old or unable to record, family members or caregivers would help to send the form. In total, we accounted for the universal applicability of MediHK when we developed it to ensure easy operation. The only complicated step was the switch between the interface. However, this was emphasized when training in the outpatient clinic.

Knowledge deficits, inadequate pain assessment, misconceptions of pain, complex environments, and infrequent communication with health care providers are barriers in pain management. A joint physician-pharmacist team operating through a digital health platform can improve it. The cancer patient pain assessment was complicated. It is necessary to select quantification tools and assess the cause, location, quality, and relieving or aggravating factors of the pain comprehensively. The time that physicians spend on each patient is limited, and it is difficult to provide long-term and continuous monitoring.
The digital platform can better solve these problems. The platform trains patients to record their pain conditions in a more standardized and targeted manner. During clinical encounters, clinicians can spend more time addressing patients’ concerns in a meaningful way, rather than running through checklists of questions [22]. In addition, this platform promotes patient self-management. It allows patients to pay attention to the daily changes in pain and offers a digital tool to seek out the help of a professional team when suffering from an intractable pain or serious ADRs.

It is essential to consider the clinical workflow, security and liability, and the time-cost. We conducted a preliminary investigation and consideration in the early stage and carried out several rounds of related process optimizations and software improvements. In addition, when patients first visited the clinic, we would state that their physicians and pharmacists would provide the home services via the platform, and patients trusted this service. Finally, patients signed an exemption agreement and informed consent to ensure medical safety before using the platform. The satisfaction of patients and medical workers on such digital health platforms matters. One study designed a module in a mobile app to survey overall satisfaction, and the questionnaire was completed by participants at the end of the study [18]. Another study also assessed patient satisfaction about the convenience and helpfulness of using mobile systems, receiving technical software support, receiving consultant and training courses, and prompt responses for help; the results indicated that patients had a high level of satisfaction toward these kinds of digital tools [21]. Our preliminary idea was to evaluate satisfaction by embedding a questionnaire. For patients, this included assessing the pattern of the platform and the pain management team’s joint management, the content of medication education, the acceptability of response time, and the overall services. For the pain management team, this included evaluating the ease of operation of the platform, the acceptability of clinical workflow interference, and working-time costs. The questionnaire could contain an open-ended question, in which both patients and the pain management team are encouraged to provide suggestions regarding improvements to MediHK.

Study Strengths
There were some strengths of this study. First, this study was a real-world randomized controlled trial conducted in a large ambulatory clinic of a tertiary hospital. All patients were clinically recruited and randomly assigned. The integration of PROMs has not been a feature of other eHealth and (web) application–related studies, allowing this digital health study to help advance this field. In addition, real-time reporting can facilitate just-in-time interventions based on an individual's current circumstance or environment. This study achieved real-time communication between ambulatory patients with cancer pain and health care providers through MediHK, extending medical services to ambulatory patients as a pathway for the self-management in home settings.

Study Limitations
The study had the several limitations. First, this study had abnormally high participation, which will not necessarily reflect what would happen when patients use the platform independently, because pharmacists would send daily notifications. Second, it was prospectively powered and conducted in a randomized manner, but inevitable confounding factors can exist in the real world. Multivariate linear regression can only explain a small part of the influence of different pain intensity types. Third, since the study was conducted in a single tertiary hospital, applying this approach in other clinical settings may require some individualization to meet specific needs. Fourth, the observation time of only 4 weeks limited the long-term application of the results. Fifth, this study lacked further assessment about buy-in from both patients and the pain management team.

Conclusions
The joint physician-pharmacist team operating through MediHK enhanced communication, optimized outcomes, and promoted self-management of patients in home settings. This study supports the feasibility of integrating the internet into patient self-management of cancer pain. In the future, it will be necessary to enlarge the sample size to further explore the long-term effects of this method on the self-management of ambulatory patients with cancer pain.

Acknowledgments
MediHK was developed by the joint efforts of the College of Mathematics and Computer Science, Hunan Normal University and the Department of Pharmacy of Xiangya Hospital, Central South University. The authors sincerely thank all of the patients and their relatives for cooperating in this study. We thank all physicians for their support and participation as well as the College of Mathematics and Computer Science, Hunan Normal University for their subsequent services.

The study was supported by Changsha Natural Science Foundation of 2020 (kq2007039), National Clinical Research Center for Geriatric Disorders, Xiangya Hospital (XYYYJSTG-15), Parallel and Distributed Processing for the Stable Support Project of the National Defense Science and Technology Key Laboratory (WDZC20205500121), The Fundamental Research Funds for the Central Universities of Central South University (2021zzts1040), and Science and Technology Development Center of Chinese Pharmaceutical Association (CMEI2021KPYJ (JZYY) 00301).
Authors' Contributions
HLM, K-KL, S-SS, X-PC, and JX designed the study. K-KL, LZ, and HLM wrote the manuscript. K-KL, HLM, H-XH, and LZ performed the research. JX and H-XH analyzed the data. Y-MH, YC, W-HL, and F-ZL contributed new reagents and analytical tools. F-ZL provided technology support.

Conflicts of Interest
HLM is on the board of directors for Vyant Bio. He is one of the founders of Clariifi LLC and a consultant to Viece and eviCORE Health Solutions.

Multimedia Appendix 1
Screenshots of the patient interface, including registration, homepage, and reminders.

Multimedia Appendix 2
Screenshots of the medical interface, including reminders, review, homepage, and message-sending.

Multimedia Appendix 3
Screenshots of the introduction of the MediHK.

Multimedia Appendix 4
Medication history list. English translation shown on right.

Multimedia Appendix 5
Adherence Assessment Measurement. English translation shown on right.

Multimedia Appendix 6
Adverse Reaction Form. English translation shown on right.

Multimedia Appendix 7
Pain diary. English translation shown on right.

Multimedia Appendix 8
The independent factors influencing worst pain intensity.

Multimedia Appendix 9
The independent factors influencing average pain intensity.

Multimedia Appendix 10
The independent factors influencing least pain intensity.

Multimedia Appendix 11
The independent factors influencing present pain intensity.

Multimedia Appendix 12
CONSORT-eHEALTH checklist (V 1.6.1).
References


Abbreviations

ADR: adverse drug reaction
BPI: Brief Pain Inventory
MediHK: Medication Housekeeper
NRS: Numerical Rating Scale
PROM: patient-reported outcome measures
ULN: upper limits of normal
Mobile Apps as Audience-Centered Health Communication Platforms

Michael Mackert¹,², PhD; Dorothy Mandell³,⁴, PhD; Erin Donovan⁵, PhD; Lorraine Walker⁶, EdD, MPH; Mike Henson-García⁷, MPH; Lindsay Bouchacourt¹, MA

¹Stan Richards School of Advertising and Public Relations, The University of Texas at Austin, Austin, TX, United States
²Center for Health Communication, The University of Texas at Austin, Austin, TX, United States
³School of Community and Rural Health, University of Texas Health Science Center at Tyler, Tyler, TX, United States
⁴Population Health, The University of Texas System, Austin, TX, United States
⁵Department of Communication Studies, The University of Texas at Austin, Austin, TX, United States
⁶School of Nursing, The University of Texas at Austin, Austin, TX, United States
⁷School of Public Health, University of Texas Health Science Center - Dallas Regional Campus, Dallas, TX, United States

Corresponding Author:
Michael Mackert, PhD
Stan Richards School of Advertising and Public Relations
The University of Texas at Austin
300 W Dean Keeton St
Austin, TX, 78712
United States
Phone: 1 512 348 8490
Email: mackert@utexas.edu

Abstract

Health communication campaigns often suffer from the shortcomings of a limited budget and limited reach, resulting in a limited impact. This paper suggests a shift of these campaigns to audience-centered communication platforms—particularly, apps on mobile phones. By using a common platform, multiple interventions and campaigns can combine resources and increase user engagement, resulting in a larger impact on health behavior. Given the widespread use of mobile phones, mobile apps can be an effective and efficient tool to provide health interventions. One such platform is Father’s Playbook, a mobile app designed to encourage men to be more involved during their partner’s pregnancy. Health campaigns and interventions looking to reach expectant fathers can use Father’s Playbook as a vehicle for their messages.

(JMIR Mhealth Uhealth 2021;9(8):e25425)  doi:10.2196/25425

KEYWORDS
health communication; mHealth; mobile apps; mobile health; prenatal health; pregnancy; audience-centered

Introduction

There are numerous public health issues, ranging from disparities in maternal mortality to the reframing of child abuse as a population health concern, where communication plays a key part in the solution. For example, the US Centers for Disease Control and Prevention’s Tips From Former Smokers was a national mass media anti-smoking campaign that profiled real people living with serious long-term health effects from smoking [1]. The campaign is known for its television spots with graphic and emotional testimonials from former smokers. This massive campaign helped approximately 1 million people successfully quit smoking [1]. A health communication example using new media is the Text4Baby smartphone app, which is aimed toward expectant and new mothers. The app provides information on topics ranging from baby milestones to nutrition to childcare tips, and it sends over 250 SMS text messages to the user’s phone with the most critical information for pregnant women and mothers [2]. Women who used the Text4Baby app felt more prepared to be new mothers [3], had higher attitudes toward prenatal vitamins [4], and had a higher level of pregnancy health knowledge [4].

Although high-profile and effective health communication interventions exist, it must be acknowledged that many do not achieve their objectives. A primary reason could be that health communication interventions are “underdosed”—most simply do not have the budget to reach enough people to make a
substantial difference. Although the Tips From Former Smokers campaign was very successful, it had a budget of US $48 million in 2012 [1], which is anomalously high among health campaign budgets. This applies to health communication campaigns that focus on broad message dissemination as well as to more complex and interactive interventions.

Health communication could benefit from a fundamental paradigm shift. Interventions are often designed appropriately by following best practices [5,6], working within the budget and evaluation constraints of a particular project. To improve the efficacy of health communication interventions, we suggest a shift to audience-centered communication platforms, which are platforms that focus on a specific audience but tailor their content to address different subject matter (in this case, health issues) or different subsets of the audience.

When considering the potential of an audience-centered platform, mobile apps specifically can be extremely effective due to their versatility in content and widespread reach. Mobile apps could provide opportunities for multiple interventions and campaigns to combine resources and increase user engagement in service of promoting health behavior change and public health. As just one example, Bright by Text, a smartphone app for parents that provides information about early childhood topics, has regional implementations [7]. These regional implementations provide location-specific activities for parents to do with their children; this geographic tailoring can make the content more specific and useful for parents who live in different regions across the United States and avoids the need for regional programs to find their own ways to spread information to those parents.

Given the widespread use of mobile devices in the general public [8], delivering health messages through mobile apps is particularly useful [9]. Mobile apps can efficiently deliver appropriate doses of health messages along with providing “in-the-moment” health interventions [10,11], which are essentially a type of just-in-time adaptive intervention [12]. Mobile apps can deliver tailored health messages to people in specific and opportune moments within their everyday settings, effectively addressing the “message dosing issues” many health campaigns face [10,11]. Mobile apps can also effectively intervene among multiple health issues in one specific group or population. Health communication interventions are often built around a single health issue, despite the fact that many health issues coincide together within specific groups of individuals [13-15]. The time and resources used on promoting many independent health communication interventions could be better spent if there was a common platform that could be shared, especially if the interventions were all targeting a common audience.

One such platform to share health information is Father’s Playbook, a smartphone app designed for men to use during and after their partner’s pregnancy. Health communication interventions and campaigns attempting to reach expectant fathers can use Father’s Playbook as a way to reach an audience that already exists.

Father’s Playbook Case Study

The Father’s Playbook smartphone app was created to fill a large gap in pregnancy-related health information—the lack of male-focused pregnancy information and resources. While women should be the main focus during pregnancy, including men has benefits for father, mother, and baby. Father involvement during the prenatal stage has shown to have many positive outcomes, including increased communication between partners, higher number of provider visits during pregnancy, and increased postpartum best practices [16,17]. Although men do feel strongly about being involved in pregnancy health [18], structural health care barriers and personal barriers may prevent them from being engaged. To help combat these barriers, Father’s Playbook was designed to engage men in prenatal health.

The Father’s Playbook app was developed using an incremental approach to the research and the app’s development. This process was cost-effective, and it allowed us to learn from mistakes and build on previous knowledge. Before the Father’s Playbook smartphone app existed, a web-based pilot was created. To obtain more generalizable results regarding attitudes and barriers to prenatal health involvement, a survey with a nationally representative sample of men was conducted [18]. Overall, men believe that it is important to be involved in pregnancy health; however, perceived barriers (eg, time restraints, unclear role, financial burdens) still exist [18]. This survey also required participants to interact with the website, and the participants made suggestions on how to improve the site. Through interviews and the survey, we were able to obtain men’s opinions on the website, along with suggestions for features that would improve the website [19]. This information allowed us to prioritize features to be developed (articles and an interactive budget calculator) and focus on user testing for new features.

Both the app’s development and future research related to the app have next steps. Presently, the app’s content is available in English and Spanish, which opens the door to a larger audience of men. Following this type of audience specification, in the future, the goal is to tailor content toward different types of fathers. For example, the experience of a stay-at-home father versus a single parent father greatly differs, and the pregnancy and fatherhood experience of transgender men who become pregnant will differ from the experience of a cisgender man [20]. On the research side, the next step will be to focus on the father’s engagement in prenatal health in a broader sense. To be able to measure the effectiveness of Father’s Playbook in improving father engagement, baseline data is needed on the current level of father engagement in prenatal health. As such, a representative survey targeting fathers and expectant fathers is being conducted to gather baseline data. Future development of the app can strengthen its potential as a platform to enable more efficient communication with expectant fathers rather than individual programs and efforts to reach this audience.

There are a wide variety of opportunities to improve paternal, maternal, and child health through improved communication with expectant fathers. There is a wealth of evidence showing...
that fathers want to be more engaged in the pregnancy and feel underprepared for fatherhood [19,21,22]. However, there is a dearth of research focused on how to engage and prepare fathers. The app provides a flexible platform to test education and communication strategies with fathers directly. For instance, directed education about nutrition and pregnancy complications may result in men assisting their pregnant partners in making healthier nutrition choices and may increase their ability to identify obstetric emergencies. Evidence also suggests that when expectant parents are unsatisfied with their partnerships, they are more likely to exhibit insensitive parenting styles once their infant is born. The app may provide a platform for disseminating communication interventions to help the quality of the relationship between expecting couples [23]. Additionally, studies have found that low levels of paternal engagement throughout the prenatal period are related to birth complications [24] and to low postnatal engagement. Positive father involvement in early childhood is tied to a variety of positive cognitive and health outcomes for children [25]; therefore, increasing the father’s engagement has the potential to result in positive health outcomes across the entire family unit.

**Father’s Playbook as an Audience-Centered Platform**

Father’s Playbook was designed using audience-centered principles, which means that the content, language, and approach used in the app are tailored to best fit the audience (new and expectant fathers). By focusing on the audience, the app can tailor information and its delivery to the needs of the audience. In its current format, Father’s Playbook is a single app that includes content pages about specific father-related information and interactive app features (such as a budget calculator) to encourage expectant fathers to become more engaged during and after their partner’s pregnancy. Future versions of the app will include tailored content, allowing the app to be more personally relevant to its user. This ability to tailor content can allow other health communication scholars to use the app to deploy tailored interventions and campaigns to the target audience of expectant and new fathers.

The shift to thinking of audience-centered platforms can broaden the reach and efficiency of interventions that have successfully developed an approach for reaching a particular audience. As an example, although Father’s Playbook is currently focused on amplifying the engagement of expectant fathers during the prenatal period, there are other issues that commonly affect expectant fathers, and at any time, the app can deploy other types of content to address the audience’s other public health needs. Other examples of public health information expectant fathers might want to consume include management of nutrition and physical activity throughout pregnancy, prevention of paternal postpartum depression, smoking cessation, and information related to paternity testing. These different health issues can be addressed specifically through tailored content in the app, without the need for researchers and professionals with interests in these issues to develop an entirely new intervention or campaign and then determine how to reach the target audience.

Along with addressing health issues of expectant and new fathers, Father’s Playbook can be used to address parenting issues and child-rearing best practices. For example, a researcher could be interested in testing an intervention that increases the amount of time a father reads to his children; as part of the intervention, participants would need to download the Father’s Playbook app. The app could add features that specifically work to reach the intervention objectives, such as content articles that discuss the benefits of fathers reading to their children, interactive games that encourage book reading, and a list of book suggestions appropriate for specific age ranges.

Father’s Playbook was built with the goal of increasing father engagement during pregnancy. Now that an audience of expectant and new fathers exists, other researchers and practitioners can access this unique audience to address a myriad of parenting and health issues, allowing a collaborative approach toward health communication campaigns and interventions. Without the use of an audience-centered platform, health professionals would only need to use their own resources to reach this audience.

**Conclusion**

Health communication can be an effective tool to help improve the health and well-being of individuals and populations. There is a strong evidence base of health communication that can be leveraged across health issues and audiences, such as the increased efficacy of targeted and tailored messages compared to more general appeals [26].

Despite the high-profile success of health communication campaigns that have achieved important and demonstrated benefits, substantial opportunities remain to advance the field through new approaches to message design and reaching audiences. The approach advocated in this paper is to focus on more audience-focused platforms, which could be a more efficient strategy for message dissemination. Taking an audience-centered approach allows for a better understanding of people, their behaviors, their contexts, and their intersections allowing for more nuanced health communication and health promotion efforts. Given that disease manifests from the compound of multiple risk factors—and said factors are differentially distributed across various lifestyles and identities—audience-centered approaches have the potential to be highly effective vehicles for health transformation and useful for any number of audiences, ranging from transgender women to Black men who have sex with men to older persons managing multiple chronic conditions to COVID-19 survivors.

It is a well-established principle of health communication that targeted and tailored communication is more effective than general messages. The approach advocated in this paper—to build audience-centered communication platforms—is a promising approach to develop more cost-effective, engaging, and effective health communication interventions.
Acknowledgments

This article is based on work resulting from participation in the Puentes 2.0 Conference sponsored by the Department of Population Health, the Lozano Long Institute of Latin-American Studies (LLILAS)–Benson Mexico Center, the Office of the Vice President of Research, and University of Texas System Collaboration for Population Health Innovation and Improvement (CoPHII) project. This project was funded through the Texas Safe Babies project (IAC 24307170) from the Texas Department of Family and Protective Services. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Texas Department of Family and Protective Services.

Conflicts of Interest

None declared.

References

7. Bright by Text. URL: https://brightbyte.org [accessed 2021-04-27]


Abbreviations

CoPHII: Collaboration for Population Health Innovation and Improvement
LLILAS: Lozano Long Institute of Latin-American Studies
Enabling Wearable Pulse Transit Time-Based Blood Pressure Estimation for Medically Underserved Areas and Health Equity: Comprehensive Evaluation Study

Venu Ganti¹, BS; Andrew M Carek¹, PhD; Hewon Jung¹, MS; Adith V Srivatsa²; Deborah Cherry³, BBA; Levather Neicey Johnson³, BSN, RN; Omer T Inan¹,², PhD

¹School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, United States
²The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, United States
³VSNS, Inc, Atlanta, GA, United States

Corresponding Author:
Venu Ganti, BS
School of Electrical and Computer Engineering
Georgia Institute of Technology
85 5th St NW
Atlanta, GA, 30308
United States
Phone: 1 2406434250
Email: vganti6@gatech.edu

Abstract

Background: Noninvasive and cuffless approaches to monitor blood pressure (BP), in light of their convenience and accuracy, have paved the way toward remote screening and management of hypertension. However, existing noninvasive methodologies, which operate on mechanical, electrical, and optical sensing modalities, have not been thoroughly evaluated in demographically and racially diverse populations. Thus, the potential accuracy of these technologies in populations where they could have the greatest impact has not been sufficiently addressed. This presents challenges in clinical translation due to concerns about perpetuating existing health disparities.

Objective: In this paper, we aim to present findings on the feasibility of a cuffless, wrist-worn, pulse transit time (PTT)–based device for monitoring BP in a diverse population.

Methods: We recruited a diverse population through a collaborative effort with a nonprofit organization working with medically underserved areas in Georgia. We used our custom, multimodal, wrist-worn device to measure the PTT through seismocardiography, as the proximal timing reference, and photoplethysmography, as the distal timing reference. In addition, we created a novel data-driven beat-selection algorithm to reduce noise and improve the robustness of the method. We compared the wearable PTT measurements with those from a finger-cuff continuous BP device over the course of several perturbations used to modulate BP.

Results: Our PTT-based wrist-worn device accurately monitored diastolic blood pressure (DBP) and mean arterial pressure (MAP) in a diverse population (N=44 participants) with a mean absolute difference of 2.90 mm Hg and 3.39 mm Hg for DBP and MAP, respectively, after calibration. Meanwhile, the mean absolute difference of our systolic BP estimation was 5.36 mm Hg, a grade B classification based on the Institute for Electronics and Electrical Engineers standard. We have further demonstrated the ability of our device to capture the commonly observed demographic differences in underlying arterial stiffness.

Conclusions: Accurate DBP and MAP estimation, along with grade B systolic BP estimation, using a convenient wearable device can empower users and facilitate remote BP monitoring in medically underserved areas, thus providing widespread hypertension screening and management for health equity.

(JMIR Mhealth Uhealth 2021;9(8):e27466) doi:10.2196/27466

Keywords
wearable sensing; pulse transit time; cuffless blood pressure; noninvasive blood pressure estimation; health equity; mobile phone
Introduction

Background

Current clinical practice regarding hypertension management and control hinges on the century-old approach of obtaining infrequent cuff-based measurements of blood pressure (BP) in clinical settings. This paradigm of the measurement being anchored to the clinical setting and requiring persons to proactively visit a medical professional to determine their hypertensive status is costly—due to the time and money spent [1]—and considered ineffective—due to the infrequency and error (ie, white coat hypertension) of office BP measurements [2,3]. Hence, we observed remarkable disparities in hypertension detection, treatment, and control across socioeconomic status and race, with populations lacking access to regular office visits and care, having nearly half the awareness of their existing hypertensive status, and enduring up to triple the rates of subsequent cardiac events [4,5]. Technologies that enable frequent, reliable, and accurate measurements of BP in ambulatory settings promise to reduce the global burden of hypertension and offer an opportunity to advance health equity [6]. Leveraging the ubiquity of smartphones and digital health technologies equipped with highly sensitive, miniaturized sensors is essential for the remote monitoring of BP [7].

Existing wearable devices that incorporate noninvasive BP methodologies offer an affordable and efficient means of tracking out-of-office BP [8]. Unfortunately, they commonly use uncomfortable techniques (ie, oscillimetry and tonometry) that demand imparting forces on blood vessels to achieve accurate measurements [9-11]. These inconveniences fail to empower users to take control of their health, posing a significant challenge to the widespread routine monitoring of BP. Instead, strategies that compute the pulse transit time (PTT), a measure of arterial stiffness, present a convenient alternative for BP estimation [12].

The PTT, the time the pressure wave propagates along the length of the arterial tree, is a cuffless surrogate for BP and can be acquired noninvasively [12]. In practice, the acquisition of noninvasive PTT requires a combination of sensors—typically an accelerometer, force sensor, light-emitting diode (LED) and photodiode, electrode, or ultrasonic transceiver—placed proximally and distally along the arterial tree and computed from fiducial points in the captured seismocardiogram (SCG), ballistocardiogram, photoplethysmogram (PPG), impedance cardiogram, impedance plethysmogram, or arterial blood pressure (ABP) waveform [13,14]. Despite their inherent convenience, these sensing modalities are naturally of concern when used in populations with intrinsic mechanical, optical, and electrical barriers, stemming from higher melanin levels and body fat percentages.

To the best of our knowledge, noninvasive PTT-based BP estimation has yet to be examined in a diverse population—a gap in our scientific understanding that presents a formidable obstacle to its adoption. Specifically, some medically underserved areas (MUAs), which stand to benefit the most from remote monitoring [15], have a large number of Black and Latino individuals with higher melanin content and obesity rates compared with White individuals [16]. Recent notable data from the Centers for Disease Control and Prevention further stress these concerns by exposing that non-Hispanic Black individuals not only have significantly higher hypertension prevalence than non-Hispanic White individuals but also witness significantly lower hypertension control rates [17,18]. Affordable remote monitoring options have the responsibility to combat not only social determinants of health, such as access to health care and income, but also in turn the existing health disparities that are byproducts of them. As a result, there exists a glaring hole in PTT-based BP monitoring—that this technology has yet to be tested on the population for whom it may be the most valuable, and until now, its continuing practice will only exacerbate existing health disparities.

Objectives

In our previous work, we designed a wearable, multimodal, wrist-worn PTT monitoring device (SeismoWatch) and validated it in both controlled lab [19] and unsupervised home [20] settings, primarily on young, healthy persons with lighter skin. In this paper, we expand upon our previous work with a community-engaged research strategy that leverages expertise from a nonprofit organization serving MUAs in Georgia and evaluated our device in a more diverse population. We present our device’s ability to accurately estimate BP in this diverse population and capture significant demographic-level differences in underlying arterial stiffness that coincide with observations from existing literature, through the calibration coefficients used in our BP estimation model. This work represents the first time that a noninvasive, cuffless, PTT-based wearable device has been extensively evaluated in a community-based diverse population as a potentially reliable and convenient monitoring option toward, ultimately, the remote screening and management of hypertension for health equity.

Methods

Study Protocol

A comprehensive breakdown of the demographics of the study population is presented in Table 1. This study was conducted under a protocol approved by the Georgia Institute of Technology institutional review board (protocol number H19251). The study was separated into two different populations (N=44 participants) referred to throughout this work as follows: (1) a young and healthy homogeneous population (first cohort=26 participants) and (2) an older, entirely Black, higher BMI, metropolitan population (second cohort=18 participants) recruited later through the help of our community outreach partners—a nonprofit organization serving medically underserved persons in the state of Georgia. For the first cohort, 26 (19 males and 7 females) young and healthy volunteers (mean age 26.7 years, SD 3.7; mean weight 73.8 kg, SD 14.1; height 173.9 cm, SD 9.6; and mean BMI 24.2 kg/m², SD 3.2) with no previous history of cardiovascular disease were recruited, and written informed consent was obtained. For the second cohort, 18 (6 males and 12 females) Black participants (mean age 44.1 years, SD 11.7; mean weight 94.4 kg, SD 18.0; mean height 169.6 cm, SD 11.5; and mean BMI 33.2 kg/m², SD 7.6) with no previous history of cardiovascular disease other than
hypertension were recruited from the Atlanta metropolitan area, written informed consent was obtained, and further demographic information was collected post hoc with verbal consent. Both hypertensive status and the use of regular prescription medications were self-reported.

Table 1. Participant demographics and cardiovascular parameters for study participants (grouped by cohort; N=44).

<table>
<thead>
<tr>
<th>Demographics and cardiovascular parameters&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Homogenous data set (first cohort; n=26; participant 1-26)</th>
<th>Community outreach (metropolitan Atlanta) data set (second cohort; n=18; participant 27-44)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years), mean (SD)</td>
<td>26.7 (3.7)</td>
<td>44.1 (11.7)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>19 (73)</td>
<td>6 (33)</td>
<td>N/A&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Female</td>
<td>7 (26)</td>
<td>12 (67)</td>
<td>N/A</td>
</tr>
<tr>
<td>Height (cm), mean (SD)</td>
<td>173.9 (9.6)</td>
<td>169.6 (11.5)</td>
<td>.19</td>
</tr>
<tr>
<td>Weight (kg), mean (SD)</td>
<td>73.8 (14.1)</td>
<td>94.4 (18.0)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>BMI&lt;sup&gt;c&lt;/sup&gt; (kg/m&lt;sup&gt;2&lt;/sup&gt;), mean (SD)</td>
<td>24.2 (3.2)</td>
<td>33.2 (7.6)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Obesity class, n (%)</strong></td>
<td></td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>I</td>
<td>1&lt;sup&gt;d&lt;/sup&gt; (4)</td>
<td>2&lt;sup&gt;e&lt;/sup&gt; (11)</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>N/A</td>
<td>3&lt;sup&gt;f&lt;/sup&gt; (17)</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>N/A</td>
<td>4&lt;sup&gt;g&lt;/sup&gt; (22)</td>
<td></td>
</tr>
<tr>
<td><strong>Race, n (%)</strong></td>
<td></td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>Black</td>
<td>1&lt;sup&gt;h&lt;/sup&gt; (4)</td>
<td>18 (100)</td>
<td></td>
</tr>
<tr>
<td>Other race</td>
<td>25 (96)</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td><strong>Hypertensive status, n (%)</strong></td>
<td></td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>Normotensive</td>
<td>26 (100)</td>
<td>15 (83)</td>
<td></td>
</tr>
<tr>
<td>Hypertensive</td>
<td>N/A</td>
<td>2&lt;sup&gt;i&lt;/sup&gt; (11)</td>
<td></td>
</tr>
<tr>
<td>Hypotensive</td>
<td>N/A</td>
<td>1&lt;sup&gt;j&lt;/sup&gt; (6)</td>
<td></td>
</tr>
<tr>
<td><strong>Current medications, n (%)</strong></td>
<td></td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>Hydrochlorothiazide (1×day)</td>
<td>N/A</td>
<td>2&lt;sup&gt;k&lt;/sup&gt; (11)</td>
<td></td>
</tr>
<tr>
<td>Lisinopril (1×day)</td>
<td>N/A</td>
<td>1&lt;sup&gt;l&lt;/sup&gt; (6)</td>
<td></td>
</tr>
<tr>
<td>Iron supplement</td>
<td>N/A</td>
<td>1&lt;sup&gt;m&lt;/sup&gt; (6)</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Statistical significance between groups in values, where applicable, was computed using an unpaired two-tailed t test.

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

<sup>c</sup>Obesity classified using the BMI per the guidelines from the National Heart, Lung, and Blood Institute of the National Institutes of Health [21] (I: BMI=30-34.9; II: BMI=35-39.9; III: BMI ≥40).

<sup>d</sup>Participant 23.

<sup>e</sup>Participants 30 and 43.

<sup>f</sup>Participants 38, 40, and 42.

<sup>g</sup>Participants 34, 36, 37, and 41.

<sup>h</sup>Participant 5.

<sup>i</sup>Participants 29 and 37.

<sup>j</sup>Participant 33.

<sup>k</sup>Participants 29 and 37.

<sup>l</sup>Participant 29.

<sup>m</sup>Participant 33.

The concept of the study design is shown in Figure 1. Although not explicitly shown, two versions of the SeismoWatch were used in this study: a previous version of the hardware with comparable sensors was used in the young, homogeneous population (ie, the first cohort), before being adapted for a more robust, portable, and multimodal wearable device used in the metropolitan Atlanta population (ie, the second cohort). Specifically, the data from these cohorts were collected during
two intervals, between which the hardware was revised to incorporate multiwavelength PPGs before investigating the performance of the sensing modality in the underrepresented population. This was essential to assess the efficacy of shorter-wavelength LEDs (i.e., those with shallower skin penetration depths) in a Black population. However, in both the correlations in Figure 2 and calibration coefficient comparisons in Figure 3, only the results derived from the infrared (IR) PPGs, available to both devices, were computed and shown. The other key sensing components and reference system components were essentially identical: (1) the first version of the device used an analog version of the accelerometer (ADXL354, Analog Devices) to acquire the SCG, whereas the second version simply used the digital version of the same sensor (ADXL355, Analog Devices) to reduce size and (2) the finger-cuff continuous BP reference system (ccNexfin, Edwards Lifesciences) along with the data acquisition module (MPU150, Biopac Systems) were identical in both studies.

Figure 1. Concept overview and study design. Sensor information and placement locations for wearable system (blue) and reference system (purple). Noninvasive pulse transit time (PTT) measurement concept overview using seismocardiogram (SCG) and photoplethysmogram (PPG) sensors. Study protocol tasks in chronological order with duration and mean (SD) of mean arterial pressure (MAP) values for each task. Sample filtered signals from the participant with the lowest signal-to-noise ratio (SNR) signals (n=37): a hypertensive, high BMI, older Black female. In order from top to bottom: electrocardiogram (ECG), SCG, infrared PPG, red PPG, green PPG signals measured from the wearable system (blue) and the synchronized ECG, and arterial blood pressure (ABP) signals measured by the reference system (purple). Systolic blood pressure (SBP; top) and diastolic blood pressure (DBP; bottom) plotted across the full protocol for participant 37, with rest periods (green) and perturbations used to modulate BP (red) highlighted in chronological order, and the location where the reference finger-cuff continuous blood pressure (BP) system was paused during the exercise indicated. ABP: arterial blood pressure; BP: blood pressure; DBP: diastolic blood pressure; ECG: electrocardiogram; LED: light-emitting diode; PD: photodiode; PPG: photoplethysmogram; PTT: pulse transit time; SBP: systolic blood pressure; SCG: seismocardiogram.
Figure 2. Wearable pulse transit time (PTT)–based blood pressure (BP) estimation results. Correlation and Bland-Altman plots between PTT-estimated BP and the finger-cuff continuous BP for mean arterial pressure, diastolic blood pressure, and systolic blood pressure estimation. The root mean squared error and the mean absolute difference for each correlation are shown. DBP: diastolic blood pressure; MAD: mean absolute difference; MAP: mean arterial pressure; PTT: pulse transit time; RMSE: root mean square error; SBP: systolic blood pressure.

Figure 3. Participant-specific diastolic blood pressure (DBP) calibration coefficients are significantly different in demographics with typical disparities in arterial stiffness. Boxplots showing the statistically significant (*$P<.05$; Mann-Whitney U) difference in the DBP K1 and K2 calibration coefficients between participants who are nonobese and obese. Boxplots showing the statistically significant (*$P<.05$; Mann-Whitney U) difference in the DBP K1 calibration coefficients between male and female participants. Boxplots showing the statistically significant (*$P<.05$; Mann-Whitney U) difference in the DBP K1 calibration coefficients between participants of other race and Black participants. Boxplots showing the difference in the DBP K1 and K2 calibration coefficients between young and older participants. DBP: diastolic blood pressure.

To acquire a timing reference for the start of a cardiac cycle, while serving as a reference for alignment to the wearable system signals, a wireless electrocardiogram (ECG) module (BN-EL50, Biopac Systems) was attached to the participant in a three-lead configuration with Ag/AgCl gel electrodes as shown in Figure 1. As depicted in Figure 1, a finger-cuff BP sensor
based on the volume-clamp methodology (ccNexfin, Edwards Lifesciences) [22,23] was placed on the same hand as the watch, acquiring a reference measurement of continuous beat-by-beat BP. Although volume-clamping continuous BP devices are not the clinical gold standard for ABP measurements, an arterial line was not feasible due to invasiveness, and a sphygmomanometer was not used because of the need for a trained professional and lack of beat-by-beat BP data. Similarly, semi-automated BP cuffs were not used as they hinge on following strict guidelines to obtain an accurate reading, such as being seated and resting the arm at heart level, which were impossible to satisfy simultaneously while acquiring watch measurements, given the need for the contralateral hand to touch the ECG electrode to activate the PTT mode [20]. In addition, it was recently demonstrated that a volume-clamping—based system had comparable accuracy with noninvasive oscillometric BP cuffs [24]. All reference system sensors were sampled at 1 kHz and interfaced to a computer using a data acquisition system (MPU150, Biopac Systems) and its corresponding software (Acqknowledge, Biopac Systems). The reference system files were saved to a desktop computer for postprocessing.

Participants were asked to change into either a V-cut T-shirt or tank top, if not wearing one already, to acquire the sternal PPGs included in the wearable designs, though not examined in this work. The watch was fitted such that the PPGs faced the radial artery on the ventral side of the wrist. To capture the PTT, the participant performed a simple maneuver to place the watch on the sternum to acquire the SCG for the proximal timing reference, as shown in Figure 1, whereas the PPGs were sampled at both the sternum and wrist. Although this offers a noncontinuous assessment, routine remote BP monitoring using oscillometric devices has already demonstrated clinical value, although it does not offer continuous BP measurement [2]. Specifically, ambulatory BP monitors, due to their superior portability and measurement frequency—comparable with what this wearable device can easily provide [20]—have become invaluable for the screening and management of hypertension [2] such that the added benefit of continuous BP measurement may only be marginal.

In order, the participants went through a 2-minute baseline period while sitting before obtaining another 2-minute baseline measurement while standing. Then, a series of perturbations with varying rest periods in between were used to modulate BP. First, a mental arithmetic exercise was used to increase BP [12], in which participants were given a three-digit integer and were told to add the sum of the digits to the number repeatedly for 1 minute. Then, a cold pressor test was conducted in which participants submerged their hand contralateral to the watch in a bucket of ice water for as long as tolerable or until the full minute. Finally, during the exercise session, the finger cuff was removed to avoid damage, and the participant performed either a stair stepping or bicycling exercise, based on personal preference, for 1 minute. As mentioned in our previous work [20], the new version of the watch enters the PTT measurement mode when the user places a finger from their hand contralateral to the watch on the positive wrist ECG electrode; therefore, we were unable to acquire PTT data during exercise for both cohorts and cold pressor for the participants in the second cohort (ie, second cohort). Although with the newer hardware, we were unable to collect PTT data during the cold pressor perturbation for the second cohort, the effect of the cold pressor—assessed directly after the hand was removed from the ice water (ie, a maximum of 1 minute after immersion)—was still well within its physiological window during the following rest period [25]. Overall, as our device is not designed to offer continuous measurements of BP, examining the effect of these perturbations in the rest period directly following them, similar to our previous work [19], still allowed for a comprehensive evaluation of the methodology in a diverse population. However, PTT data from the first cohort during the cold pressor were still used. As the BP data from the cold pressor test were still acquired for the second cohort, as the continuous BP cuff was still on, the mean arterial pressure (MAP) values were factored into the ones displayed in Figure 1. To do so, a 50 ms moving average filter was applied to the measured continuous BP signal, ensemble averages of 10 heartbeats with 50% overlap were taken, and the BP beat with the highest signal-to-noise ratio (SNR) was selected.

Signal Processing

The signal processing pipeline is shown in Figure 4. All signal processing and statistical analyses were performed in MATLAB R2018a (MathWorks). Before preprocessing the SCG and PPG signals acquired from the wearable system, it was imperative to time-align them to the continuous BP signal from the reference system using the respective ECGs to ensure proper temporal comparison. Specifically, the ECGs from each system were first filtered using a digital finite impulse response bandpass filter (BPF; fpass=10-40 Hz) to remove baseline wander due to postural sway and extract the R-wave, which was then identified using a simple peak detection algorithm. Then, cross-correlation was used to align the R-peaks of the two ECG readings by detecting the amount of lead and truncating either the wearable or reference signals depending on the condition. After alignment, the dorsoventral axis of the SCG (ie, z-axis acceleration) and green, red, and IR wrist PPGs were filtered using a digital finite impulse response BP filter with bandwidths of 1-40 Hz and 1-8 Hz, respectively, to remove their out-of-band noise and baseline wander due to respiration. In addition, the continuous ABP waveform was smoothed using a 50 ms moving average filter.
Figure 4. Signal processing pipeline. Block diagram of signal processing overview showing signal alignment using electrocardiogram signals acquired from the wearable system (blue) and reference system (purple) before bandpass filtering, heartbeat windowing, and photoplethysmogram (PPG) selection. After beat selection and signal quality assessment, the pulse transit time is computed as the aortic valve opening point of the seismocardiogram to the diastolic foot of the PPG. Calibration is used to estimate blood pressure (BP) using the arterial blood pressure waveform acquired from the continuous BP finger-cuff. Block diagram of the custom PPG selection algorithm, locating beats with greater systolic upstrokes and signal-to-noise ratio (SNR). ABP: arterial blood pressure; AO: aortic valve opening; BP: blood pressure; BPF: bandpass filter; ECG: electrocardiogram; PPG: photoplethysmogram; PTT: pulse transit time; SBP: systolic blood pressure; SCG: seismocardiogram; SNR: signal-to-noise ratio.

Next, the filtered and aligned SCG, PPG, and ABP waveforms were split into separate heartbeats using the detected R-R intervals of the synchronized ECG. Then, these heartbeat-indexed signals were ensemble-averaged using 10-beat windows with 50% overlap before assessing the signal quality to select the highest quality beat per task for each participant, similar to the methods used in our previous works [19,20]. Given the number of high BMI participants in this population, the SCG not only had a lower mean SNR when compared with our previous studies but was also observed to have less variability than the PPG SNR; hence, an emphasis was placed on determining the optimal PPGs first. In addition, upon an initial assessment of signal quality, it was observed that when the PPG signal had the highest SNR, typically, the SCG signal did as well—perhaps because acquiring a clean PPG signal inherently hinges on applying consistent pressure. The optimum PPG was selected using a physiologically inspired algorithm to first identify the beats with the top 10% of systolic upstrokes (ie, maximum of the derivative of the PPG waveform) and then select the remaining beat with the maximum SNR. The SNR
was calculated using a noise-to-signal ratio detection algorithm detailed in Inan et al [26]. The methods used to determine the timing references for PTT calculation, the foot of the PPG, and the aortic valve opening (AO) point of the SCG were the same as those used in our previous studies [19,20]. Specifically, the foot of the PPG was computed using the intersecting tangent method described in the study by Mukkamala et al [12], and the AO point was assumed to be the first peak in each ensemble-averaged window before the foot of the PPG. Occasionally, the SCG signal was manually annotated to impose realistic constraints for the AO point or to ensure that the same morphological peak was consistently chosen for all tasks per participant. Participant-specific SNR thresholds were set to retain only high-fidelity signals; if the SNR of the SCG, PPG, or ABP beats was not greater than the prescribed cutoff, or if the foot of the PPG was not within a realistic range, then the respective ensemble-averaged waveforms were deemed too noisy for use and that task was not used for PTT calculation. Notably, the continuous reference BP allowed for the ability to evaluate the SNR of the ABP signal and incorporate this quality assessment into our signal processing pipeline to remove beats with low SNR reference measurements. After the entire signal quality assessment process, at least four of the tasks were used for BP estimation per participant. Finally, the PTT was calculated as the difference of the proximal timing reference, AO point of the paired SCG, and distal timing reference, the foot of the selected PPG. In addition to wavelength comparisons, the green and red wavelength PPGs were not used as the IR wavelength wrist PPGs had the highest mean SNR, because of the greater indifference of the IR wavelength to melanin absorption and the ability to capture more pulsatile arteries deeper in the tissue than cutaneous capillaries [12,27].

In addition, the postexercise recovery period was separated into an early and late rest period based on when the BP reached a consistent value. This allowed us to capture both the immediately heightened cardiac output–induced BP increase postexercise and the recovery back to baseline, while opportunoely adding another PTT and BP data point for linear regression.

Statistical Analysis

Simple linear regression was performed independently between wearable participant-specific inverse PTT (PTT−1) and reference diastolic blood pressure (DBP), MAP, or systolic blood pressure (SBP) value pairs, to calculate the calibration coefficients necessary to estimate each of the three BP components per participant; nonlinear models, whereas potentially more accurate, dictate the need for more calibration points [12,28]. Therefore, the resulting calibration coefficients—used to estimate BP from the conventional PTT-based BP estimation model shown in equation 1—are merely the slope (ie, slope calibration coefficient \(K_1\)) and y-intercept (ie, Y-intercept calibration coefficient \(K_2\)) of the line of best fit [12]. This was identical to the calibration methods used in our previous work [19,20].

\[
BP = (K_1 / PTT) + K_2 \quad (1)
\]

The mean absolute difference (MAD) was computed from the mean of the absolute value of the difference between the estimated and reference BP. The benchmarks for MAD were chosen based on the Institute for Electronics and Electrical Engineers (IEEE) standard for wearable cuffless BP measuring devices [29]. In addition, the root mean square error (RMSE), calculated as the root mean square of the difference between the estimated BP and measured BP, was computed because of its enhanced sensitivity to outliers.

We stratified the entire study population for the demographic comparisons of the calibration coefficients shown in Figure 3, based on four factors (ie, obesity, sex, race, and age) known to affect arterial stiffness [30-35] and therefore the PTT. The participants were split into nonobese and obese groups based on the guidelines from the National Heart, Lung, and Blood Institute of the National Institutes of Health defining a BMI ≥30 kg/m² as obese [21]. Thus, the nonobese group had a BMI ≤30. To assess differences due to age, we separated the participants into younger (aged ≤40 years) and older groups (aged ≥40 years). Statistical significance (\(P<.05\)) between demographic data for each cohort was assessed using an unpaired two-sample, two-tailed \(t\) test, as shown in Table 1.

For the demographic DBP calibration coefficient comparisons, a one-sample Kolmogorov-Smirnov test was used on each data point to test for normality, which determined that none of the data for the comparisons were normally distributed. Then, a Mann-Whitney \(U\) test (ie, Wilcoxon Rank Sum test) was used to assess statistical significance (\(P<.05\)) among the unpaired data.

For the PPG wavelength DBP estimation comparisons—only applicable to the second cohort population due to the differences in hardware used—first, a one-sample Kolmogorov-Smirnov test was used on each data point to test for normality, which determined that none of the data for the comparisons were normally distributed. Then, a Wilcoxon Signed Rank test was used to assess statistical significance (\(P<.05\)) among the paired data.

Results

Multimodal Engineering Mechanics of the SeismoWatch

The previous version of the watch, not shown in this work, consisted of a 3D printed case embedded with an accelerometer, PDs, and IR LEDs. All sensors were connected to a small external circuit box with straps for the participant to wear around the waist. The output of the analog accelerometer (ADXL354, Analog Devices) was connected to an analog front end (AFE) in the circuit box. To amplify the SCG signal and prevent saturation of the alternating current components owing to the varying direct current levels, the AFE separated the direct current and alternating current components using a low pass (f<1 Hz; G=−10 dB) and BPF (fpass=0.2 Hz-40 Hz) in parallel. An analog adder recombined both components. For PPG measurements, the cathode of the PDs (S2386-18K, Hamamatsu Photonics) was connected to transimpedance amplifiers configured as a low-pass filter (fc=12 Hz; G=110 dB) followed
by gain and filter stages (fpass=0.5-12 Hz; G=59 dB). Finally, the ECG was acquired by placing 3 copper dry electrodes on the wrist band of the watch with 2 on the inside in contact with the wrist and 1 on the outside to place the index and middle finger. The 2 on the inside act as the right leg drive electrode and the positive lead, whereas the outside electrode is the negative lead. All electrodes were connected to an AFE (AD8232, Analog Devices) for ECG measurements. A microcontroller (Teensy 3.6, PJRC LLC) sampled the output of the accelerometer, PPG, and ECG AFE at 1 kHz. An onboard SD card was used to store the raw data for postprocessing, and a 1.2 Ah lithium-ion rechargeable battery was used to power the system. All instrumentation details were adopted from our previous work, with minor revisions [19].

The updated hardware, pictured in Figure 5, added modalities of sensing (ie, a gyroscope), included multiple wavelengths of LEDs for comparison with IR, improved the form factor for comfort and ease of use, and featured embedded systems innovations leveraged in this study. A more thorough description of the revised hardware is available in our most recent work [20]. An example of the serviceable automatic LED current scaling algorithm, detailed in our previous work [20], is highlighted in Figure 5. This proved to be an integral part of enabling this work; by adaptively adjusting the LED drive current, we were able to prevent saturation and variable PPG signal quality caused by varying contact pressure and, more importantly, prominent differences in skin tone among participants.

**Figure 5.** Pertinent multimodal hardware block diagram and adaptive light-emitting diode (LED) scaling. Main board with ATSAM4LS8 microcontroller (μC), ADXL355 triaxial accelerometer, BMG250 triaxial gyroscope, and BME280 environmental sensor using the serial peripheral interface for fast communication supporting higher sample rates. Sensor board used to acquire wrist photoplethysmogram (PPG) and electrocardiogram signals. Automatic LED current scaling in operation during data collection: showing an increase in contact pressure and subsequent saturation of the photodiode, mitigated by an automatic decrease in LED current and overall consequential improvement in PPG signal quality. ECG: electrocardiogram; LED: light-emitting diode; PD: photodiode; PPG: photoplethysmogram; RLD: right leg drive; SD: Secure Digital; SPI: Serial Peripheral Interface.
Human Subject Studies in a Diverse Population

All applicable results are presented as mean (SD). Figure 2 illustrates the correlation and Blund-Alman plots for our wearable PTT-based BP estimation of MAP, DBP, and SBP across all participants (N=44). The MAD was 2.90 mm Hg, 3.39 mm Hg, and 5.36 mm Hg for DBP, MAP, and SBP, respectively. The mean RMSE was 3.41 (SD 2.01) mm Hg, 3.95 (SD 2.42) mm Hg, and 6.28 (SD 3.44) mm Hg for DBP, MAP, and SBP, respectively. DBP and MAP estimation had better 95% CIs than SBP at 7.99 mm Hg, 9.42 mm Hg, and 14.59 mm Hg, respectively. The mean Pearson correlation coefficient (PCC) was 0.67 (SD 0.16), 0.63 (SD 0.31), and 0.50 (SD 0.41) for PTT-based DBP, MAP, and SBP estimation, respectively.

The MAD for the individual study populations (first cohort=26 participants and second cohort=18 participants) was 2.69 mm Hg and 3.20 mm Hg, 3.21 mm Hg and 3.64 mm Hg, and 5.17 mm Hg and 5.63 mm Hg for DBP, MAP, and SBP estimation, respectively. The mean RMSE for the individual study populations (first cohort=26 participants and second cohort=18 participants) was 3.19 (SD 1.64) mm Hg and 3.73 (SD 2.48) mm Hg, 3.78 (SD 2.06) mm Hg and 4.18 (SD 2.90) mmHg, and 6.26 (SD 3.25) mm Hg and 6.32 (SD 3.80) mm Hg for DBP, MAP, and SBP estimation, respectively. The mean PCC for the individual study populations (first cohort=26 participants and second cohort=18 participants) was 0.69 (SD 0.15) and 0.65 (SD 0.17), 0.68 (SD 0.23) and 0.55 (SD 0.38), and 0.38 (SD 0.33) and 0.39 (SD 0.49) for DBP, MAP, and SBP estimation, respectively.

The MAD for the 19 Black participants was 3.18 mm Hg, 3.72 mm Hg, and 5.84 mm Hg for DBP, MAP, and SBP estimation, respectively. The mean RMSE for all 19 Black participants was 3.72 (SD 2.41) mm Hg, 4.29 (SD 2.86) mm Hg, and 6.69 (SD 4.03) mm Hg for DBP, MAP, and SBP estimation, respectively. The mean PCC for all 19 Black participants was 0.64 (SD 0.17), 0.53 (SD 0.38), and 0.37 (SD 0.48) for DBP, MAP, and SBP estimation, respectively.

The MAD for the 10 participants who were obese was 2.69 mmHg, 3.17 mm Hg, and 5.02 mm Hg for DBP, MAP, and SBP estimation, respectively. The mean RMSE for all 10 participants who were obese was 3.28 (SD 2.59) mm Hg, 3.69 (SD 3.00) mm Hg, and 5.71 (SD 4.18) mm Hg for DBP, MAP, and SBP estimation, respectively. The mean PCC for all 10 participants who were obese was 0.65 (SD 0.18), 0.52 (SD 0.48), and 0.39 (SD 0.58) for DBP, MAP, and SBP estimation, respectively.

Figure 3 depicts the boxplots of the DBP calibration coefficients from our estimation model, $K_1$ and $K_2$, for four different demographic factors known to affect arterial stiffness: obesity, sex, race, and age. The DBP $K_1$ and $K_2$ values for nonobese ($N=34$) versus obese ($N=10$) participants are 2.38 (SD 1.99) mm Hg/s versus 1.20 (SD 0.88) mm Hg/s and 61.02 (SD 18.03) mm Hg versus 74.31 (SD 5.14) mm Hg, respectively. The DBP $K_1$ and $K_2$ values for male ($N=25$) versus female ($N=19$) participants are 2.65 (SD 2.18) mm Hg/s versus 1.40 (SD 0.98) mm Hg/s and 60.16 (SD 20.64) mm Hg versus 69.14 (SD 8.29) mm Hg, respectively. The DBP $K_1$ and $K_2$ values for non-Black ($N=25$) versus Black ($N=19$) participants are 2.63 (SD 2.21) mm Hg/s versus 1.44 (SD 0.94) mm Hg/s and 60.66 (SD 20.29) mm Hg versus 68.49 (SD 9.98) mm Hg, respectively. The DBP $K_1$ and $K_2$ values for young ($N=31$) versus older ($N=13$) participants are 2.38 (SD 2.09) mm Hg/s versus 1.47 (SD 0.87) mm Hg/s and 61.96 (SD 19.03) mm Hg versus 69.00 (SD 9.22) mm Hg, respectively.

Both $K_1$ and $K_2$ were significantly different between the nonobese and obese populations ($P=0.045$ and $P=0.008$, respectively). The female $K_1$ values were significantly ($P=0.04$) lower than those of their male counterparts. The $K_2$ values for Black participants were significantly ($P=0.047$) lower than those of the other races.

For the participants in the second cohort—all Black—with whom we used the newer version of the hardware [20] that included green and red LEDs in addition to the IR, the PCC for DBP estimation was 0.38 (SD 0.34), 0.59 (SD 0.44), and 0.65 (SD 0.17) when using the green $\lambda=526$ nm, red $\lambda=660$ nm, and IR $\lambda=950$ nm wavelength PPGs for the distal timing reference, respectively. The PCC for the IR and red wavelength PPGs was significantly ($P=0.01$ and $P=0.048$) higher than that of the green wavelength PPGs. However, the corresponding mean DBP RMSE was 3.95 (SD 2.53) mm Hg, 3.11 (SD 2.33) mm Hg, and 3.73 (SD 2.48) mm Hg for green, red, and IR, respectively.

Discussion

Principal Findings

To the best of our knowledge, this is the first study to accurately estimate DBP and MAP using noninvasive PTT measurements acquired from a holistic population, with considerable differences in body fat percentage, melanin levels, and vascular stiffness associated with age and hypertension. Furthermore, our SBP estimation is sufficient to be clinically recommended for monitoring [29,36]. We demonstrated the reliability of a convenient method for estimating BP and observed that our calibration coefficients were significantly different in characteristic demographic groups known to have increased arterial stiffness. This work represents a necessary advancement toward remote monitoring for persons in MUAs by enabling wearable PTT-based BP estimation, including through the comprehensive evaluation of a watch-based form factor conducive to obtaining ambulatory BP measurements in low-resource settings.

Accurately Estimating BP in a Diverse Population Using a Multimodal Wearable Device

We demonstrated the performance of our wrist-worn PTT-based device when used to estimate BP within a diverse population over the course of multiple unique perturbations. Our results for MAP and DBP passed the acceptable benchmarks for the BP estimation error set by the IEEE standard on wearable cuffless BP estimation devices [29]. We were still able to achieve a reliable correlation between PTT and BP even with several demographic factors such as age, melanin levels, and BMI inherently influencing the measured optical-PPG and mechanical-SCG signals.
The DBP estimation remained the most accurate, similar to our previous studies [19,20]; the foot of the PPG waveform, used as the distal timing reference, indicates the arrival of the pulse wave during end diastole. Similarly, the SBP estimation continued to perform the worst, as the peak of the pulse wave is the fiducial marker of the PPG that occurs during systole; however, the peak is not frequently extracted, as its true timing can be confounded by wave reflection interference, leading to unreliable PTT estimates [12]. Recent studies have demonstrated that the PTT computed using the diastolic foot of the PPG outperforms that using the systolic maximum for both DBP and SBP [37].

The DBP RMSE was relatively similar at both low and high values of DBP, which indicates that the diastolic foot was a dependable timing reference for calculating the PTT, irrespective of inherent participant-specific differences in BP. Although our SBP estimation was just outside the acceptable limits set forth by the IEEE standard (ie, MAD=5.36 mm Hg vs 5.0 mm Hg) [29], this error translates to a grade B classification [29] and therefore would still be clinically recommended for monitoring SBP [36]. Furthermore, the SBP range studied in this work was greater than 100 mm Hg, substantially higher than that reported in previous studies in the literature for wearable cuffless BP estimation, and a combination of different perturbations was used to modulate BP. Using a single perturbation would have led to an improved correlation [12,14], as in our previous work where we had only used exercise [19]. However, a comprehensive evaluation of this methodology would be incomplete without a procedure consisting of a wide variety of perturbations with different known physiological responses and pathways to modulate BP [14]. In addition, as noted in Figure 1, the exercise perturbation did not apparently produce a marked difference in BP due to several factors: (1) technical limitations in rapid calibration for the reference measurement (ie, finger-cuff continuous BP) and increased motion artifacts following exercise led to a greater percentage of beat removal in the early exercise section than any other task and (2) exercise does not necessarily consistently modulate BP in a predictable manner due to differences in participant-specific vasoactivity and contractility [12,38].

Only the DBP was examined for further analyses conducted below because, as previously mentioned, the distal timing reference used in this work (ie, the foot of the PPG waveform) occurs during diastole and therefore provides the most reliable estimate of DBP out of the three BP components [12]. The dependability of the diastolic foot and our robust DBP estimation were necessary before performing in-depth analyses with confidence. Although elevated SBP is considered to be the greatest predictor of future cardiovascular risk [39,40], elevated DBP has nonetheless been shown to independently increase the risk of subsequent cardiac events [39,41]. In addition, DBP is a greater contributor to MAP, which in older patients with isolated systolic hypertension, when compared with an equivalent increase in pulse pressure, has been shown to be a comparable independent predictor of both stroke and all-cause mortality [42]. Finally, DBP has been shown to be a more significant predictor than SBP of new-onset hypertension in adults aged ≤50 years [40,43-45]. This suggests that accurate DBP estimation using a wearable device can efficiently be used to incentivize people to make healthy lifestyle modifications earlier in life, central to the World Health Organization’s effort to reduce the global prevalence of hypertension [46].

**Essential Device Novelties Enabling Reliable PTT Computation**

For the first time, we demonstrated that noninvasive PTT measurements are reliable estimators of BP across a wide range of skin tones and BMI. Both DBP and MAP estimation for the 10 participants who are obese and 19 Black participants in this study were well under the IEEE requirement [29]. This was enabled by the highly sensitive hardware, multisensor approach, and automated LED current scaling that our custom wearable device offers [20]. The PPG array and adaptive LED current scaling allowed us to automatically mitigate poor signal quality issues due to misplacement, inherent differences in skin tone, and applied pressure that typically corrupts PPG signals. However, the most integral components of our PPG hardware were the IR wavelength LEDs.

We leveraged longer wavelengths of light for deeper penetration into the tissue to robustly acquire the PPG signal from arteries located deeper than the cutaneous vascular bed [12]. Cutaneous arteries are greatly affected by the changes in vascular tone expected from the perturbations we used to modulate BP herein (ie, cold pressor and exercise). Furthermore, as IR PPGs are more susceptible to motion artifacts than lower wavelength ones [12,47], our PPG-first signal quality assessment not only avoided these motion artifact corrupted waveforms because of their low SNR but also avoided moments where the SCG quality would naturally suffer as well. However, even the red PPGs had a considerably larger SD in their PCC than the IR PPGs, possibly because the IR wavelength, when compared with red, is less sensitive to the oxygen content of hemoglobin and has approximately half the skin absorption coefficient in Black individuals [12,27]. Despite statistically significant differences in the PCC using IR and red PPGs rather than green PPGs, the actual DBP RMSEs were comparable. This implies that when using the green PPGs for participants with a low PCC, our signal quality assessment algorithm removed beats with greater BP variability, resulting in a lower SD of DBP and consequent RMSE. Although even green wavelength PPGs have demonstrated the ability to reliably extract heart rate across a wide variety of skin tones [48], our data suggest that these shorter wavelengths cannot be used to dependable compute the PTT in a diverse population. In addition, although unconventional, our watch was placed on the ventral side of the wrist, which allows for both higher quality, convenient SCG acquisition and enhanced PPG SNR due to viable access to the radial artery and less melanin content than the dorsal side [49]. Therefore, existing smartwatches, beginning to slowly incorporate cuffless, noninvasive BP methodologies, may face even greater difficulties in achieving accurate PPG measurements across a broad range of skin tones.

Finally, our physiologically inspired PPG selection algorithm—to first select the PPG signals with the greatest systolic upstrokes—had an important role in reducing the BP estimation error. PPG waveforms with greater systolic upstrokes...
(ie, maximum derivative of the PPG waveform) offer improved PTT estimates and are key indicators of BP stemming from larger, more pulsatile, elastic arteries with greater distensibility [12,50]. In addition, several recent machine learning (ML) approaches to use the PPG signal for BP estimation have shown that the systolic upstroke is one of the most important features of the waveform [51,52]. Hence, the selection algorithm, by extracting information from these more reliable and clinically important arteries, was a central part of our ability to notice the demographic differences in arterial stiffness rooted in our calibration coefficients.

**Calibration Coefficients Capture Demographic Differences in Arterial Stiffness**

We observed that the participant-specific calibration coefficients used in the standard linear PTT-BP estimation model for DBP, shown in equation 1 (ie, $K_1$ and $K_2$), are significantly different between subpopulations with large variations in demographic factors known to affect arterial stiffness. We selected the four demographic categories (ie, obesity, sex, race, and age) based on the literature, emphasizing these as major determinants of differences in arterial stiffness and therefore risk factors for hypertension [16,30-33,35,53,54].

The $K_1$ value (ie, the slope of the line of best fit) is indicative of the underlying baseline vascular stiffness, whereas $K_2$ (ie, the intercept) represents the inherently correlated bias in baseline BP [12,55,56]. At the same BP, persons with greater arterial stiffness have inherently faster pulse wave velocities (PWVs) and therefore shorter PTTs than persons with normal arterial stiffness [12]. The $K_1$ value mitigates these differences in PTT-based estimation by capturing the intrinsic participant-specific arterial stiffness to output similar BP values. Therefore, with increasing arterial stiffness, we expected to find a lower $K_1$ value and a higher $K_2$ value, as observed in the PWV literature [55,56].

Obesity was the only comparison for which the differences in the $K_1$ and $K_2$ calibration coefficients were statistically significant. This coincides with the literature stating that obesity is one of the greatest age-normalized risk factors and contributors to arterial stiffness [57]. Otherwise, only the $K_1$ values in the sex and race comparisons were statistically significant between the groups. Although it has been shown that both females and Black individuals have greater arterial stiffness than similar-age males and White individuals [31,34,35], these two comparisons should be re-evaluated after increasing our recruitment. Approximately 47% (9/19 participants) of both the female and Black population were also obese. The age comparison was not statistically significant, although the older population followed a similar trend of a lower $K_1$ and higher $K_2$. This finding is not surprising, as significant differences in arterial stiffness and substantial augmentations in arterial remodeling are typically examined in participants aged ≥50 years [32,58].

**Limitations and Future Work**

**Refining Population Demographics and Investigating PTT, $K_1$, and $K_2$ as Potential Digital Biomarkers of Arterial Stiffness**

Overall, although this data set captured a more representative population in the range of end users for which consistent BP monitoring is recommended [59], our PTT-based device should be further tested in an exclusively older (ie, age >50 years), morbidly obese (ie, BMI >40 kg/m²), and hypertensive population—with even distributions across sex, race, and skin tones along the Fitzpatrick scale—to truly understand the limits of this technology and supplement the findings herein.

Early vascular remodeling due to the demographic factors investigated in this work, not to mention socioeconomic factors affecting MUAs [15,16], predispose individuals who are obese and Black individuals to greater lifetime cardiovascular risk [30,35,57,60,61]. Therefore, future PTT-based BP estimation studies should closely monitor the calibration coefficients, $K_1$ and $K_2$, as potential intermediate digital biomarkers for longitudinal monitoring and the comparison of arterial stiffness among different persons [7]. Eventually, even PTT measurements, as PWV is already an independent predictor of arterial stiffness [62], may indicate subclinical differences in vascular resistance due to early stage arterial remodeling, the main precursor to hypertension [32].

**Reducing the Burden of Calibration**

Consistent recalibration poses a practical concern for PTT-based BP estimation. Hence, future studies should focus on evaluating the timeframe for which participant-specific calibration curves can reliably estimate BP and whether interparticipant and population-level curves can be sufficient. However, given the value of interpreting the calibration coefficients presented in this work, caution should be exercised due to the trade-off of sacrificing this potential usefulness when using generalized interparticipant models. Furthermore, the individual effects of the perturbations used to modulate BP in this experiment should be scrutinized, along with other exercises shown to substantially change BP [63-65]. The goal is to use perturbations that can consistently be leveraged to increase the dynamic range of BP measurements for calibration—critical to achieving optimal estimations at home in our previous work [20] and are achievable in low-resource settings.

**Leveraging ML and Hardware Advancements for Robust SCG AO Detection**

Similarly, to the instrumental role of the physiologically inspired PPG selection algorithm in this work, further exploration into automated SCG fiducial point detection algorithms may help extract the most informative SCG signals. Specifically, the SCG can be greatly affected by inaccurate placement of the watch; however, recent advancements using ML techniques have shown that the SCG waveform is modulated in a predictable manner during these placement inaccuracies [66]. Therefore, by interpreting these findings, one might be able to convert the measured SCG to the archetypal SCG or use a template-matching localization approach [67] for each
participant before extracting salient features from the optimal waveform.

In addition, annotating the exact AO point can be challenging because the signal not only has appreciable interparticipant variability, especially in a population with considerable differences in BMI, but can also be corrupted by motion artifacts. Although our technique for extracting the AO point has led to a high correlation between PTT and BP, in both our recent work [20] and this one, for a few sessions, we manually annotated the SCG to impose realistic constraints for the range of the pre-ejection period (PEP) and selected a consistent morphological peak across all tasks per participant. Eventually, robust identification of this timing reference is necessary for reliable automatic PTT computation, as the main advantage of using the PTT over the pulse arrival time (ie, the time from the R-wave of the ECG to the diastolic foot of the PPG) for BP estimation is its ability to account for changes in the nonnegligible cardio-electromechanical delay, that is, the PEP [12,68]. Furthermore, examining the other sensor data available at our disposal, such as filtering the SCG in a higher bandwidth (ie, fpass=30-125 Hz) to retain the phonocardiogram signal indicative of valve closures, using the three-axis gyrocardiogram or simply the other axes of the SCG, could prove to help with improving PEP estimation as shown in previous work [19,69].

Conclusions
We have demonstrated that a wrist-worn device, using noninvasive PTT estimates, can reliably and conveniently track BP in a diverse population. Leveraging the ubiquity of wearable devices can empower users to make healthy lifestyle modifications such as exercise, which can contribute to a significant reduction in arterial stiffness [30,70] by providing consistent feedback on progress [71-73]. In addition, digital health technologies that accurately estimate BP could potentially be used to titrate BP medications for patients with hypertension from the comfort of their homes [7,74]. In addition to these broader impacts, the knowledge gained from this study—especially when combined with the advent of low-profile, flexible electronics capable of robustly detecting physiological biosignals [75-78]—represents a significant step toward the unobtrusive monitoring of BP in ambulatory settings and health equity for persons in MUAs.

Acknowledgments
This work was funded in part by the National Institute of Health, National Institute of Biomedical Imaging and Bioengineering under grant 1U01EB018818-01, by the National Institute of Biomedical Imaging and Bioengineering Point-of-Care Technologies Research Network, Atlanta Center for Microsystems-Engineered Point-of-Care Technologies under grant 1U54EB027690, and by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1TR002378.

Authors’ Contributions
VGG codeveloped the newer version of the hardware, performed human subject studies on the second cohort, analyzed the collected data, and coprepared the manuscript. AMC developed the older version of the hardware, codeveloped the newer version of the hardware, co-conducted human subject studies on the first cohort, and assisted in advising the analysis and editing of the manuscript. HJ co-conducted human subject studies on the first cohort and assisted in both human subject studies on the second cohort and editing of the manuscript. AVS assisted in both conducting human subject studies on the second cohort and editing of the manuscript. DC assisted in both participant recruitment for the second cohort and editing of the manuscript. LNJ assisted in advising the study, participant recruitment for the second cohort, and editing of the manuscript. OTI guided the study and coprepared the manuscript.

Conflicts of Interest
OTI is a cofounder of and scientific advisor at Cardiosense, Inc, and a scientific advisor at Physiowave, Inc. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References


https://mhealth.jmir.org/2021/8/e27466


Abbreviations

- ABBP: arterial blood pressure
- AFE: analog front end
- AO: aortic valve opening
- BP: blood pressure
- BPF: bandpass filter
- DBP: diastolic blood pressure
- ECG: electrocardiogram
- IEEE: Institute for Electronics and Electrical Engineers
- IR: infrared
- LED: light-emitting diode
- MAD: mean absolute difference
- MAP: mean arterial pressure
- ML: machine learning
- MUA: medically underserved area
- PCC: Pearson correlation coefficient
- PEP: pre-ejection period
- PPG: photoplethysmogram
- PTT: pulse transit time
- PWV: pulse wave velocity
- RMSE: root mean square error
- SBP: systolic blood pressure
- SCG: seismocardiogram
- SNR: signal-to-noise ratio
General Practitioners' Perceptions of the Use of Wearable Electronic Health Monitoring Devices: Qualitative Analysis of Risks and Benefits

Lucia Volpato¹, MSc; María del Río Carral¹, PhD; Nicolas Senn², MD, PhD; Marie Santiago Delefosse¹, PhD

¹Research Centre for Psychology of Health, Aging and Sport Examination, Institute of Psychology, University of Lausanne, Lausanne, Switzerland
²Department of Family Medicine, Centre for Primary Care and Public Health, University of Lausanne, Lausanne, Switzerland

Corresponding Author:
María del Río Carral, PhD
Research Centre for Psychology of Health, Aging and Sport Examination
Institute of Psychology
University of Lausanne
Géopolis Quartier UNIL - Mouline
Lausanne, 1015
Switzerland
Phone: 41 21 692 32 67
Email: maria.delriocarral@unil.ch

Abstract

Background: The rapid diffusion of wearable electronic health monitoring devices (wearable devices or wearables) among lay populations shows that self-tracking and self-monitoring are pervasively expanding, while influencing health-related practices. General practitioners are confronted with this phenomenon, since they often are the expert-voice that patients will seek.

Objective: This article aims to explore general practitioners’ perceptions of the role of wearable devices in family medicine and of their benefits, risks, and challenges associated with their use. It also explores their perceptions of the future development of these devices.

Methods: Data were collected during a medical conference among 19 Swiss general practitioners through mind maps. Maps were first sketched at the conference and their content was later compared with notes and reports written during the conference, which allowed for further integration of information. This tool represents an innovative methodology in qualitative research that allows for time-efficient data collection and data analysis.

Results: Data analysis highlighted that wearable devices were described as user-friendly, adaptable devices that could enable performance monitoring and support medical research. Benefits included support for patients’ empowerment and education, behavior change facilitation, better awareness of personal medical history and body functioning, efficient information transmission, and connection with the patient’s medical network; however, general practitioners were concerned by a lack of scientific validation, lack of clarity over data protection, and the risk of stakeholder-associated financial interests. Other perceived risks included the promotion of an overly medicalized health culture and the risk of supporting patients’ self-diagnosis and self-medication. General practitioners also feared increased pressure on their workload and a compromised doctor–patient relationship. Finally, they raised important questions that can guide wearables’ future design and development, highlighting a need for general practitioners and medical professionals to be involved in the process.

Conclusions: Wearables play an increasingly central role in daily health-related practices, and general practitioners expressed a desire to become more involved in the development of such technologies. Described as useful information providers, wearables were generally positively perceived and did not seem to pose a threat to the doctor–patient relationship. However, general practitioners expressed their concern that wearables may fuel a self-monitoring logic, to the detriment of patients’ autonomy and overall well-being. While wearables can contribute to health promotion, it is crucial to clarify the logic underpinning the design of such devices. Through the analysis of group discussions, this study contributes to the existing literature by presenting general practitioners’ perceptions of wearable devices. This paper provides insight on general practitioners’ perception to be considered in the context of product development and marketing.

(JMIR Mhealth Uhealth 2021;9(8):e23896) doi:10.2196/23896
KEYWORDS
mHealth; wearable devices; health wearables; activity trackers; health monitoring; self-tracking; general practitioners; mind maps; qualitative research; health psychology

Introduction
Over the last few years, the development of new health-related technologies has been particularly rapid and prosperous [1,2]. In particular, wearable electronic health monitoring devices (henceforth referred to as wearable devices or wearables) are designed to support health management by monitoring bodily vital signals, as well as tracking an individual’s activity and habits [3,4]. User-friendliness and gamification play an important role in the appealing and engaging design of wearable devices, which are often paired with mobile phone apps [5]. Their association with mobile phones is at the origin of the terms mHealth (or mobile health) and continuous connection with wireless devices has been associated with self-surveillance and self-tracking mentality [6]. Yet, the difference between technologies targeting lay populations and the ones designed to monitor specific medical conditions is not always clearly defined [3]. In both fields—health promotion and intervention—the logic underpinning wearable devices’ marketing aims to improve health by promoting behavior change through self-tracking, on the basis of feedback mechanisms [4,7,8].

More specifically, the marketing discourse on wearable devices is strongly based on the promise of benefits regarding personalized health management programs that claim to promote patient’s self-responsibility and autonomy [6] by fostering a less hierarchical relationship between the user and the health professional [7]. Furthermore, wearable manufacturing companies have been sponsoring large-scale medical studies [9,10] with little consideration of the consequences associated with the introduction of self-tracking devices in everyday life [11]. This adds further complexity to the picture because traditional health care systems risk being detoured by other—often profit-led—motives [11]. While it has been argued that wearable devices may contribute to fostering users’ autonomy [12], self-tracking has not always been associated with empowerment [13,14]. In fact, wearable use could hinder the autonomy of users, who would increasingly rely on these devices for daily, health-related decision making [4]. According to Andreassen and colleagues [15], feelings of domestication and resistance co-exist in the user–object relationship. Furthermore, wearable devices have been associated with high abandonment rates after only 6 months of use [16]. It is, therefore, clear that, beyond manufacturers’ promises, the concrete use of these technologies in the health sector is subject to negotiation processes that depend on complex dynamics regarding the object–user relationship and the doctor–patient relationship [17,18].

Activity trackers have affected how treatment, visits, and health management are established during health consultations [19]. Their growing presence may constitute a positive addition to the relationship between patients and their general practitioners (GPs), because the devices could support effective transmission of health-related information [19]. Historically, the introduction of technological devices has defined medical practices [20], and wearable devices are no exception. The notion of the Quantified Self—associated with activity trackers—is pervasively shaping health norms through self-surveillance across life domains [21]. Moreover, the contemporary trend of healthism, which values individual responsibility and surveillance in health management, has expanded over the past decades [22]. This societal discourse constitutes a fertile ground for the production and marketing of wearables.

GPs are an inherent part of the digital health revolution, given their role as health experts, citizens, and sometimes, wearable consumers [23]. The intention to adopt activity trackers and recommend their use are shaped by views and attitudes that individuals may have toward such technologies [24,25]. In light of these findings, it is necessary to further explore perception and views on wearables among GPs, because such devices may change existing medical practices and contribute to shape new ones. This is particularly relevant within the Swiss context, where, similarly to other European countries [26,27], GPs are in the front-line when patients access the health system. In this sense, often GPs are the first health professionals to interact with many wearable device users.

Some studies have investigated health professionals’ attitudes toward technologies that are specific to chronic health conditions, such as epilepsy [28], asthma [29], arthritis [30], and other chronic diseases [31]. Another body of literature on GPs’ perspectives has examined their experience with a wide array of eHealth innovations, beyond the specific use of wearables [32]. With respect to wearables, few studies have explored GPs’ attitudes with a set of predefined topics using individual semistructured interviews [33,34] and web-based surveys [19]. We further investigated GPs’ perspectives and contribute to the existing literature on the role of wearables in family medicine. We aimed to explore how GPs perceive wearable devices—both for health promotion and clinical use—in the context of their medical practice, by focusing on perceived benefits and risks. To do so, we used an innovative qualitative methodology with mind maps to analyze group discussions that took place during a medical conference on family medicine. Mind maps have been described as being particularly suitable for analyzing group discussions in the field of health care [35]. We discuss salient elements to consider in the future development of these technologies.

Methods
Research Context and Sampling
This study’s aims were defined by the authors in collaboration with health psychologists and physicians working at the University of Lausanne, as well as GPs working in the French-speaking regions of Switzerland. Data were collected in a symposium—New Technologies in Family Medicine—that took place in Switzerland in 2018, as part of a medical conference, mainly targeting GPs. Given the qualitative nature
of our study, we aimed at in-depth understanding and contextualization of data, rather than generalization. For data collection and analysis, we followed the quality criteria for qualitative research defined in the field of health psychology [36,37].

We used convenience sampling: 19 GPs (7 female and 12 male) working in family medicine in the French-speaking regions of Switzerland. Participants were informed about the symposium’s goal, involving the definition of potential research perspectives regarding the use of wearables in family medicine, based on their perception. GPs were formally informed that group discussions would be recorded for further analysis, and oral consent was obtained. Under Swiss ethical regulations, no written consent was required as no biomedical information was collected.

Regarding participants’ background on wearables use, the vast majority reported not having actively introduced them in their clinical practice and that any discussion on wearables was usually initiated by patients themselves. Cited examples included patients who monitored their menstrual cycle through apps and the tracking of physical activity through smartwatches. Only 2 GPs reported that they used mobile apps for sleep monitoring and for diagnostic procedures via symptom-input mechanisms. It was highlighted that such apps were offered by official health providers. Some participants were familiar with such technologies through life experiences beyond their professional practice as GPs. With respect to their personal use, 1 GP reported using a smartwatch for performance monitoring during sports training. In contrast, another participant reported deactivating all tracking functions on their mobile device because of mistrust of the app’s use of personal information.

**Group Discussions**

GPs were enrolled in group discussions on smartwatches, wearable devices, and health apps. These topics had been previously defined, so that participants could join any group discussion, based on their personal interests. Each group (average of 6 participants per group) was moderated by 1 health psychologist and 1 GP. Discussions lasted approximately 1 hour and were audiorecorded.

In each group, participants were invited to briefly present themselves and were informed that the discussions were going to address the role of multiple mHealth technologies in family medicine. The 2 moderators introduced a brief explanation of the specific discussion topic (either smartwatches, wearable devices, or health apps, each discussed within a different group). These 3 groups of technologies were chosen for their high interconnectedness and interdependence within the broad category of mHealth [6]. For instance, smartwatches may be considered a wearable device category and are often supported by a smartphone app for data collection and analysis [5]. Participants were asked to discuss the following questions within each group: (1) What is the role of such technologies within your practice, according to your experience? (2) What risks and benefits do you identify in relation to such technologies? (3) Which challenges would you associate to the concrete use in your professional practice?

While the differences with other methods of data collection (ie, group interviews or focus groups) may be subtle, group discussions are less bound to structured interview guides and the emerging discussion topics often result from the interactions among group members, rather than from detailed predetermined questions grid [38]. Group discussions are also particularly suited for data collection among individuals who belong to the same group, for example, a professional category [38,39]. Moreover, the role of the moderators in group discussions is to provide topics to stimulate interactions among participants in a nondirective way [38].

To facilitate participants’ interactions, moderators took part in the discussions and summarized the material produced from their group. Summaries were approved and validated by participants of each group, resulting in specific descriptive reports [40-42]. Participant validation has shown to be a critical stage of qualitative research, because it provides more solidity and pertinence to the collected data [37].

**Mind Maps**

The potential of mind maps has been recently underlined for their use as research methods for data collection and analysis in the field of health [35]. A mind map can be defined as “a diagram used to represent concepts, ideas or tasks linked to and arranged radially around a central key word or idea [35].” Mind maps present information in a hierarchical way [43] through a synthetic visually engaging format [44]. Beyond their use in data collection, they can facilitate the data analysis by identifying and representing thematic and conceptual patterns in a nonlinear form [45], while showing associations between ideas and topics [43]. During the symposium, an overt participant–observant researcher (the main author) circulated among the different discussion groups, taking notes of part of the ongoing discussions, and sketching preliminary mind maps.

Through participant observation, further notes were taken to identify the links between the raised concepts and capture the contextual dimensions of verbal exchanges [46].

Inspired by Burgess-Allen and Owen-Smith [35], we considered a separate mind map for each of the themes used for the 3 groups canvases: benefits, risks, and some insights for the future. In addition, mind maps drawn during the group discussions revealed a recurring substructure in the discussion of the themes: doctor–patient relationship, patient–device relationship and GPs’ broad concerns, and final mind maps reflected this structure. The content of each mind map was then assembled inductively, and narrative contents were systematically compared by assessing their semantic similarities and differences. The proceedings and the audiorecordings from the group discussions facilitated integration of any missing relevant information from the preliminary notes and mind map drafting. This was particularly helpful to confirm the accuracy of the qualitative material and the mind map analysis. Finally, the 3 mind maps were compared to one another to identify common issues raised across group discussions regarding the potential benefits and risks, as well as, some insights for the future of wearable devices. This technique allowed for data analysis according to a theme-categories-subcategories structure, analogous to inductive thematic analysis, where mind mapping
is a preliminary stage [47]. In this study, mind maps were first sketched on paper for conceptualization purposes and were later digitally reproduced with FreeMind software (version 1.0.1).

**Results**

**General**

Regarding GPs’ perceptions of smartwatches, wearable devices, and health apps in family medicine, the first mind map (Figure 1) summarizes the perceived benefits of wearable devices. Here, participants used the conditional verb tense, which suggested that their arguments often applied to hypothetical scenarios and specific conditions. The second map (Figure 2) shows perceived risks that wearable devices usage and promotion may entail. The third map (Figure 3) presents insights that should be considered in the future production and use of wearables.

**Figure 1.** Mind map presenting the benefits that general practitioners associate with wearable use.

**Figure 2.** Mind map presenting the risks that general practitioners associate with wearable use.
Benefits That GPs Associate With Wearable Use

**General Benefits Associated With Wearable Devices’ Characteristics**

Overall, GPs positively evaluated wearable devices that were considered user-friendly to be used in a variety of health situations, thereby representing an attractive solution for different populations. According to GPs, wearable devices could be easily used by both patients and health professionals due to their simple, intuitive designs. In particular, doctors appreciated devices whose parameters can be easily adapted and personalized to fit a patient’s personal health characteristics. In fact, these features could enable a personalized approach to health management. Moreover, GPs considered that wearables could benefit those who wished to have regular feedback about their personal health and measure their physical activity through performance monitoring. According to GPs from the 3 group discussions, the widespread use of wearables could allow for large-sample data collection, which would be especially useful for medical research. In fact, a potential benefit concerned the strong statistical power that wearable devices could enable through research conducted among a high number of users.

**Patient–Wearable Relationship**

Wearable devices were also discussed in relation to patients’ empowerment. According to different participants across the 3 group discussions, wearables could help raise awareness among patients on their overall health condition. Due to the feedback and reward mechanisms that define certain devices, wearables could train users to make informed health decisions. In this sense, results reveal that wearables may enhance patient’s self-responsibility and be a concrete partner for health promotion. In fact, wearables were described as a potential mean for behavior change through the implementation of new health behaviors, through consistent self-management. As stated by a participant:

*Smart watches could motivate people for being more active, because there’s a certain degree of satisfaction in seeing the [step] counting going up.*

Moreover, some GPs asserted that wearables could help patients to keep track of their medical history and develop an unprecedented awareness regarding their own bodies. For instance, wearable users could establish links between their own feelings and the data provided by the device. In their views, wearables may, in fact, concretely support self-education.

**GP–Patient Relationship**

Wearables were described as potential health care partners in the light of the rapid information-transmission processes that such devices enable, allowing patients to have a more central role in their health management. Furthermore, the time-efficient data exchange between patients and health professionals could be beneficial for the management of certain health conditions, such as epilepsy or cardiac diseases. These health conditions were mentioned given the capacity of wearables to transfer data in real time. From GPs’ views, this represented a helpful feature to prevent seizures or enable screening procedures. Wearables were also perceived as potential allies for telehealth, since these devices could help GPs reach patients living in geographically remote areas.

One GP affirmed:

*We don’t need to make a trip to their place if they can measure themselves blood tension, glycemia, the...*
heartrate at home and then transfer information via the internet.

This was considered a particularly important aspect in the Swiss context, where, due to alpine geography, certain patients may need to travel long distances to receive medical care. Wearables were described as a possible means of connection within the health system between different professionals and caregivers, as well as a useful solution for those who do not have family or social support in daily health care.

**Risks That GPs Associate With Wearable Use**

**General Risks Associated With Wearable Devices’ Characteristics**

Although wearables’ potential advantages were thoroughly discussed, the debate raised several potential risks linked to their use. For instance, participants expressed fears regarding the unclear degree of reliability and accuracy of commercial wearables that are increasingly available. GPs shared their professional experience in stressing that devices that are able to accurately record biometrics are often more expensive and more complex. Therefore, their use requires specific training and an understanding of data collection. Furthermore, GPs highlighted the scarcity of scientific studies on wearables’ validity and reliability. To them, this represented an obstacle that impeded them from actively promoting the use of instruments that are not supported by scientific evidence.

Furthermore, the lack of accurate information regarding the management of biomedical data by manufacturers was considered to be a serious danger for patients. The risk of compromising an individual’s privacy was a major concern with respect to these technologies since, as affirmed by a participant:

> Third-party use of personal data is still very poorly regulated.

In an era where personal data is becoming widely commodified, several industries can profit from wearable use without being genuinely concerned by users’ health. According to GPs, the promotion of wearables could thus imply that financial profit is valued over health.

The role of health insurance companies and their possible relation with the wearable device industry were also considered in group discussions, since the Swiss health system relies on compulsory private not-for-profit health insurance companies. According to some participants, the collaboration between the 2 stakeholders could encourage the development of incentive-based medicine rationalities, that is, of a health philosophy, by which patient behavior could be rewarded or punished by insurance companies as a consequence of the degree of behavior compliance determined by the wearable’s design. For instance, health insurance companies could be inclined to reward so-called good users for having achieved the health aims set by the device or punish so-called bad users who have failed to do so. This mechanism deserves better attention, because it may also potentially reinforce health inequalities from a socioeconomic perspective. These reflections raised a debate on which ethical principles should underpin family medicine, as well as, on the rights and responsibilities of each actor.

**Patient–Wearable Relationship**

Regarding user–device relationship, GPs argued that continuous self-monitoring could stress an overmedicalization of life, generated by an excessive intellectualization of the user’s physical condition. From their perspective, the prioritization of self-monitoring practices in the field of health would inevitably confront users with the paradoxes of our culture: while health-related practices are aimed at reducing stress in daily life by helping users to slow their pace, wearables would be the symbol of a society that values rapid information exchange, and hence, would contribute to reinforcing a fast-paced life. In this context, GPs raised the risk of overestimating the value and the role of individual data in coming to conclusions about a person’s general state of health. According to GPs, the continuous measurement of biomedical information appeared to be also potentially anxiety triggering. A participant feared that

> People may end up spending more time preoccupying about their health instead of living.

Patients with apprehensive personalities could particularly risk developing overmonitoring practices, to the detriment of their mental health.

Regarding the level of trust toward certain devices, some participants feared that wearables would induce the users to gradually feel estranged from their body. In this sense, wearables could provide digital information that does not correspond to the users’ subjective perceptions on their own body and health. This mismatch between the wearable’s feedback and the embodied sensations could induce the users to mistrust their subjective perception and thus feel disconnected from their own body. According to GPs, this risk would also interfere with the principle of patient autonomy, whose appraisal of their own body would therefore mainly depend on the wearable verdict instead of their own perception. In this scenario, the patient and the caregiver would need to invest even more resources to set up a process of bodily re-appropriation. From the participants’ view, these risks would result in a reversed power relationship with the device that could be dangerous and that should be avoided. A participant feared a

> Very likely tendency towards over-training during a sport session

while seeking positive feedback from an activity tracker. An important element of the debate concerned wearable data interpretation. Participants agreed on the fact that, given a decrease of exchanges between users and health professionals, the former would be confronted to increased uncertainties regarding the interpretation of their personal physiological values, which is considered to be as dangerous for a user’s health. In fact, on the basis of the wearable data, patients could be tempted with self-diagnosis or self-medication solutions, something that ought to be avoided, especially when medical expertise is essential.

**GP–Patient Relationship**

In the light of the increased production of patient-specific medical information, participants highlighted the risk of devoting their working hours to time-consuming data analyses.

https://mhealth.jmir.org/2021/8/e23896
We can collect plenty of data, but then what will do about them? wondered a participant. In fact, the instantaneous nature data transmission could intensify the expectation of an immediate reply from health professionals, which would amplify the pressure on GPs’ daily practices, without any verified benefit for concerned patients. Participants also expressed the danger of inconsistency between the information recorded by wearables and data provided by other devices measuring biometrics. This type of divergence could, in fact, entail a progressive mistrust on the part of patients regarding the information provided by other the medical instruments, and GPs may suffer from credibility loss.

Insights for the Future of Wearables According to GPs

**GPs’ Professional Needs**

Participants expressed a sense of inevitability toward the introduction of wearable devices into contemporary medical practices within the Swiss context, regardless of the outcomes of current research in the field. In this sense, several GPs highlighted the urgent need, and their personal interest, to become more involved in the development of wearables. In their view, a synergy between producers and health professionals is necessary to enable the design of beneficial instruments. GPs also expressed the wish to receive training to better understand how these technologies work (especially concerning data collection and storage of information) and to be better informed on the news regarding the health-wearable market. GPs particularly valued the importance of understanding patients’ concerns, of identifying patients’ health needs, and of answering to their questions.

When worried patients come [to the consultation] with such instruments, what do we do? We have to give them an answer,

stated a participant. They also expressed the importance to them as professionals to be aware and to remain critical of the usefulness these devices in health care.

**Future Developments**

Overall, participants from the 3 group discussions raised the urgent need for firm legislation to guide future design, production, and marketing of wearables. In particular, they showed a high degree of mistrust regarding the confidentiality of their patients’ data that wearable companies could guarantee. To them, a priority for the future would therefore, be legislation to ensure data protection, as well as an overt policy to safeguard the rights of users, because the latter are in a vulnerable position within the health care system. Furthermore, the legislation should also concern the homologation of wearables’ internal parameters. Alongside scientific research, these measures could help to produce more valid and reliable devices for health self-management. With respect to the role that wearables should have, participants expressed the vision of wearables as partners that could help improve the care of patients. As emphasized by certain participants, well-oriented and accurate feedback is central to medical practice, because it can facilitate the learning process. In this sense, the analyses showed that wearables would not only be a tool for information transmission but could solidify partners in the promotion of behavior change. Moreover, according to the participants, the use of wearables should be limited to patients suffering from certain health conditions (although no examples were explicitly given during the symposium), instead of monitoring healthy people. In this sense, health would be achieved by conducting a digital-free, slow-paced life, where the person is not dependent on self-tracking devices.

**Open Questions**

According to the participants, several questions remained unanswered yet would be worth exploring. For instance, some GPs wondered how to assess which patients would best benefit from wearable use, and on the contrary, which patients would feel disconnected from their own body, that is, lose their personal autonomy toward the interpretation of their own embodied feelings. In this sense, a participant asserted that

To be useful, such products could be adapted to the patient’s profile in the future.

GPs also wished to know how wearables may affect user’s sense of responsibility. More generally GPs also raised the following issue: How will health wearables affect the GP–patient relationship? These questions closed the debates across groups, highlighting the need of additional analyses before establishing any further statements on the role of health wearables for our contemporary societies.

**Discussion**

**Principal Findings and Comparison With Prior Work**

GPs play a crucial role in the health care system by promoting and prescribing specific health practices. We aimed to explore their perceptions on wearables in the context of family medicine, by directly addressing the risks and benefits associated with these technologies and reflecting on possible future developments in the field of health care. The methodology adopted in this research was qualitative. This perspective was particularly suited to generate exploratory and contextualized knowledge. While group discussions allowed us to capture GPs’ views throughout their spontaneous interactions, mind maps enabled an iterative and efficient process of data collection and analysis given our research setting [35].

**Wearables as Information Providers**

The effects of digital health wearables on the doctor–patient relationship appeared to be both beneficial and risky, highlighting their ambiguous potential. While wearables were viewed as suitable for information transmission, coordination, and general illness management, GPs also feared that these technologies would put increased pressure on their role and expertise as health professionals. Indeed, GPs anticipated longer consultations that would be dedicated to data analysis and data interpretation stemming from patients’ wearables. From this perspective, our results confirm those of recent studies showing that wearables are considered to be particularly useful for information-transmission and general illness management but that time-consuming data interpretation continue to be important concerns among health professionals [19,34,48]. With respect
to digital information, GPs also expressed their concern regarding product reliability and patients’ data protection. As recent studies have argued, developers need to consider these key issues when designing health-monitoring technologies [19,48,49].

Participants perceived wearables as user-friendly devices that could foster patients’ empowerment and support them throughout behavior change processes [33,34]. However, the use of wearables for patient education and empowerment has also been associated with a patronizing view of the doctor–patient relationship [6]. Preventing such repercussions represents a concrete challenge faced by research in the health sector. In this context, wearable use can be envisaged in relation to the concept of continuity in health care, defined as informational, relational, or management related [50]. In this sense, wearables constitute tools that can positively contribute to ensuring informational and health management continuity. Nevertheless, these tools alone may not be able to support the multifaceted relationship continuity between the doctor and the patient and would hence need to be adapted.

Self-tracking: A Catalyst for Healthism

GPs were also concerned about the role that wearables could play in patients’ everyday lives outside medical consultations. For instance, GPs highlighted the potential risk of promoting a dominant social discourse or life-philosophy, where self-tracking and self-monitoring become practices that are encouraged, even among individuals who are healthy or who do not suffer from specific health conditions. This overmedicalization of life can be compared to what Gabriels and colleagues [33] have coined as entertainment medicine, where self-tracking devices become responsible for producing “medically unnecessary data that belong more to the fitness or wellness than to the medical realm.” Echoing past literature [32], GPs stressed the importance of understanding patients’ needs in order to address their concerns more effectively.

More generally, self-tracking in the medical field has been previously argued to have culturally and structurally transformed the ways in which health-related practices are being defined [15]. In this sense, an important contribution of our findings to the debate is GPs’ strong resistance to incentive-based medicine, in which healthy behaviors are implemented within a reward versus punishment mechanism. This posture contests 2 aspects. The first is with respect to the global trend across stakeholders to collect information produced by wearable devices for financial purposes [51], which causes ethical concerns to be raised by GPs. The second refers to the philosophical and pedagogical premises underpinning incentive-based medicine. In GPs’ views, this type of medicine contrasts with the value of patient autonomy and risks to promoting an undesirable obsessive compliance with health standards set by wearables. In the contemporary dominant culture of healthism that values self-management [22], this risk becomes increasingly important. Through subtle imperatives, wearables may indeed respond to patients’ intention to take control over their own health [21], while simultaneously triggering feelings of apprehension and self-inadequacy. GPs’ intentions of promoting patient autonomy emphasizes the urgent need to develop alternative approaches in health care that can facilitate behavior change. Indeed, as in the case of other social practices, health practices are subject to ambiguity, contradictions, and ultimately, continuous change [18]. We argue that these premises should be considered in the design of wearable technologies.

Future Perspectives

The rapid expansion of wearables has entailed changes that remain unchallenged regarding their social, psychological and cultural implications for individual and public perceptions of health within our Western societies dominated by healthism [22]. In this sense, it is essential to clarify the rationale underpinning the development and marketing of such devices, whose extensive use may not necessarily be desirable from a GP’s perspective. A clear legal frame guiding the production and distribution of wearables for medical usage might help guide the effectiveness and clinical safety for users and health professionals. For instance, the concept of Health Technology Assessment [52] offers a useful illustration of how this frame could be conceptualized. This study calls for future research to deconstruct and analyze the logic behind the conceptualization, development, and use of health wearables, from the perspective of health professionals, users, and technology developers. In this context, it would be interesting to compare these results with patients’ views, in order to identify possible differences, with an aim toward better integration of wearables in general medical practice. Indeed, our study confirms the necessity for researchers and developers to question the values and logic guiding wearable design.

Limitations

This study is not exempt from limitations. Given its exploratory nature, our qualitative results require further analysis regarding other contexts and methodologies. Moreover, while appropriate to our research setting, mind maps allow limited in-depth data analysis compared to other qualitative methods [35]. In addition, the visually synthetic characteristic of mind maps does not allow for data saturation claims and does not allow the integration of specific details. Rather, mind maps constitute an exploratory step in research that can be complemented by other techniques [47]. Nonetheless, this method is useful to develop hypotheses that can be tested in future research.

Conclusion

This study found that GPs are willing to be more actively engaged as collaborators in the design, development, and promotion of wearables, alongside producers and end-users. Our research contributes to broadening current understanding of wearables and self-tracking technologies in the field of family medicine, by emphasizing the role of wearables as key information providers. Indeed, GPs are neither passive spectators of—nor opponents to—digital health developments, which are perceived to be increasingly more important and inevitable. In spite of the important role of wearables, this study underlined the irreplaceable character of the doctor–patient relationship, which remains a central dimension in family medicine. GPs manifested their opposition to the logic of self-monitoring that GPs considered to have a negative impact on patients’ global well-being and autonomy. Regarding research perspectives, it
seems crucial to reflect upon the definition of health that is being shaped by wearables and similar self-tracking technologies. These perspectives would enable an informed comparison across main actors in health care and contribute to collective coordinated efforts to improve individual and public health while reducing health-related costs.

Acknowledgments
We would like to thank the general practitioners who participated to the workshops. We would also like to thank the moderators: Vincent Gesbert, PhD; Roane Keller, MD; Thierry Mathieu; Daniel Widmer, MD; Baptiste Pedrazzini, MD; and Vanlisa Bourqui. We would like to express our gratitude toward Les Editions Médecine et Hygiène for their kind support in the organization of the symposium.

Conflicts of Interest
None declared.

References
5. Altenhoff B, Vaigneur H, Caine K. One step forward, two steps back: the key to wearables in the field is the app. 2015 Presented at: 9th International Conference on Pervasive Computing Technologies for Health care (PervasiveHealth); May 20-23; Istanbul, Turkey p. 241-244. [doi: 10.4108/icst.pervasivehealth.2015.259049]


27. Damoiseaux RAMJ. [E-health and the GP as gatekeeper; are the two compatible?]. Ned Tijdschr Geneeskd 2014 Dec;4:2010-2013 [FREE Full text] [doi: 10.1016/j.pmedr.2013.05.004] [Medline: 23780804]


42. Pedrazzini B, Bourqui V. Place de l'objet de santé connecté en médecine de famille. Revue Médicale Suisse 2018 Dec 05;14(S21-S29):13-14 [FREE Full text]
45. Biktimirov E, Nilson L. Show them the money: using mind mapping in the introductory finance course. J Fin Educ 2006;32:72-86 [FREE Full text]

Abbreviations

GP: general practitioner
mHealth: mobile health

©Lucia Volpato, María del Río Carral, Nicolas Senn, Marie Santiago Delefosse. Originally published in JMIR mHealth and uHealth (https://mhealth.jmir.org), 09.08.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR mHealth and uHealth, is properly cited. The complete bibliographic information, a link to the original publication on https://mhealth.jmir.org/, as well as this copyright and license information must be included.
Assessing Electrocardiogram and Respiratory Signal Quality of a Wearable Device (SensEcho): Semisupervised Machine Learning-Based Validation Study

Haoran Xu¹, MSc; Wei Yan², MD; Ke Lan³, MSc; Chenbin Ma⁴, BSc; Di Wu¹, MSc; Anshuo Wu⁵, BSc; Zhicheng Yang⁶, PhD; Jiachen Wang¹, BSc; Yaning Zang⁷, BM; Muyang Yan², PhD; Zhengbo Zhang⁸, PhD

¹Medical School of Chinese PLA, Beijing, China
²Department of Hyperbaric Oxygen, The First Medical Center, Chinese PLA General Hospital, Beijing, China
³Beijing SensEcho Science & Technology Co., Ltd., Beijing, China
⁴School of Biological Science and Medical Engineering, Beihang University, Beijing, China
⁵University of Washington, Seattle, WA, United States
⁶PAII Inc., Palo Alto, CA, United States
⁷Department of Kinesiology, Shanghai University of Sport, Shanghai, China
⁸Centre for Artificial Intelligence in Medicine, Medical Innovation Research Department, Chinese PLA General Hospital, Beijing, China

Corresponding Author:
Zhengbo Zhang, PhD
Centre for Artificial Intelligence in Medicine
Medical Innovation Research Department
Chinese PLA General Hospital
28 Fuxing Road, Haidian District, Beijing
Beijing, 100853
China
Phone: 86 13693321644
Email: zhengbozhang301@gmail.com

Abstract

Background: With the development and promotion of wearable devices and their mobile health (mHealth) apps, physiological signals have become a research hotspot. However, noise is complex in signals obtained from daily lives, making it difficult to analyze the signals automatically and resulting in a high false alarm rate. At present, screening out the high-quality segments of the signals from huge-volume data with few labels remains a problem. Signal quality assessment (SQA) is essential and is able to advance the valuable information mining of signals.

Objective: The aims of this study were to design an SQA algorithm based on the unsupervised isolation forest model to classify the signal quality into 3 grades: good, acceptable, and unacceptable; validate the algorithm on labeled data sets; and apply the algorithm on real-world data to evaluate its efficacy.

Methods: Data used in this study were collected by a wearable device (SensEcho) from healthy individuals and patients. The observation windows for electrocardiogram (ECG) and respiratory signals were 10 and 30 seconds, respectively. In the experimental procedure, the unlabeled training set was used to train the models. The validation and test sets were labeled according to preset criteria and used to evaluate the classification performance quantitatively. The validation set consisted of 3460 and 2086 windows of ECG and respiratory signals, respectively, whereas the test set was made up of 4686 and 3341 windows of signals, respectively. The algorithm was also compared with self-organizing maps (SOMs) and 4 classic supervised models (logistic regression, random forest, support vector machine, and extreme gradient boosting). One case validation was illustrated to show the application effect. The algorithm was then applied to 1144 cases of ECG signals collected from patients and the detected arrhythmia false alarms were calculated.

Results: The quantitative results showed that the ECG SQA model achieved 94.97% and 95.58% accuracy on the validation and test sets, respectively, whereas the respiratory SQA model achieved 81.06% and 86.20% accuracy on the validation and test sets, respectively. The algorithm was superior to SOM and achieved moderate performance when compared with the supervised models. The example case showed that the algorithm was able to correctly classify the signal quality even when there were...
complex pathological changes in the signals. The algorithm application results indicated that some specific types of arrhythmia false alarms such as tachycardia, atrial premature beat, and ventricular premature beat could be significantly reduced with the help of the algorithm.

Conclusions: This study verified the feasibility of applying the anomaly detection unsupervised model to SQA. The application scenarios include reducing the false alarm rate of the device and selecting signal segments that can be used for further research.

(JMIR Mhealth Uhealth 2021;9(8):e25415) doi:10.2196/25415

KEYWORDS
signal quality; electrocardiogram; respiratory signal; isolation forest; machine learning; mobile health

Introduction

Background

Wearable devices have been widely adopted for daily health care monitoring during the past decades. Many researchers utilize wearable sensors to continuously monitor physiological signals for mobile health (mHealth) and ubiquitous health (uHealth) app studies [1-3]. Recently, wearable devices have shown their potential in providing early warning of disease deterioration, chronic disease self-management, rehabilitation assessment, among others [4-7]. For example, some clinical deterioration changes in physiological signals could be often present 8-24 hours before a severe life-threatening event such as an unplanned intensive care unit admission or sudden cardiac death [8,9]. In these scenarios, signal quality is essential to acquire the valuable information from the time-series physiological signals which are very sensitive to noise. Signal quality assessment (SQA) facilitates reducing the high false alarm rate caused by signal quality [10] and can be applied to automatically screen the “real-world” data for further research. However, SQA of wearable physiological signals has not been well investigated. Such inadequate studies on signal quality reliability limit the further clinical deployment of these devices in the medical sector [11]. Therefore, it is important to develop a feasible method to evaluate the signal quality from wearable physiological monitoring systems and SQA is one of the basics of mHealth research and apps.

Related Work

It is widely recognized that the electrocardiogram (ECG) and respiratory signals are crucial for both patient monitoring and health status identification, and thus are being extensively investigated. Various solutions have been proposed to accomplish ECG SQA [12,13]. Some early studies, such as those by Langley et al [14] and Johannesen [15], considered the poor quality of ECG signals when their waveform features exceed the preset thresholds [16]. Several signal quality indices (SQIs) such as kSQI (the kurtosis of the distribution), sSQI (the skewness of the distribution), and pSQI (the relative power in the QRS complex) were introduced [17-19], which use the features from the time domain and the frequency domain of the ECG signals to assess the quality [20]. Another approach to ECG SQA is based on template matching. Researchers usually compare the similarity between the signals and a template that is fixed or derived from historical data [21]. In recent years, leveraging the machine learning technology in the medical domain, many researchers used the time–frequency domain features and SQIs to build machine learning models to achieve ECG SQA [16,21-23]. For example, Zhao et al [23] provided an algorithm based on convolutional neural networks, which aimed at identifying noisy segments from wearable ECG recordings. Zhang et al [16] compared the performance of random forest (RF), support vector machine (SVM), and their variants for ECG SQA with nonlinear features. For respiratory signals, Charlton et al [24] developed an SQI for the impedance pneumography respiratory signal by using the breath duration variations and by examining whether the peaks and troughs are clear and similarity of breath morphologies. However, research on respiratory SQA remains in its infancy. Few studies have investigated this topic so far to our knowledge.

Challenges

Owing to the rapid development of wearable devices, there is an explosion of the volume of data being acquired and available for research studies. However, the importance of the SQA process has been underestimated. The limitations of previous studies and the challenges we are currently facing are summarized as follows: For ECG SQA, first, signal quality is often judged subjectively, which lacks objective quantitative criteria, and the standard of signal quality was relatively fuzzy in previous studies [25,26]. Second, most of the SQAs were conducted under well-designed laboratory conditions by using simulated signals [27], or assessed the signals from bedside monitors. Thus, signals are highly different from those measured by wearable devices in daily lives because the noise in the laboratory was relatively single and controllable, or the signal quality was good for most of the time. Third, although most of the methods have good performances on ECG SQA, the dominant methods are still supervised machine learning models [16]. There is a concern that these models are at a high risk of overfitting, leading to unsatisfying model generalization. Moreover, when using supervised models, it is quite challenging to prepare tons of labeled data and even impossible for each research group to use the fixed open-source data sets, such as the MIT-BIH Arrhythmia Database (MITDB), to build models, which were not built for SQAs. In addition, hardware designs of wearable devices are diverse, resulting in aggravating incomplete generalization of data and poor migration performance of models. One possible solution to this problem is to build dedicated models using specific wearable devices and the data they collected. For respiratory SQAs, the challenge lies in the various respiratory patterns. Compared with ECG signals, respiratory signals have more diverse forms, broader spectral distribution, and different noise sources.

https://mhealth.jmir.org/2021/8/e25415

JMIR Mhealth Uhealth 2021 | vol. 9 | iss. 8 | e25415 | p.77
(page number not for citation purposes)
**Study Objectives**

To address the above problems, we pioneered the idea that the SQA process can be seen as an anomaly detection. The basic hypothesis of our study was that the decline of the signal quality can be quantified with the increase of the anomaly and can be detected by the machine learning model. The application scenarios we expected of the algorithm include reducing the false alarms caused by poor signal quality and selecting the high-quality signal segments for further research. The objectives and main components of this paper are to:

- design an algorithm based on the unsupervised machine learning model, isolation forest (IF), to classify the ECG and respiration signal quality into 3 different grades: good, acceptable, and unacceptable.
- quantitatively evaluate the performance of the algorithm on a small amount of labeled data. Further validation of the algorithm was implemented on several cases of data to prove its feasibility.
- apply the SQA algorithm to real-world data to demonstrate that the algorithm has the potential to reduce the false alarms caused by poor signal quality.

**Methods**

**The Wearable Device and Data Sources**

The medical-grade wearable device we used was a self-developed physiological signal monitoring system, SensEcho (Figure 1) [28], which has received clearance from the China Food and Drug Administration (CFDA) and has been deployed in the general wards of the Hyperbaric Oxygen (HBO) Department in Chinese PLA General Hospital (PLAGH) since 2018. The core wearable device of SensEcho is a vest, which provides a single-lead ECG signal, chest and abdominal respiratory signals via the respiratory inductive plethysmography (RIP) technology, and triaxial acceleration signals. It also allows for communication with other third-party wearable devices such as oximeters and blood pressure monitors. Its battery supports continuous monitoring for a minimum of 24 hours. For detailed information about SensEcho and the monitoring system, please refer to [29]. At the time of writing, SensEcho has collected more than 1000 records from patients and healthy individuals. Each record contains nearly 24-hour physiological signal monitoring results; thus, a large pool of data is available for research purposes. Data collection was carried out in a clinical environment for patients and from daily lives for healthy individuals without restriction of movement and activity. In this study, we used the single-lead ECG signal and chest respiration signal from the data pool to establish and evaluate the algorithm. This study was approved by the ethics committee of PLAGH (No. S2018-095-01).

*Figure 1.* Picture of SensEcho, including third-party oximeter and cuff blood pressure monitor.
Signal Quality Classification

Overview

The definition of signal quality was indistinct in previous studies, but some of the studies have proposed a few quantitative criteria. Inspired by [26] and the results of our pre-experiment, 10- and 30-second segments of ECG and respiratory signals were considered sufficient for our study. In early SQA studies, 5 quality groups (excellent, good, adequate, poor, and unacceptable) [15], 3 quality groups (acceptable, indeterminate, and unacceptable) [18,30,31], and 2 quality groups (acceptable and unacceptable) [32-35] were investigated. Based on previous studies, we defined 3 grades of signal quality for different requirements: (1) good signal quality refers to that in which the signal waves are clear, and signal of this grade can be analyzed automatically in follow-up studies and have confidence high enough for waveform feature analysis; (2) acceptable signal quality refers to that in which the R peak in ECG signal and peaks and troughs of respiratory signal can be accurately located by the algorithm, and the signal of this grade can be used for relative accurate heart rate and respiratory rate analysis. In addition, this grade is often the most difficult to distinguish and the signal availability depends on the specific apps where further manual determination might be needed; (3) unacceptable signal quality refers to that in which the waveform in the window is chaotic, and this grade of signal should be dropped because of the unreliable results obtained in signal analysis.

A brief description of characteristics of signal noise sources and their patterns is summarized in the following subsections [12,22,36,37].

Baseline Wander

ECG signals are affected by respiratory motion, body movement, and poor electrode contact. Respiratory signals are more sensitive to movement and breath pattern than ECG signals. One final major expression in signals is different levels of baseline wander.

High-Frequency Noise

For ECG signals, high-frequency noise usually includes power line interference, myoelectricity interference, and movement artifact. For respiratory signals measured by the RIP, the noise often is from vibrations caused by movement, such as moving or speaking.

Signal Loss

This is also a pervasive pattern in daily signal acquisition, which usually appears as a straight line. Based on the noise source and expression analysis, the quantitative evaluation criteria defined by clinical and engineering experts in our study are listed in Table 1.

<table>
<thead>
<tr>
<th>Quality grade</th>
<th>Electrocardiogram</th>
<th>Respiratory signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>• ECG rhythm is clear; each QRS waveform can be distinguished with naked eyes.</td>
<td>• Regular waveform lasts for more than three-fourth of the observation window.</td>
</tr>
<tr>
<td></td>
<td>• No signal loss in the observation window.</td>
<td>• Maximal baseline wander amplitude is less than the signal amplitude in the observation window.</td>
</tr>
<tr>
<td></td>
<td>• Maximal baseline wander amplitude is less than one-third of signal amplitude in the observation window.</td>
<td>• High-frequency noise can be easily filtered and does not affect the judgment of the respiratory signal waveform.</td>
</tr>
<tr>
<td></td>
<td>• Pathological changes do not influence the signal quality assessment; the recognized obvious pathological patterns can be classified as good quality, such as ventricular premature beats.</td>
<td></td>
</tr>
<tr>
<td>Acceptable</td>
<td>• Low-intensity high-frequency noise; the R waves in signal can be recognized accurately.</td>
<td>• One-half to one-fourth of the signal is clear; respiratory rhythm can be identified.</td>
</tr>
<tr>
<td></td>
<td>• No more than 2 high-frequency impulse noises occur in the observation window.</td>
<td>• Time for signal loss or hold breath lasts less than one-half of the observation window.</td>
</tr>
<tr>
<td></td>
<td>• Less than 2-second signal loss in the observation window.</td>
<td>• High-frequency noise has only a little impact on the judgment of the overall waveform trend.</td>
</tr>
<tr>
<td></td>
<td>• The maximal baseline wander amplitude is below the signal amplitude.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Fewer than 2 cardiac cycles in which the QRS waves cannot be recognized are allowed.</td>
<td></td>
</tr>
<tr>
<td>Unacceptable</td>
<td>• Full of noise.</td>
<td>• The pattern of respiratory waveform is difficult to recognize.</td>
</tr>
<tr>
<td></td>
<td>• More than 2 R peaks in the observation window cannot be distinguished.</td>
<td>• Severe baseline wander.</td>
</tr>
<tr>
<td></td>
<td>• Excessive baseline wander.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Signal loss lasts more than 2 seconds.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Suspected pathological changes, but the cause is not clear.</td>
<td></td>
</tr>
</tbody>
</table>

Isolation Forest

IF is an unsupervised anomaly detection model that has been applied to many fields such as streaming data processing and mineral mapping [38,39]. IF grows an ensemble of binary trees to estimate the degree of being an anomaly of an instance. As anomalies are more susceptible to isolation, they have a short path length [38,40]. Furthermore, an anomaly score can be
obtained by measuring/estimating the average path height of the ensemble of binary trees (in [40], the authors named them iTree). The IF model is based on 2 fundamental assumptions and premises. The first one is that the anomalies should be “few and different.” If a pattern occurs frequently in the training set, it will be more likely to be perceived as normality, although it is indeed an anomaly manually determined. The second one is that the training set should conclude as many normal patterns of the signals as possible. It is necessary to guarantee that the training set has a large enough variety, especially for normal signals; otherwise the model will be more likely to classify a brand-new pattern as an anomaly.

Based on the above theory, the general framework of the SQA algorithm is shown in Figure 2. We built models for ECG and respiratory SQA, respectively, and both models were trained and evaluated independently. The preprocess included filtering, removing the outliers, removing the baseline, and normalization. We then selected 8 and 18 features from the time and frequency domains of the ECG and respiratory signals, respectively. Skewness, kurtosis, and distances of adjacent waveforms calculated using the dynamic time warping method [41] were the key features we used, which also have been widely adopted as the key variables to construct the SQIs [17,18,42,43]. The skewness and kurtosis are defined as Equations (1) and (2). Other features we used in this study were the features from amplitude of the signal in the time–frequency domain, power spectrum distribution, and power spectral density.

\[
(1) \quad \text{where} \quad N \quad \text{is the sample points of the signal, is the mean value, and } \sigma \text{ is the SD.}
\]

\[
(2)
\]

**Experiment Design**

**Overview**

The experimental process involved 4 key steps. The model training and validation were conducted on 4 nonoverlapping data sets extracted from the sizable volume data pool and possessed different functions: (1) training set, which was used to train the IF model; (2) validation set, which was used to find the thresholds that map the anomaly scores obtained by the model to the triclassification SQA results; (3) test set, which was used to quantitatively measure the generalization ability of the model; and (4) case set, which was used to qualitatively evaluate the model’s performance by feeding a whole case of data to it. Some details of these 4 data sets are specified in the following sections.

**Training Set**

We selected a set of 24-hour monitoring records which met the following inclusion criteria: (1) signal acquisition was stable by manual determination; (2) no signal loss for extended periods (over 10 minutes) during monitoring; and (3) no persistent atrial fibrillation during monitoring. Based on these, 30 records were included and we selected 3-10 of them randomly to construct the training set with their whole data. We repeated the selection process 20 times for each epoch, that is, we randomly selected 3 records to construct the training set 20 times to find the best performance of the model.

**Validation Set**

We used the data from 16 patients and 8 healthy individuals to construct this data set, expecting that the pathological changes were more complex and the proportion of anomaly was relatively high. We selected 10,000 windows of signals from the records and then removed half of them that were obviously of high quality. The data set was labeled independently by 3 pretrained graduate students of biomedical engineering according to the criteria in the above section. To guarantee label accuracy, we used the agreed result to define the final label, and dropped the windows of signals that had conflicting label results. Moreover, we asked clinical specialists to mark whether the ECG signals in the data set were pathological. If pathological manifestations of the signal, such as arrhythmia or ST-segment elevation, were confirmed, the number of this signal segment was recorded additionally. After the manual annotation of the data set is completed, the anomaly scores of the labeled data can be obtained by feeding the signals to the trained SQA model. Then, thresholds T1 and T2 were set to map the anomaly scores to the signal quality grades. We adjusted the values of T1 and T2, respectively, to find the best performance thresholds, which were fixed and used in the next step.

**Test Set**

Test set data came from 8 patients and 9 healthy individuals, because we expected the test set to be somewhat different from the validation set and to be closer to practical use. We extracted 1 window of signals every 6 minutes and this data set initially comprised 5500 windows of signals, which were labeled in the same way as the validation set. We used the T1 and T2 values
determined by the validation set to obtain the classification results of the model, and then quantitatively evaluated the generalization ability of the model. The basic information about the individuals involved in the validation and test sets is summarized in Table 2.

Table 2. Basic information about the individuals utilized in the validation and test sets.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Validation set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patients (n=16)</td>
<td>Healthy individuals (n=8)</td>
</tr>
<tr>
<td><strong>Demography</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male, n (%)</td>
<td>9 (56)</td>
<td>8 (100)</td>
</tr>
<tr>
<td>Age (year), mean (Q1-Q3)</td>
<td>56 (52-60)</td>
<td>27 (25-33)</td>
</tr>
<tr>
<td>Height (cm), mean (Q1-Q3)</td>
<td>168 (160-170)</td>
<td>174 (171-176)</td>
</tr>
<tr>
<td>Weight (kg), mean (Q1-Q3)</td>
<td>68 (55-76)</td>
<td>68 (59-74)</td>
</tr>
<tr>
<td><strong>Comorbidity, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coronary heart disease</td>
<td>12 (75)</td>
<td>—</td>
</tr>
<tr>
<td>Hyperlipemia</td>
<td>9 (56)</td>
<td>—</td>
</tr>
<tr>
<td>Hypertension</td>
<td>9 (56)</td>
<td>—</td>
</tr>
<tr>
<td>Diabetes</td>
<td>8 (50)</td>
<td>—</td>
</tr>
<tr>
<td>Pulmonary nodule</td>
<td>4 (25)</td>
<td>—</td>
</tr>
<tr>
<td>Sleep apnea syndrome</td>
<td>2 (13)</td>
<td>—</td>
</tr>
</tbody>
</table>

**Case Set**

We fed several cases of data to the model. Different grades of signal quality segments were marked in different colors. We looked at several observation windows in detail to determine whether the model classification results were correct. Note that we are particularly concerned about the pathological changes in the cases, because we expected pure pathological changes to be not misclassified as poor signal quality.

**Data Set Descriptions**

After data labeling, we obtained the final validation and test sets. The validation set consisted of 3460 and 2086 ECG and respiratory labels (all agreed), respectively. Of the 3460 ECG labels, 3022 (87.34%) were good, 189 (5.46%) were acceptable, and 249 (7.20%) unacceptable. Of the 2086 respiratory labels, 1308 (62.70%) were good, 511 (24.50%) acceptable, and 267 (12.80%) unacceptable. The test set consisted of 4686 and 3341 ECG and respiratory labels, respectively. Of the 4686 ECG labels, 3767 (80.39%) were good, 284 (6.06%) acceptable, and 635 (13.55%) unacceptable, compared with 2255 (67.49%), 587 (17.57%), and 499 (14.94%), respectively, for respiratory labels. Some typical examples of the labeled ECG and respiratory signals are shown in Figures 3 and 4.

Meanwhile, for the pathological ECG labels, a total of 661/3460 (19.10%) windows of ECG signal in the validation set were marked. Of these, 648 (98.0%) were labeled as having good quality and the rest (13/661, 1.9%) as acceptable quality. In the test set, 634/4686 (13.53%) windows of signal were pathological; of these, 618 (97.5%) were of good quality and the rest (16/634, 2.5%) were of acceptable quality.
Figure 3. Typical examples of the labeled electrocardiogram signals. (a) & (b) are the normal, good-quality signals; (c) is suspected of arrhythmia while (d) is an expression of ventricular premature beats (VPBs); (e) – (h) show examples of baseline wander, power line interference and impulse noise; (i) – (k) show examples of severe noise and signal loss; (l) is suspected of VPBs but the signal is unclear.

Figure 4. Typical examples of the labeled respiratory signals. (a) – (d) show clear and regular respiratory waves; signals in (e) – (h) do not have enough regularity, apnea occupies some small segments in the observation windows; (i) – (l) show severe noise and signal loss in the observation windows.
Performance Evaluation

The programming language we used was Python (version 3.6.5) and the major library in this study is scikit-learn (version 0.23.1). The proposed algorithm contained 2000 trees and had 5% anomaly proportion as parameters. We first evaluated the algorithm’s performance according to its accuracy score, which is defined as the number of correctly classified samples divided by the total number of samples. Some additional evaluation indicators included mean precision rate, recall rate, and F1 score (marco-F1). To further evaluate the performance of the algorithm, we compared the algorithm with the self-organizing maps (SOMs) [44] and 4 classical supervised machine learning models, namely, logistic regression (LR), SVM, RF, and extreme gradient boosting (XGB). It should be noted that the SOM is an unsupervised model based on artificial neural network and has been applied in several health care–related signal processing fields such as photoplethysmogram signal classification [45, 46] and health situation monitoring [47, 48]. The SOM library used in this study was MiniSom (version 2.2.7) and the SOM model was trained using 10,000 interactions and a 10 × 10 grid on the training set with the learning rate of 0.05. For RF, we used 1000 trees, whereas for XGB, we chose the following hyperparameters: “binary: softmax” as the logistic function and “approx” as the tree method. The other parameters of the models were default. Features were normalized before being fed to LR, SVM, and SOM.

According to our evaluation strategy, for unsupervised models, we trained the models on the training set and found best thresholds on the validation set. For supervised models, we trained the models on the whole validation set. We then compared the performance of both supervised and unsupervised models on the test set. The accuracy, precision, recall, and F1 scores are calculated.

We also investigated the performance of the proposed model with fewer labels in comparison with that of the reference model. We randomly selected 200, 600, and 1000 labels in the validation set to find the thresholds for the unsupervised models and train the supervised models, and then test these on the whole test set. Each random selection is repeated 30 times, and then the mean and SD of the accuracy of the models are computed.

Algorithm Application

We applied the designed SQA algorithm to 1144 cases of data collected in the HBO Department of PLAGH; each of the cases had a dynamic ECG record of nearly 24 hours. Each record of data was read by a clinical expert to give an overall signal quality evaluation result. According to the results, the data were divided into 3 groups, representing different grades of quality of the whole signals. We also scanned the data with an arrhythmia detection algorithm, which is commonly used in automatic dynamic ECG analysis, and the real-time alarm function of SensEcho. The core technology of the arrhythmia detection algorithm is traditional signal processing methods, including filtering and wavelet decomposition. We learned about the type, onset, and duration of each arrhythmia alarm detected by the arrhythmia detection algorithm. For the purpose of this study, a false alarm was defined as the onset of 1 arrhythmia alarm marked with poor signal quality. The proportion of different quality of signals, the number of various arrhythmia alarms, and the percentage of false alarms in each group were calculated.

Results

Model Performance

For the training set that is important for the IF model, we randomly selected monitoring records as described in the “Experiment Design” section and built the training sets to train the model to guarantee the variety and find the best performance of the model. Quantitative evaluation results of the model performance on the validation and test sets are shown in Figure 5. For ECG signals, the model performed at the same level on both validation and test sets, but for respiratory signals, the model performed slightly better on the test set than on the validation set. This is reasonable because the two data sets were constructed differently; thus, the test set was easier for SQA classification. Models that performed the best on the test set were selected for further study. The scores gained from the best model for ECG SQA and the best classification thresholds are shown in Figure 6, in which the accuracy reached 94.97% and 95.58% on the validation and test sets, respectively. The confusion matrices are shown in Figure 7. Similarly, the scores for respiratory SQA and the thresholds are shown in Figure 8. This model achieved 81.06% and 86.20% accuracy on the validation and test sets, respectively. Figure 9 shows the confusion matrix of the results.

The results regarding the classification efficiency of the pathological ECG signal are summarized as follows: in the validation set, 100% (648/648) of good-grade and 23% (3/13) of acceptable-grade pathological ECG signals were classified correctly; however, 77% (10/13) of acceptable-grade signals were misclassified as good quality. In the test set, 99.8% (617/618) of good-grade and 31% (5/16) of acceptable-grade pathological signals were classified correctly; however, 1 sample of good-quality signal was misclassified as acceptable grade and 69% (11/16) of acceptable-grade signals were misclassified as good quality. The above results showed that the model also had a good classification effect on pathological signals: In this study, the vast majority of pathological signals were correctly classified and the misclassification will not increase false-negative decisions.
Figure 5. Quantitative evaluation of the model performance on the validation set and test set. ECG: electrocardiogram.

Figure 6. Electrocardiogram (ECG) signal anomaly scores on the validation set and test set, and the best performance thresholds.

Figure 7. The electrocardiogram confusion matrixes of the results. 0: Good; 1: Acceptable; 2: Unacceptable.
Performance Evaluation Results

The classification results of the desired algorithm and reference models of the test set are summarized in Tables 3 and 4. From Table 3, it can be found that, for supervised models, the LR model performed the worst for both ECG and respiratory signals. Meanwhile, RF and XGB performed slightly better than the proposed algorithm. Understandably, supervised models generally have better performance than unsupervised models. For unsupervised models, SOM performed worse than the proposed model. For ECG SQA, the SOM achieved 0.91 accuracy and 0.55 F1 score on the validation set, indicating an insufficient generalization ability of the thresholds in this scenario for the model. We speculated that the complex pathological changes and noise in the data set made it difficult for SOM to perform dimensionality reduction and correctly map the model outputs to the SQA results. From Table 4, it can be found that the proposed model had a better performance when the number of labels is small. When the number of labels is greater than 1000, the performance of the supervised models was better than that of the proposed model. In other words, when we do not have enough labeled data, the unsupervised model is superior. However, we still recommend preparing slightly more labels as possible to guarantee the stability and generalization ability of the thresholds.
### Table 3. Model performance on the test set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Electrocardiogram</th>
<th>Respiratory signal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Precision</td>
</tr>
<tr>
<td>Supervised models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic regression</td>
<td>0.59</td>
<td>0.55</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>0.60</td>
<td>0.57</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>Extreme gradient boosting</td>
<td>0.97</td>
<td>0.95</td>
</tr>
<tr>
<td>Unsupervised models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-organizing maps</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>Isolation forest, proposed unsupervised model</td>
<td>0.96</td>
<td>0.90</td>
</tr>
</tbody>
</table>

### Table 4. The accuracy on the test set of models with fewer labeled data.

<table>
<thead>
<tr>
<th>Number of labels</th>
<th>Logistic regression</th>
<th>Support vector machine</th>
<th>Random forest</th>
<th>Extreme gradient boosting</th>
<th>Self-organizing maps</th>
<th>Isolation forest, proposed unsupervised model</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECG*, mean (SD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.80 (0.00)</td>
<td>0.85 (0.05)</td>
<td>0.86 (0.06)</td>
<td>0.84 (0.06)</td>
<td>0.80 (0.01)</td>
<td>0.89 (0.06)</td>
</tr>
<tr>
<td>600</td>
<td>0.80 (0.00)</td>
<td>0.86 (0.05)</td>
<td>0.89 (0.06)</td>
<td>0.88 (0.05)</td>
<td>0.81 (0.01)</td>
<td>0.90 (0.06)</td>
</tr>
<tr>
<td>1000</td>
<td>0.81 (0.00)</td>
<td>0.90 (0.04)</td>
<td>0.93 (0.04)</td>
<td>0.92 (0.04)</td>
<td>0.81 (0.01)</td>
<td>0.93 (0.02)</td>
</tr>
<tr>
<td>Respiratory signal, mean (SD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.71 (0.05)</td>
<td>0.71 (0.06)</td>
<td>0.80 (0.04)</td>
<td>0.79 (0.03)</td>
<td>0.70 (0.02)</td>
<td>0.82 (0.04)</td>
</tr>
<tr>
<td>600</td>
<td>0.75 (0.04)</td>
<td>0.75 (0.06)</td>
<td>0.85 (0.02)</td>
<td>0.84 (0.02)</td>
<td>0.73 (0.03)</td>
<td>0.84 (0.02)</td>
</tr>
<tr>
<td>1000</td>
<td>0.77 (0.04)</td>
<td>0.76 (0.06)</td>
<td>0.87 (0.01)</td>
<td>0.86 (0.01)</td>
<td>0.72 (0.07)</td>
<td>0.85 (0.01)</td>
</tr>
</tbody>
</table>

*ECG: electrocardiogram.

### Case Validation

To further evaluate the performance of the algorithm on SQA, the algorithm was tested on several cases. In this paper, ECG and respiratory signals of a patient are illustrated. The patient is a 65-year-old male, standing 170 cm tall, and weighing 68 kg when admitted, and had been monitored by the SensEcho in the general ward of the HBO Department. He was diagnosed with coronary heart disease, posterior mitral valve prolapse, hypertension risk level 2, hyperuricemia, and fatty liver disease.

As shown in Figures 10 and 11, the different signal quality grades classified by the algorithm were marked in 3 colors: the green segments stand for the good quality, the yellow segments for the acceptable quality, and the red segments for the unacceptable quality. Furthermore, in these figures, 4 windows of the monitoring signals were selected to elaborate and illustrate the detailed signals and the classification results, respectively. It can be seen that the monitoring lasted for up to 24 hours, but there was not much high-quality data available in this case. Signal loss was the most common unacceptable signal quality expression and the segments were all marked in red. ECG and respiratory signals of the last few hours were full of noise, so it was suspected that the patient might have removed the device ahead of time.

We found that the pathological changes in ECG did not influence the SQA process directly (Figure 10). Most of the observation windows with ventricular premature beats (VPBs) were also marked in green and yellow correctly, that is, in this case the pathological changes were not filtered which met our expectations. In Figure 11, acceptable and unacceptable signal quality segments are more numerous and dispersed for respiratory signals compared with ECG signals. The good-quality segments were mainly concentrated during the patient’s bed rest period, as breath was more controllable and vulnerable to noise during the day. In conclusion, the algorithm demonstrated an excellent performance in this case and it can be used to automatically screen out the good-quality segments for further research.
Algorithm Application Results

The algorithm application results are summarized in Table 5. The types of arrhythmia alarm we were concerned about were bradycardia, tachycardia, atrial premature beat (APB), VPB, atrial bigeminy, and atrial trigeminy. The “count” column represents the number of cases with a specific arrhythmia alarm detected; for example, bradycardia was detected in 525 cases out of the total 1144 cases. From Table 5, it can be seen that the age, weight, and height of the 3 groups of patients were basically on the same level, whereas the proportion of females increased in the medium and worst groups, indicating that the quality of ECG signal measured from female users might be poor due to hardware. The proportion of different signal quality grades in these cases means that the best group of patients has the highest percentage of good quality and the lowest percentage of unacceptable quality, whereas the worst group of patients has the lowest percentage of good quality and the highest percentage of unacceptable quality. Among these cases, the median [Q1-Q3] for good, acceptable, and unacceptable quality proportion was 90.0% [81.4%-95.9%], 4.8% [2.1%-8.0%], 4.0% [1.1%-9.3%], respectively. These results have 2 implications: First, the desired SQA algorithm is consistent with the common knowledge of people, which can be used to analyze the quality of signals measured by SensEcho automatically and quantitatively. Second, the vast majority of ECG signals measured by SensEcho are usable, which demonstrates that the
wearable device can effectively monitor patients' ECG signal for most of the time.

For the arrhythmia alarm results, ideally, the number of various arrhythmia alarms within each group should be similar. However, it was observed that the number of APBs and VPBs increased significantly (\(P=.02\) and <.001, respectively), suggesting that the signal quality did affect the accuracy of the arrhythmia detection algorithm and that some of the alarms might have been caused by poor signal quality. For the defined false alarm results, the APBs and VPBs increased significantly (\(P<.001\) for both) in the medium and worst groups, and the false alarm of VPBs even accounted for 60.4\% [23.9\%-87.3\%] in the worst group, compared with 18.2\% [0.0\%-61.5\%] for the VPBs among all cases. In addition, it was found that tachycardia had a very high false alarm proportion, probably due to the movement of patients with poor signal quality. We considered that the aforementioned types of false alarms can be detected and effectively reduced by the desired SQA algorithm. Meanwhile, it was also found that for some types of arrhythmia alarms such as those for atrial bigeminy and atrial trigeminy, the arrhythmia detection algorithm was accurate and rarely affected by the signal quality.

Table 5. Results of the SQA algorithm and the arrhythmia detection algorithm applied to the data collected from the Hyperbaric Oxygen Department.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Best (n=671)</th>
<th>Medium (n=365)</th>
<th>Worst (n=108)</th>
<th>Total (n=1144)</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demography</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>239 (35.6)</td>
<td>145 (39.7)</td>
<td>45 (41.7)</td>
<td>429 (37.5)</td>
<td>1144</td>
</tr>
<tr>
<td>Age (year), median (Q1-Q3)</td>
<td>59.3 (53.0-66.8)</td>
<td>61.5 (54.0-67.9)</td>
<td>60.6 (53.5-66.7)</td>
<td>60.1 (53.5-67.1)</td>
<td>1144</td>
</tr>
<tr>
<td>Weight (kg), median (Q1-Q3)</td>
<td>70.0 (62.0-78.5)</td>
<td>71.0 (65.0-80.0)</td>
<td>70.0 (63.0-81.0)</td>
<td>70.3 (63.0-80.0)</td>
<td>1144</td>
</tr>
<tr>
<td>Height (cm), median (Q1-Q3)</td>
<td>168.0 (160.0-173.0)</td>
<td>168.0 (160.0-173.0)</td>
<td>167.5 (160.0-174.0)</td>
<td>168.0 (160.0-173.0)</td>
<td>1142</td>
</tr>
<tr>
<td>Proportion of different signal quality grades detected by the algorithm (%), median (Q1-Q3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td>93.2 (87.0-97.3)</td>
<td>85.9 (78.6-92.0)</td>
<td>75.5 (61.9-86.2)</td>
<td>90.0 (81.4-95.9)</td>
<td>1141</td>
</tr>
<tr>
<td>Acceptable</td>
<td>3.8 (1.7-7.0)</td>
<td>5.8 (3.2-9.4)</td>
<td>5.8 (4.3-9.0)</td>
<td>4.8 (2.1-8.0)</td>
<td>1141</td>
</tr>
<tr>
<td>Unacceptable</td>
<td>2.1 (0.6-5.4)</td>
<td>7.0 (2.9-12.7)</td>
<td>15.2 (6.6-29.2)</td>
<td>4.0 (1.1-9.3)</td>
<td>1141</td>
</tr>
<tr>
<td>Arrhythmia alarm count, median (Q1-Q3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bradycardia</td>
<td>4.0 (2.0-7.0)</td>
<td>3.0 (2.0-6.0)</td>
<td>2.0 (1.0-4.5)</td>
<td>4.0 (2.0-6.0)</td>
<td>525</td>
</tr>
<tr>
<td>Tachycardia</td>
<td>1.0 (1.0-2.0)</td>
<td>1.0 (1.0-2.0)</td>
<td>1.0 (1.0-2.0)</td>
<td>1.0 (1.0-2.0)</td>
<td>224</td>
</tr>
<tr>
<td>APB&lt;sup&gt;a&lt;/sup&gt;</td>
<td>11.0 (4.0-34.5)</td>
<td>15.0 (6.0-42.0)</td>
<td>17.0 (4.0-46.0)</td>
<td>13.0 (5.0-39.0)</td>
<td>1103</td>
</tr>
<tr>
<td>VPB&lt;sup&gt;b&lt;/sup&gt;</td>
<td>6.0 (2.0-25.2)</td>
<td>14.0 (4.0-54.0)</td>
<td>25.0 (5.0-76.0)</td>
<td>9.0 (3.0-39.0)</td>
<td>987</td>
</tr>
<tr>
<td>Atrial bigeminy</td>
<td>4.0 (1.0-13.0)</td>
<td>5.0 (2.0-9.0)</td>
<td>2.5 (2.0-6.0)</td>
<td>4.0 (2.0-10.0)</td>
<td>79</td>
</tr>
<tr>
<td>Atrial trigeminy</td>
<td>4.0 (2.0-10.5)</td>
<td>5.0 (1.0-10.8)</td>
<td>6.0 (2.2-11.2)</td>
<td>4.5 (1.8-11.0)</td>
<td>88</td>
</tr>
<tr>
<td>Defined false alarm proportion (%), median (Q1-Q3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bradycardia</td>
<td>0.0 (0.0-0.0)</td>
<td>0.0 (0.0-0.0)</td>
<td>0.0 (0.0-0.0)</td>
<td>0.0 (0.0-0.0)</td>
<td>525</td>
</tr>
<tr>
<td>Tachycardia</td>
<td>50.0 (0.0-100.0)</td>
<td>100.0 (0.0-100.0)</td>
<td>100.0 (50.0-100.0)</td>
<td>100.0 (0.0-100.0)</td>
<td>224</td>
</tr>
<tr>
<td>APB</td>
<td>0.0 (0.0-9.9)</td>
<td>5.6 (0.0-28.6)</td>
<td>14.1 (0.0-70.6)</td>
<td>0.4 (0.0-19.1)</td>
<td>1103</td>
</tr>
<tr>
<td>VPB</td>
<td>6.2 (0.0-50.0)</td>
<td>35.6 (1.9-71.7)</td>
<td>60.4 (23.9-87.3)</td>
<td>18.2 (0.0-61.5)</td>
<td>987</td>
</tr>
<tr>
<td>Atrial bigeminy</td>
<td>0.0 (0.0-0.0)</td>
<td>0.0 (0.0-0.0)</td>
<td>0.0 (0.0-0.0)</td>
<td>0.0 (0.0-0.0)</td>
<td>79</td>
</tr>
<tr>
<td>Atrial trigeminy</td>
<td>0.0 (0.0-0.0)</td>
<td>0.0 (0.0-0.0)</td>
<td>0.0 (0.0-0.0)</td>
<td>0.0 (0.0-0.0)</td>
<td>88</td>
</tr>
</tbody>
</table>

<sup>a</sup>APB: atrial premature beat.
<sup>b</sup>VPB: ventricular premature beat.

**Discussion**

**Contributions and Principal Findings**

Our highlights and key contributions are summarized as follows:

- We achieve the ECG and respiratory SQA by using an unsupervised model, IF, which has not been applied in SQA before. Furthermore, we attempted to verify the idea that the SQA process can be viewed as an anomaly detection. In this study, the proposed algorithm was superior than SOM and achieved moderate performance when compared with the supervised models.

- We applied the SQA algorithm to a large data set with 1144 records of ECG signal. The results demonstrate that the arrhythmia alarm accuracy could be influenced by the signal quality, and the SQA algorithm has the potential to reduce
some specific types of arrhythmia false alarms such as tachycardia, APB, and VBP caused by poor signal quality.

- To our knowledge, this is one of the earliest studies that focuses on the quality of respiratory signals measured via the RIP technology. It provides a method to automatically select the high-quality segments of respiratory signal for further studies.

One featured point in our study is that 3 data sets that have different functions were used to construct and quantitatively validate the algorithm. In the workflow of our study, the training set was a large volume data set in which ideally all the patterns of the signal could be enumerated, while the validation set and the test set were unseen by the model when we trained it. We also conducted a very small experiment, where we directly trained the models on the validation set, found the best performance thresholds, and then evaluated the performance of the models on the test set. The results showed that for the ECG signal, the model achieved 0.92 accuracy and 0.72 F1 score, whereas for the respiratory signal, the model achieved 0.72 accuracy and 0.68 F1 score, which are lower than the current performance in the “Results” section. These results demonstrate that the diversity of patterns in the training set ensures the generalization performance of the unsupervised model. In fact, in an era of big data, it is easy to obtain a training set with a large sample size, yet lacking labels. The workflow we proposed in this study provides a feasible way to take advantage of the large sample size that can be applied in follow-up studies.

What should be emphasized is that we included the respiratory signal measured via RIP in this study for 2 reasons. First, the respiratory signal is an important physiological signal, which contains abundant personalized information, indicating the health status and disease deterioration of a person. More importantly, the quality of respiratory signal measured via RIP is not well investigated compared with ECG. In our study, we would like to point out that a signal with relatively little research and no fixed waveform could also be assessed by this method, which has the potential to be extended to other SQA scenarios such as impedance pneumography respiratory signal, dynamic blood pressure, and photoplethysmogram. That is, our study provides a practical workflow for other time-series physiological signal research groups to develop their own applicable SQA algorithms.

**Limitations**

There are also some limitations to our work. First, the model we used was an unsupervised machine learning model, which lacks enough interpretability and the performance is largely determined by the quality of the training set. We attempted several construction methods of the training set, yet it was hard to guarantee that the models achieved the best performance. Second, the classification results of the models for the medium grade of signal quality were not good. The sensitivities of the grade of signal quality were not good. The sensitivities of the algorithm for this grade are only 0.34 for ECG and 0.57 for respiratory signals, respectively, which seriously lower the overall F1 scores of the models. This is because the medium level of signal quality is always the hardest to classify even manually. We tried some approaches such as data augmentation and constructing an artificial training set. However, the results showed no significant improvement. It is worth mentioning that the SOM showed moderate performance in the unsupervised methods, perhaps because, in our study, the framework, especially the training and generalization methods, was not suitable for this model. Further to this point, SOM and the rapidly evolving deep learning methods are worth being investigated after further accumulation of data. Third, as the validation of the algorithm on pathological signals was insufficient, although the results in this study were good, we still consider that the algorithm has the risk of misclassifying pathological changes as abnormal as a result of noise. We thus need to further validate the algorithm, which demands more pathological data accumulation and long-term feedback of actual use from clinicians.

**Future Work**

Our future research includes the following. First, the algorithm calls for more comprehensive experimental validation. Accordingly, we should further verify the performance of the model in the presence of pathological changes and quantify how much the model can reduce the false alarm rate. It requires long-term usage and more data collection, especially from patients with specific diseases such as arrhythmia and chronic obstructive pulmonary disease. Second, we will test the time usage and real-time performance of the algorithm. To our knowledge, the IF model operation does not take too much time when the thresholds are determined, yet the feature extraction process is more time-consuming. As we preliminarily tested, the whole SQA process for ECG signal takes 0.3-0.5 seconds on server for every observation window (10 seconds). For respiratory signal, it takes less than 0.1 seconds for every observation window (30 seconds). We will integrate the algorithm into the server to achieve the real-time SQA. Third, there are many mHealth and uHealth apps nowadays, but there is a lack of assessment of the data measured under nonlaboratory conditions and their usability. Based on the algorithm we developed, we will further evaluate the value of the wearable device, SensEcho, in daily life situations from a signal quality perspective, find the cause of the decrease in signal quality, and improve the hardware and software of the wearable device. We believe that this will further promote the application of mHealth and uHealth.

**Conclusions**

In this study, the results verified our hypothesis that the SQA problem can be seen as an anomaly detection. We built a model based on the unsupervised machine learning model, IF, to avoid heavy data annotation work and to realize ECG and respiratory SQA. What distinguishes us from other studies that used the IF model is that we used a small amount of labeled data to enable the mapping of model scores to human cognitive classification results. Our validation results indicate that the proposed algorithm is superior than SOM and shows a moderate performance compared with supervised models. Meanwhile, the proposed algorithm has the advantages of flexibility, easy adjustment, and better performance with few labeled data. In addition, the pathological changes in our case are correctly classified, demonstrating the model’s good application effect. The algorithm application results on 1144 cases from the clinic...
suggest that the proposed algorithm has the potential to reduce some types of arrhythmia false alarms such as tachycardia, APB, and VBP.

Middle-aged and elderly people, such as patients in the HBO Department in this study, often suffer from complex chronic diseases and are at relatively high risk even in hospitals. Therefore, the adoption of wearable devices in clinics and the advancement of data analysis could provide easily accessible health care that can greatly benefit this population. We consider that the proposed algorithm can advance the clinical apps of wearable devices and facilitate follow-up mHealth and uHealth studies of various time-series physiological signals.

Acknowledgments
This work was supported by NSF of China (61471398), Beijing Municipal Science and Technology (Z181100001918023), Special Grant for Health Care (16BJZ23), CERNET Innovation Project (NGH20160701), Big Data Research & Development Project of Chinese PLA General Hospital (2018MBD-09), and National Key Research and Development Project (2017YFC0114001).

Authors’ Contributions
HX and WY contributed equally to this work. HX designed the algorithm, analyzed the results, and wrote the manuscript; WY guided the data collection and data labeling; ZZ and MY provided the general guidance; KL designed and validated the algorithm; DW designed the ECG SQA algorithm; AW and ZY revised the manuscript; CM, JW, and YZ cleaned and labeled the data.

Conflicts of Interest
This work was done during ZY’s internship at Beijing SensEcho Science & Technology Co, Ltd, Beijing, China, when he was a PhD candidate at University of California, Davis, CA, USA. The other authors have no conflicts to declare.

References


Abbreviations

APB: atrial premature beat
CFDA: China Food and Drug Administration
ECG: electrocardiogram
HBO: hyperbaric oxygen
IF: isolation forest
LR: logistic regression
mHealth: mobile health
MITDB: MIT-BIH Arrhythmia Database
PLAGH: Chinese PLA General Hospital
RF: random forest
RIP: respiratory inductive plethysmography
SOM: self-organizing maps
SQAs: signal quality assessment
SQIs: signal quality indices
SVM: support vector machine
uHealth: ubiquitous health
VPB: ventricular premature beat
XGB: extreme gradient boosting
Review

Evaluating the Validity and Utility of Wearable Technology for Continuously Monitoring Patients in a Hospital Setting: Systematic Review

Vikas Patel¹, BMSc, MD; Ani Orchanian-Cheff², BA, MIS; Robert Wu¹,³, MSc, MD, FRCPC

¹Faculty of Medicine, University of Toronto, Toronto, ON, Canada
²Library and Information Services, University Health Network, Toronto, ON, Canada
³Division of General Internal Medicine, University Health Network, Toronto, ON, Canada

Corresponding Author:
Vikas Patel, BMSc, MD
Faculty of Medicine
University of Toronto
1 King’s College Cir
Toronto, ON, M5S 1A8
Canada
Phone: 1 4169756585
Email: vik.patel@mail.utoronto.ca

Abstract

Background: The term posthospital syndrome has been used to describe the condition in which older patients are transiently frail after hospitalization and have a high chance of readmission. Since low activity and poor sleep during hospital stay may contribute to posthospital syndrome, the continuous monitoring of such parameters by using affordable wearables may help to reduce the prevalence of this syndrome. Although there have been systematic reviews of wearables for physical activity monitoring in hospital settings, there are limited data on the use of wearables for measuring other health variables in hospitalized patients.

Objective: This systematic review aimed to evaluate the validity and utility of wearable devices for monitoring hospitalized patients.

Methods: This review involved a comprehensive search of 7 databases and included articles that met the following criteria: inpatients must be aged >18 years, the wearable devices studied in the articles must be used to continuously monitor patients, and wearables should monitor biomarkers other than solely physical activity (ie, heart rate, respiratory rate, blood pressure, etc). Only English-language studies were included. From each study, we extracted basic demographic information along with the characteristics of the intervention. We assessed the risk of bias for studies that validated their wearable readings by using a modification of the Consensus-Based Standards for the Selection of Health Status Measurement Instruments.

Results: Of the 2012 articles that were screened, 14 studies met the selection criteria. All included articles were observational in design. In total, 9 different commercial wearables for various body locations were examined in this review. The devices collectively measured 7 different health parameters across all studies (heart rate, sleep duration, respiratory rate, oxygen saturation, skin temperature, blood pressure, and fall risk). Only 6 studies validated their results against a reference device or standard. There was a considerable risk of bias in these studies due to the low number of patients in most of the studies (4/6, 67%). Many studies that validated their results found that certain variables were inaccurate and had wide limits of agreement. Heart rate and sleep were the parameters with the most evidence for being valid for in-hospital monitoring. Overall, the mean patient completion rate across all 14 studies was >90%.

Conclusions: The included studies suggested that wearable devices show promise for monitoring the heart rate and sleep of patients in hospitals. Many devices were not validated in inpatient settings, and the readings from most of the devices that were validated in such settings had wide limits of agreement when compared to gold standards. Even some medical-grade devices were found to perform poorly in inpatient settings. Further research is needed to determine the accuracy of hospitalized patients’ digital biomarker readings and eventually determine whether these wearable devices improve health outcomes.

(JMIR Mhealth Uhealth 2021;9(8):e17411) doi:10.2196/17411
KEYWORDS
wearable; inpatient; continuous monitoring

Introduction

Background
Most physiologic parameters, such as vital signs or activity, are routinely monitored a few times each day in hospital ward settings [1]. Some parameters, such as sleep, are not routinely monitored at all [2,3]. More frequent monitoring could allow for the timely identification of the deteriorating health of patients and spur efforts for improving patients’ overall health through increased sleep and activity. Since subtle changes in vital signs are often present 8 to 24 hours before a life-threatening event, such as intensive care unit admission or cardiac arrest, vital sign surveillance has the potential to detect clinical deterioration at an earlier phase, thereby permitting clinicians to make corrective interventions [4-7]. This includes identifying patients with poorly controlled pain and recognizing arrhythmias. The term posthospital syndrome has been used to denote the deleterious effects of acute illnesses that are compounded with poor sleep and low activity and occur during hospital stay [8]. Measuring sleep and activity could improve the recognition of such issues and encourage health providers to introduce interventions that improve patients’ experiences in hospitals by encouraging mobilization and to identify targets for sleep-promoting interventions [9-11]. In addition, access to other digital biomarkers (eg, heart rate, blood pressure, oxygen saturation, etc) would allow clinicians to determine underlying etiologies and make tailored interventions.

The rapid uptake of affordable wearables, such as fitness bands, may provide a method for continuously measuring sleep; activity; and vital signs, such as heart rate [12-15]. However, existing literature that describes wearable devices is mostly limited to ambulatory settings and focuses on the management of chronic diseases [16,17]. More inpatient data are needed on both the validity of wearables and patient adherence. Although wearable testing has been conducted with healthy volunteers, it will be important to validate these signals in inpatient settings, where algorithms for processing sensor data into digital signals, such as those for sleep, heart rate, and activity, may be less accurate [18]. Despite the proposed benefit of intensive monitoring, many wearable studies have found issues with patient adherence [18-20]. Adherence is a crucial barrier to acquiring data and can be influenced by device convenience, the comfort of use, and interaction requirements [19]. Studies of wearable devices worn by hospitalized inpatients have been limited by large dropout rates [20].

Although there have been systematic reviews of the monitoring of patients’ physical activity in hospitals [21-23], there are no reviews of the use of wearables that can reliably measure other health parameters. Therefore, in this review, we aimed to expand our search by including articles that used wearables to assess parameters other than physical activity and to assess the adherence of patients in inpatient settings.

Objective
For the purposes of this review, a wearable was considered to be any electronic device that has at least 1 sensor and can be worn on the body [24]. Wearables were examined for their ability to measure digital biomarkers, which are defined as digitally collected physiological and behavioral measures (eg, heart rate, average sleep duration, and daily step count) that explain, influence, or predict health-related outcomes [18]. Consistent with previous research, patient adherence was objectively assessed by reporting the mean proportion of patients who completed a given study [25]. The primary objectives of this review were to determine patients’ adherence to using wearable devices in hospitals and to examine the validity of wearable-derived biomarker readings.

Methods

Identification and Selection of Studies
A comprehensive search strategy was developed to identify articles on the three main concepts of our question—wearables, monitoring, and inpatients. The initial search strategy was developed for Ovid MEDLINE by using a combination of database-specific subject headings and text words (Multimedia Appendix 1). Additional key words were generated based on input from the subject specialists on the team, and the revised search strategy was customized for each database.

Searches of the following databases were executed on August 16, 2018: Ovid MEDLINE, Ovid MEDLINE Epub Ahead of Print and In-Process & Other Non-Indexed Citations, Cochrane Database of Systematic Reviews, Cochrane Central Register of Controlled Trials, Health Technology Assessment database (Ovid), and CINAHL with Full Text. The search in Ovid Embase was not executed until September 5, 2018, due to issues with the vendor’s August database reload. Additional search methods included reviewing the cited references of eligible studies via Web of Science (May 6, 2019) and the reference lists of eligible studies. There were no restrictions on publication period. Limits were imposed to ensure that only English-language studies and those with adult populations were included in this review. No other limits were applied to the literature search.

Article Selection and Exclusion Criteria
Records were screened by two reviewers (VP and RW) independently. For selected studies, full-text articles were obtained and evaluated for eligibility [26]. The eligibility criteria for inclusion in this review were as follows:

- Medical or surgical inpatients aged >18 years
- Device studied in the article must be a wearable (such as a watch, vest, pendant, jewelry, headset, and wristband)
- Articles must describe an element of continuous monitoring for at least 24 hours or greater
- Articles must describe the measurement of 1 or more digital biomarkers other than just physical activity or standard hospital telemetry for heart rate recording.

References were obtained for full-text articles and then imported into EndNote X7. A second search was conducted in all databases with the updated date of December 13, 2019. Reports with no English title were translated into English using Google Translate. All article titles and abstracts were independently screened for eligibility by two reviewers (VP and RW) using standardized inclusion criteria. The eligibility criteria were designed to include studies that assessed patient adherence in inpatient settings and which utilized wearables as monitors of health parameters other than physical activity and standard hospital telemetry for heart rate recording.

Conclusion
This review of wearable devices used in hospitals and to examine the validity of wearable-derived biomarker readings identified a significant number of articles that met the eligibility criteria. The majority of these studies were conducted in inpatient settings and used wearables to monitor a variety of health parameters, including sleep, heart rate, and physical activity. These findings underscore the potential of wearables to improve patient adherence and the recognition of health-related issues in inpatient settings.

Appendix 1

| Medical or surgical inpatients aged >18 years |
| Device studied in the article must be a wearable (such as a watch, vest, pendant, jewelry, headset, and wristband) |
| Articles must describe an element of continuous monitoring for at least 24 hours or greater |
| Articles must describe the measurement of 1 or more digital biomarkers other than just physical activity or standard hospital telemetry for heart rate recording |

<table>
<thead>
<tr>
<th>Eligibility criteria</th>
<th>Inclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical or surgical inpatients aged &gt;18 years</td>
<td>Yes</td>
</tr>
<tr>
<td>Device studied in the article must be a wearable (such as a watch, vest, pendant, jewelry, headset, and wristband)</td>
<td>Yes</td>
</tr>
<tr>
<td>Articles must describe an element of continuous monitoring for at least 24 hours or greater</td>
<td>Yes</td>
</tr>
<tr>
<td>Articles must describe the measurement of 1 or more digital biomarkers other than just physical activity or standard hospital telemetry for heart rate recording</td>
<td>Yes</td>
</tr>
</tbody>
</table>

https://mhealth.jmir.org/2021/8/e17411
We excluded articles that were not considered original research, such as letters to the editor, comments, and reviews. We also excluded articles that monitored less than 3 patients, described the monitoring of a very specialized system in the body (eg, insole devices, ventricular assistive devices, and cochlear implants), involved the monitoring of patients in rehabilitation hospitals, or used wearables as tools for therapy (eg, insulin delivery).

**Data Extraction**

Two reviewers (RW and VP) independently extracted the data and resolved any disagreements by discussing the findings and making a collective decision. The data extracted for each article included the year of publication, study setting and design, number of participants, gender ratio, mean age of participants, digital biomarkers measured in the study, average and maximum duration that the wearable was worn by participants in each study, and patient completion rate (the proportion of patients that wore the wearable for the minimum monitoring duration that was set by the study authors). For studies that used a reference standard, any participants who were missing data from the wearable or the standard were determined to be incomplete measurement pairs and were omitted from the final count of patients who completed the study. Furthermore, we extracted the types of wearables that were worn by the participants in each study along with the placement sites on the body. Devices were classified as medical grade (approved or cleared by the US Food and Drug Administration), research grade (typically used in research settings only), and consumer grade (used by general consumers).

Validation data were also collected for each article by assessing whether the authors compared the accuracy of their digital readings to a reference standard. To determine the validity of measures that were compared to a reference standard, correlation coefficients, mean differences, and limits of agreement were extracted from each study.

**Risk of Bias Assessment**

All articles that assessed for validated readings were independently assessed for their risk of bias by two independent reviewers (VP and RW) using a modification of the validation subscale from a checklist for assessing the methodological quality of studies on the measurement properties of health status measurement instruments (Consensus-Based Standards for the Selection of Health Status Measurement Instruments [COSMIN]) [27] (Table 1). All discrepancies were resolved by discussion and consensus. The quality evaluation included 5 study design and methodology components (the percentage of missing data, missing data management, adequate sample size, acceptable criterion comparison, and design or methodological flaws) and 1 analysis component (acceptable accuracy analyses). We rated the quality of each dimension as excellent, good, fair, or poor based on a priori modifications to the COSMIN validation subscale for scoring criteria that are appropriate for accuracy studies (Multimedia Appendix 2) [28].

<table>
<thead>
<tr>
<th>Study</th>
<th>Mean or % difference</th>
<th>Correlation</th>
<th>LOA Percent of missing data</th>
<th>Adequate sample size patients</th>
<th>Adequate sample size measurements</th>
<th>Acceptable reference comparison</th>
<th>Acceptable accuracy analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bloch et al [29]</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Excellent</td>
<td>Poor</td>
<td>Poor</td>
<td>Excellent</td>
</tr>
<tr>
<td>Breteler et al [30]</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Excellent</td>
<td>Poor</td>
<td>Excellent</td>
<td>No</td>
</tr>
<tr>
<td>Gallo and Lee [13]</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Excellent</td>
<td>Fair</td>
<td>Fair</td>
<td>No</td>
</tr>
<tr>
<td>Kroll et al [11,31]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Excellent</td>
<td>Good</td>
<td>Excellent</td>
<td>No</td>
</tr>
<tr>
<td>Steinhubl et al [32]</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Excellent</td>
<td>Poor</td>
<td>Excellent</td>
<td>No</td>
</tr>
<tr>
<td>Weenk et al [4]</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Excellent</td>
<td>Poor</td>
<td>Good</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 1. Risk of bias assessment for studies that validated their wearable readings.

*aLOA: limits of agreement.*

**Results**

**Characteristics of Included Studies**

Our literature search identified 2754 article citations. After excluding duplicate records, 2012 records were deemed eligible for screening. A total of 83 studies were selected based on abstracts and underwent full-text review. After applying our inclusion and exclusion criteria, 15 articles that described 14 studies were selected for this review (Figure 1).
All of the articles included were prospective cohort studies (Table 2) [4,11,13,14,20,29-38]. Overall, 9 different types of commercial wearables were described across the 14 studies, and 7 different health variables were assessed collectively by the 14 studies (Table 3). The wearable devices that were described by the studies came in various different forms and were attached to a range of sites on the body (Figure 2). A total of 13 articles included both men and women as the study participants; the other two papers assessed sleep changes in postpartum women [13,34]. Kroll et al [11,31] published two articles from the same study. Both articles analyzed different aspects of the continuous monitoring of inpatients (ie, they used the same cohort of patients) but were included as the same study entry in this review (Table 2).

Collectively, the mean patient completion rate across all 14 studies was over 90%. Of the 8 articles that included a qualitative analysis as a part of their methodology, 7 reported that wearables were well received by either or both patients and clinicians.

Of the 14 studies, 6 validated wearable measurements against another standard device or measure (Table 3). The studies conducted by Bloch et al [29], Gallo and Lee [13], and Steinhubl et al [32] used intermittent measurements (nurse or questionnaires) for their reference standard. Further, Breteler et al [30] used a continuous reference (continuous electrocardiography and impedance pneumography) to compare the wearable readings for heart rates and respiratory rates [30]. Weenk et al [4] and Kroll et al [11,31] validated their wearable readings against both intermittent and continuous reference measurements. Of the 9 wearables included in the studies, 6 were cleared or approved by the US Food and Drug Administration as medical devices (ViSi Mobile [Sotera Wireless], Hidalgo EQ02 [Equivital], wrist actigraphy [Ambulatory Monitoring Inc; Actigraph LLC], LifeTouch [Isansys Lifecare], Zephyr Biopatch [Medtronic], and HealthPatch [VitalConnect]).
Table 2. Summary of included studies.

<table>
<thead>
<tr>
<th>Study</th>
<th>Year published</th>
<th>Setting (ward)</th>
<th>Methodology</th>
<th>Patients, N</th>
<th>Male %: Female % ratio, mean age (years)</th>
<th>Variables measured</th>
<th>Number of days device was worn, average (maximum)</th>
<th>Patient completion rate, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lee and Lee [34]</td>
<td>2007</td>
<td>Obstetric</td>
<td>Prospective cohort</td>
<td>21</td>
<td>Females only, 32</td>
<td>Sleep</td>
<td>2a</td>
<td>100</td>
</tr>
<tr>
<td>Gallo and Lee [13]</td>
<td>2008</td>
<td>Obstetric</td>
<td>Prospective cohort</td>
<td>39</td>
<td>Females only, 29</td>
<td>Sleep</td>
<td>2 (2)</td>
<td>100</td>
</tr>
<tr>
<td>Bloch et al [29]</td>
<td>2011</td>
<td>Geriatric</td>
<td>Prospective cohort</td>
<td>10</td>
<td>Males and females, 83</td>
<td>Falls</td>
<td>21a</td>
<td>90</td>
</tr>
<tr>
<td>Chiu et al [33]</td>
<td>2013</td>
<td>Neurosurgery</td>
<td>Prospective cohort</td>
<td>60</td>
<td>65:35, 35</td>
<td>Sleep</td>
<td>7a</td>
<td>87</td>
</tr>
<tr>
<td>Watkins et al [37]</td>
<td>2015</td>
<td>Medicine and surgical</td>
<td>Prospective cohort</td>
<td>236</td>
<td>Males and females</td>
<td>HR, RR, SpO2, and BP</td>
<td>3 (3)</td>
<td>100</td>
</tr>
<tr>
<td>Jeffs et al [20]</td>
<td>2016</td>
<td>Medicine</td>
<td>Prospective cohort</td>
<td>208</td>
<td>72:28c</td>
<td>HR, RR, SpO2, temperature, and accelerometry</td>
<td>(14)b</td>
<td>32</td>
</tr>
<tr>
<td>Steinhubl et al [32]</td>
<td>2016</td>
<td>Medicine</td>
<td>Prospective cohort</td>
<td>26</td>
<td>65:35, 33</td>
<td>HR, RR, and temperature</td>
<td>3 (3)</td>
<td>100</td>
</tr>
<tr>
<td>Razjouyan et al [35]</td>
<td>2017</td>
<td>Hematology and oncology</td>
<td>Prospective cohort</td>
<td>35</td>
<td>45:55, 55</td>
<td>HR and fall risk</td>
<td>1a</td>
<td>94</td>
</tr>
<tr>
<td>Weenk et al [4]</td>
<td>2017</td>
<td>General internal medicine and surgical</td>
<td>Prospective cohort</td>
<td>20</td>
<td>65:35, 50</td>
<td>HR, RR, BP, SpO2, and temperature</td>
<td>2.5 (3)</td>
<td>100</td>
</tr>
<tr>
<td>Kroll et al [11,31]</td>
<td>2017</td>
<td>Intensive care unit</td>
<td>Prospective cohort</td>
<td>50</td>
<td>52:48, 64</td>
<td>HR, sleep</td>
<td>1a</td>
<td>96</td>
</tr>
<tr>
<td>Weller et al [36]</td>
<td>2017</td>
<td>Neurology and neurosurgery</td>
<td>Prospective cohort</td>
<td>736</td>
<td>54:46c</td>
<td>HR, RR, SpO2, and BP</td>
<td>1.7 (9)</td>
<td>100</td>
</tr>
<tr>
<td>Breteler et al [30]</td>
<td>2018</td>
<td>Surgical</td>
<td>Prospective cohort</td>
<td>33</td>
<td>72:28, 63</td>
<td>HR and RR</td>
<td>2.6 (3)</td>
<td>76</td>
</tr>
<tr>
<td>Yang et al [14]</td>
<td>2018</td>
<td>Oncology</td>
<td>Prospective cohort</td>
<td>11</td>
<td>64:36c</td>
<td>Sleep</td>
<td>16a</td>
<td>91</td>
</tr>
<tr>
<td>Duus et al [38]</td>
<td>2018</td>
<td>General surgery</td>
<td>Prospective cohort</td>
<td>50</td>
<td>58:42, 71</td>
<td>HR, RR, and SpO2</td>
<td>3.1 (4)</td>
<td>100</td>
</tr>
</tbody>
</table>

*aThe maximum number of days was not reported in the study.

bThe study included both male and female participants but did not report a ratio.

cMean age was not reported in the study.

dHR: heart rate.

eRR: respiratory rate.

fSpO2: oxygen saturation

gBP: blood pressure.

hThe average number of days was not reported in the study.
Table 3. Distribution of the health variables that were assessed for accuracy in each study.

<table>
<thead>
<tr>
<th>Study</th>
<th>Device characteristics</th>
<th>Digital biomarkers</th>
<th>Respiratory rate</th>
<th>SpO₂</th>
<th>Skin temperature</th>
<th>Blood pressure</th>
<th>Fall risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gallo and Lee [13]</td>
<td>Wrist Actigraph, Ambulatory Monitoring Inc</td>
<td>Yes</td>
<td>_</td>
<td>c</td>
<td>R=0.53</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Lee and Lee [34]</td>
<td>Mini-Motionlogger-Actigraphy, Ambulatory Monitoring Inc</td>
<td>Yes</td>
<td>—</td>
<td>Not validated</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Chiu et al [33]</td>
<td>Actigraph GT1M, Actigraph LLC</td>
<td>Yes</td>
<td>—</td>
<td>Not validated</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Yang et al [14]</td>
<td>Actigraph GT3X+ watch, Actigraph LLC</td>
<td>Yes</td>
<td>—</td>
<td>Not validated</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Kroll et al [11,31]</td>
<td>Fitbit Charge HR, Fitbit Inc</td>
<td>—</td>
<td>LoA (sinus): 23.9 to 21.9 beats per minute</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Breteler et al [30]</td>
<td>HealthPatch, VitalConnect</td>
<td>Yes</td>
<td>LoA: −8.8 to 6.5 beats per minute</td>
<td>—</td>
<td>—</td>
<td>Not validated</td>
<td>—</td>
</tr>
<tr>
<td>Jeffs et al [20]</td>
<td>Hidalgo EQ02, Equival</td>
<td>Yes</td>
<td>Not validated</td>
<td>—</td>
<td>Not validated</td>
<td>Not validated</td>
<td>—</td>
</tr>
<tr>
<td>Duus et al [38]</td>
<td>LifeTouch, Ipsys Lifecare</td>
<td>Yes</td>
<td>Not validated</td>
<td>—</td>
<td>Not validated</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Steinhubl et al [32]</td>
<td>MultiSense patch, Rhythm Diagnostic Systems</td>
<td>—</td>
<td>R=0.75</td>
<td>—</td>
<td>R=0.83</td>
<td>—</td>
<td>R=0.99</td>
</tr>
<tr>
<td>Bloch et al [29]</td>
<td>Vigi’Fall, Vigilio Telemedical</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Weenk et al [4]</td>
<td>ViSi Mobile, Sotera Wireless</td>
<td>Yes</td>
<td>LoA: −11.1 to 10.7 beats per minute</td>
<td>—</td>
<td>−5.5 to 7.9 breaths per minute</td>
<td>−3.1% to 3.3%</td>
<td>Not validated</td>
</tr>
<tr>
<td>Weenk et al [4]</td>
<td>HealthPatch, VitalConnect</td>
<td>Yes</td>
<td>—12.6 to 9.5 beats per minute</td>
<td>—</td>
<td>−10.3 to 9.0 breaths per minute</td>
<td>Not validated</td>
<td>Not validated</td>
</tr>
<tr>
<td>Weller et al [36]</td>
<td>ViSi Mobile, Sotera Wireless</td>
<td>Yes</td>
<td>Not validated</td>
<td>—</td>
<td>Not validated</td>
<td>Not validated</td>
<td>—</td>
</tr>
<tr>
<td>Watkins et al [37]</td>
<td>ViSi Mobile, Sotera Wireless</td>
<td>Yes</td>
<td>Not validated</td>
<td>—</td>
<td>Not validated</td>
<td>—</td>
<td>Not validated</td>
</tr>
<tr>
<td>Razjouyan et al [35]</td>
<td>Zephyr BioPatch, Medtronic</td>
<td>Yes</td>
<td>Not validated</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

aFDA: US Food and Drug Administration.
bSpO₂: oxygen saturation.
cNot available.
dLOA: limits of agreement.
eSteinhubl et al [32] did not report limits of agreement.
SBP: systolic blood pressure.
DBP: diastolic blood pressure.

Figure 2. Illustration of the types of and body locations for used wearable devices.

Risk of Bias
Of the 6 studies in the risk of bias assessment, 4 were ranked as poor due to a small sample size (participants: N<30). The study conducted by Gallo and Lee [13] used sleep questionnaires as a reference measure and therefore received a fair rating for the “acceptability of reference” criterion, whereas the other five studies were ranked as excellent (ie, they used intermittent nurse readings or other validated methodologies). Further, in terms of assessing the accuracy analyses, only the study conducted by Bloch et al [29] did not report mean differences, correlations, and limits of agreement.

Validation by Digital Biomarker

Heart Rate
A total of 5 studies assessed heart rate accuracy. Breteler et al [30] found that the bias and 95% limits of agreement for heart rate were −1.1 beats per minute (BPM) and −8.8 to 6.5 BPM, respectively, for 55,565 heart rate pairs [30]. Specifically, the wearable sensor accurately detected tachycardia with a sensitivity of 90% and a specificity of 97% [30]. In a cohort of intensive care unit patients, Kroll et al [11,31] found that the Fitbit (Fitbit Inc)-derived heart rate values were slightly lower than those derived from continuous electrocardiography monitoring but that 73% of the readings were within 5 BPM of the electrocardiogram value (average bias: −1.14 BPM; R=0.74; P<.001; heart rate pairs: n=12,358) [11]. Overall, the limit of agreement for the Fitbit device was 24 BPM, but its performance was significantly better in patients in sinus rhythm than in those who were not in sinus rhythm (average bias: −0.99 BPM vs −5.02 BPM, respectively; P=.02; limits of agreement: 22.9 BPM vs 46.4 BPM, respectively; P=.049) [11]. Kroll et al [11,31] also found that the Fitbit was very specific when it detected tachycardia (sensitivity=70%; specificity=99%) [31]. Steinhubl et al [32] demonstrated that manual and automated heart rate readings correlated well (R=0.75; measurements: n=111), but limits of agreement were not reported [32]. Weenk et al [4] reported that heart rate readings were generally consistent when compared to the nurse recordings; the limits of agreement for the ViSi Mobile and the HealthPatch were −11.1 to 10.7 BPM and −12.6 to 9.5 BPM, respectively (86 measurements).

Sleep
A total of 6 studies used wearables to assess sleep, of which 2 assessed whether wearable readings were reliable. Gallo and Lee [13] found that self-reported sleep correlated with the actigraphy-recorded number of awakenings (R=0.53; P=.01) [13]. Kroll et al [11,31] found that there was a moderate correlation between wearable-derived sleep duration and questionnaire-derived sleep quality (R=0.33; P=.03) [31].
Respiratory Rate

Of the 8 articles that used different wearables to measure the respiratory rate of patients, 3 assessed the wearables’ accuracy. Breteler et al [30] found that for respiratory rate, the bias was −2.3 breaths per minute, and wide limits of agreement were reported (−15.8 to 11.2 breaths per minute; measurement pairs: n=56,674) [30]. Steinhubl et al [32] reported that there was a strong correlation between wearable and manual respiratory rate readings (R=0.83; P<0.001; measurements: n=111), but limits of agreement were not reported [32]. Weenk et al [4] described wide limits of agreement for respiratory rate based on 86 measurements (ViSi Mobile limits of agreement: −5.5 to 7.9 breaths per minute; HealthPatch limits of agreement: −10.3 to 9.0 breaths per minute) [4].

Other Measures

Only 1 study, which was conducted by Weenk et al [4], assessed the accuracy of oxygen saturation and blood pressure readings from ViSi Mobile by comparing them to HealthPatch readings as well as intermittent nurse measurements. From 86 measurements, they found that the automated readings for the systolic blood pressure, diastolic blood pressure, and oxygen saturation had wide limits of agreement (systolic blood pressure: −23.1 to 24.0 mm Hg; diastolic blood pressure: −27.5 to 11.5 mm Hg; oxygen saturation: −3.1% to 3.3%) [4]. Of the 6 articles that used wearables that measured skin temperature, only Steinhubl et al [32] validated the results against a reference standard to conclude that the automated readings were reliable (R=0.99; n=112), but bias and limits of agreement were not reported [32]. Of the 3 articles in this review that detected falls by using wearables, only Bloch and colleagues [29] assessed accuracy and found that the Vigi’Fall system had a low sensitivity (37.5%) to fall risk [29].

Discussion

Principal Findings

We conducted a systematic review that evaluated the utility of wearable technology in continuously monitoring hospitalized patients for a wide variety of health parameters. Our review focused on the breadth of devices used and the signals measured in hospitalized patients and included consumer, research, and medical-grade devices. There was evidence to support the use of Fitbit, ViSi Mobile, and the HealthPatch to measure heart rate [4,11,31], since the readings were validated against both intermittent and continuous reference standards. This review demonstrated that the validity of the data did not necessarily correlate with the classification of the device because even some medical-grade devices did not perform well and yielded data with wide limits of agreement. We found that only 6 studies validated the accuracy of wearable-derived health data from hospitalized patients by comparing the readings against a reference standard. Overall, the quality of most of these studies was excellent in terms of the reporting of missing data (6/6, 100%) and the use of acceptable accuracy evaluations (5/6, 83%). However, there was a considerable risk of bias in these studies due to the low number of participants in most of the studies (4/6, 67%). Many studies reported wide limits of agreement for other digital biomarkers, such as respiratory rate and blood pressure. Of note, we also found that the majority of studies (8/14, 57%) did not validate the studied device or parameter measured.

Of the various health parameters, the best evidence of validity was in the monitoring of heart rate in hospitalized patients. We also found that, in hospital settings, limits of agreement for medical-grade devices ranged from 16.4 to 21.8 BPM, whereas the limit for a Fitbit consumer device that uses photoplethysmography signals was 24 BPM. Further, during Fitbit-based continuous electrocardiogram monitoring, 73% of the readings were within 5 BPM of electrocardiogram readings. In a systematic review of 158 studies that measured heart rate by using consumer wearable devices, 71% and 51% of Apple Watch (Apple Inc) readings (used in 49 studies) and Fitbit readings (used in 71 studies), respectively, were within 3% of electrocardiogram readings in controlled settings [39]. Moreover, in 3 free-living studies, the wrist-worn Fitbit Charge had a mean absolute error percentage of 10% [39]. A systematic review of wrist-worn devices that measure heart rate via plethysmography found limits of agreement of 8.4 BPM at rest, 30.1 BPM while on a treadmill, and 41.5 BPM while cycling [40]. Overall, our findings found large limits of agreement for all devices, and inpatient results were consistent with the wide limits of agreement found in free-living environments or with activity.

We found that sleep only had a moderate correlation with sleep survey results from inpatient settings the use research and consumer devices. A recent systematic review of Fitbit-based sleep assessments found that readings from more recently developed devices correlated well with polysomnography readings for assessing sleep episodes [41]. It is unclear whether the lower correlation that we found was due to inpatient settings with high nighttime interruptions, patient factors that were perhaps associated with acute illness, or issues with sleep surveys (or a combination of these three factors) [3]. With respect to respiratory rate, 2 studies of 2 medical-grade devices provided limits of agreement. Wider limits of agreement were found in the study that had over 50,000 measurement pairs and used a gold standard (27 breaths per minute) compared to those in the study that had less than 100 measurement pairs and used clinician-reported vitals (13.4-19.3 breaths per minute) [30,32]. Additionally, previous studies found that medical-grade devices were only accurate under laboratory conditions or at-home conditions [42,43]. There was a limited number of studies on oxygen saturation, temperature, blood pressure, and fall risk.

Limitations and Future Research

There are a few limitations that should be noted for our systematic review. There is a considerable risk of bias, as the number of participants in the studies was low. Further, the studies included were observational in design and had a high degree of heterogeneity in terms of the objectives, populations, and outcomes reported. Thus, the data analysis methods were limited to broad categorization and the extraction of the common themes and trends that emerged from the results. Reports of wearable monitoring from individual studies should be viewed based on their methodological limitations. Although patient adherence has been found to correlate well with patients’
acceptability of wearables devices in inpatient settings, we realize that studying factors such as data loss, the duration of data gaps, and qualitative feedback from nurses and patients would further strengthen the generalizability of the results. Finally, it is important to note that wearable studies are being increasingly performed, and more relevant articles will become increasingly available.

This review also identifies gaps in knowledge that still exist within literature and provides information about what is required for further research. Specifically, the further validation of digital biomarkers by using gold standard comparators, such as polysomnography for assessing sleep and continuous electrocardiogram monitoring for assessing heart rate, is required. Ideally, large participant sample sizes and large numbers of measurement pairs within a population of interest should be used to assess parameters such as vital signs. The use of 2 reference standards to validate each health parameter, such as a heart rate, has also been recommended [44]. Moreover, data that are derived under real life conditions are still needed to better understand the factors that may contribute to between-patient heterogeneity when comparing the accuracy of wearable readings, such as those for patient activity, posture, gait type and velocity, locations of wearables, and patients’ diagnoses (eg, seizures). Future studies can aim to further qualify the process of retrieving data by using wearables to explore other barriers and avenues that might hinder the collection of reliable health information (ie, a weak Bluetooth connectivity, a lack of patient digital health literacy, the added burden that the process of taking wearable readings has on clinicians, the learning curve required to operate a wearable, etc) Finally, while we found that some digital biomarkers appeared to be valid for the monitoring of inpatients via wearables, we were unable to find any studies that supported the use of wearables in inpatient settings to improve clinical outcomes.

Conclusions

Overall, the assessment of studies in this review suggested that wearable devices show promise for monitoring the heart rate and sleep of patients in hospitals. The results show that many devices were not validated in inpatient settings, and the readings from most of the devices that were validated in such settings had wide limits of agreement. Further research is needed to determine the accuracy of the digital biomarker readings of hospitalized patients and to eventually determine whether wearable devices improve the health outcomes of hospitalized patients.

Acknowledgments

RW is supported by an award from the Mak Pak Chiu and Mak-Soo Lai Hing Chair of General Internal Medicine, University of Toronto. Funding for this study was kindly provided by the University Health Network Foundation Complex Care Fund.

Authors’ Contributions

VP, AOC, and RW designed and planned the review. AOC conducted the search strategy. VP and RW screened the articles and conducted the data analysis. VP, AOC, and RW wrote and revised the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Search methods.

[DOCX File, 64 KB - mhealth_v9i8e17411_app1.docx ]

Multimedia Appendix 2

Modified Consensus-Based Standards for the Selection of Health Status Measurement Instruments (COSMIN) criteria used for the risk of bias assessment.

[DOCX File, 30 KB - mhealth_v9i8e17411_app2.docx ]

References


Abbreviations

BPM: beats per minute

COSMIN: Consensus-Based Standards for the Selection of Health Status Measurement Instruments
Comparison of the Validity and Generalizability of Machine Learning Algorithms for the Prediction of Energy Expenditure: Validation Study

Ruairi O'Driscoll¹, MSc; Jake Turicchi¹, BSc; Mark Hopkins², PhD; Cristiana Duarte¹, PhD; Graham W Horgan³, PhD; Graham Finlayson¹, PhD; R James Stubbs¹, PhD

¹Appetite Control and Energy Balance Group, School of Psychology, University of Leeds, Leeds, United Kingdom
²School of Food Science and Nutrition, Faculty of Environment, University of Leeds, Leeds, United Kingdom
³Biomathematics & Statistics Scotland, Aberdeen, United Kingdom

Corresponding Author:
Ruairi O'Driscoll, MSc
Appetite Control and Energy Balance Group
School of Psychology
University of Leeds
Woodhouse
Leeds
United Kingdom
Phone: 44 113 343 2846
Email: psrod@leeds.ac.uk

Abstract

Background: Accurate solutions for the estimation of physical activity and energy expenditure at scale are needed for a range of medical and health research fields. Machine learning techniques show promise in research-grade accelerometers, and some evidence indicates that these techniques can be applied to more scalable commercial devices.

Objective: This study aims to test the validity and out-of-sample generalizability of algorithms for the prediction of energy expenditure in several wearables (ie, Fitbit Charge 2, ActiGraph GT3-x, SenseWear Armband Mini, and Polar H7) using two laboratory data sets comprising different activities.

Methods: Two laboratory studies (study 1: n=59, age 44.4 years, weight 75.7 kg; study 2: n=30, age=31.9 years, weight=70.6 kg), in which adult participants performed a sequential lab-based activity protocol consisting of resting, household, ambulatory, and nonambulatory tasks, were combined in this study. In both studies, accelerometer and physiological data were collected from the wearables alongside energy expenditure using indirect calorimetry. Three regression algorithms were used to predict metabolic equivalents (METs; ie, random forest, gradient boosting, and neural networks), and five classification algorithms (ie, k-nearest neighbor, support vector machine, random forest, gradient boosting, and neural networks) were used for physical activity intensity classification as sedentary, light, or moderate to vigorous. Algorithms were evaluated using leave-one-subject-out cross-validations and out-of-sample validations.

Results: The root mean square error (RMSE) was lowest for gradient boosting applied to SenseWear and Polar H7 data (0.91 METs), and in the classification task, gradient boost applied to SenseWear and Polar H7 was the most accurate (85.5%). Fitbit models achieved an RMSE of 1.36 METs and 78.2% accuracy for classification. Errors tended to increase in out-of-sample validations with the SenseWear neural network achieving RMSE values of 1.22 METs in the regression tasks and the SenseWear gradient boost and random forest achieving an accuracy of 80% in classification tasks.

Conclusions: Algorithms trained on combined data sets demonstrated high predictive accuracy, with a tendency for superior performance of random forests and gradient boosting for most but not all wearable devices. Predictions were poorer in the between-study validations, which creates uncertainty regarding the generalizability of the tested algorithms.

(JMIR Mhealth Uhealth 2021;9(8):e23938) doi:10.2196/23938

KEYWORDS

bioenergetics; energy balance; accelerometers; machine learning; validation
**Introduction**

**Background**

Participation in physical activity results in increased energy expenditure [1] and represents a key modifiable risk factor for cardiovascular disease, obesity, diabetes mellitus, cancer, and mortality [2]. Thus, longitudinal, unobtrusive, and accurate measurement of intraindividual physical activity energy expenditure would be highly valuable for health research. Activity trackers offer a scalable means for the continuous collection of physical activity data in free-living environments and, by extension, the measurement of energy expenditure. Unfortunately, the accuracy of activity trackers varies greatly between devices and activities [3,4], which limits their use when quantifying energy balance and activity behaviors.

The potential of machine learning techniques to model the complex interactions of accelerometer data, physiological variables, and the rate of energy expenditure has been recognized for some time. Rothney et al [5] trained an artificial neural network using raw accelerometer data as input to predict the energy expenditure in a whole-body calorimetry chamber. Pober et al [6] used quadratic discriminant analysis and a hidden Markov model to classify activity and subsequently estimated the proportion of time performing different activities. Research groups have built on these early findings and have reported highly accurate algorithms for a variety of activities [7-11]. Researchers often take two broad approaches when modeling physical activities: first, attempting to predict the rate of energy expenditure, and second, classifying a minute as sedentary activity, light physical activity, or moderate-to-vigorous physical activity (MVPA), both of which are important for health research. Regression approaches can be used to derive the total energy expenditure for a subject and this can subsequently be incorporated into energy balance models to calculate energy intake [12]. Alternatively, accurately determining the time an individual spends in broader categories of activity or the intensity of that activity can be important for public health guidance. For example, successful weight maintenance in the National Weight Control Registry and weight management recommendations are often defined based on the time an individual spends in MVPA [13]. Machine learning algorithms have the potential to enhance physical activity assessment beyond that of traditional count-based methods, which despite being more accessible, may not be sufficiently accurate for the assessment of energy expenditure and intensity classifications [14].

Recently, we demonstrated in a laboratory validation study that accelerometer and physiological sensor outputs can be modeled using random forests to predict the rate of energy expenditure (as a multiple of resting energy expenditure) in commercial and research-grade activity monitors. We demonstrated a low error in the prediction of energy expenditure [15]. The number of activities in which energy expenditure was measured in this study was limited, and the generalizability of these algorithms remains uncertain. A method for continued refinement of predictive algorithms is to obtain more than one data set [16] to provide larger, more diverse training data with more activities. More data present a new optimization problem, which (because of different assumptions made by different algorithms) means that there is no guarantee that any algorithm will minimize error on all problems [17]. For machine learning models to be used in general health research settings, it is critical to evaluate the generalizability of prediction algorithms. The extent to which an algorithm will generalize is influenced by the characteristics of the sample, activity types, size, and quality of the training data. One approach that addresses each of these limitations is to evaluate prediction algorithms on different samples using data collected under different conditions. In addition to generalizability, a combination of heterogeneous data sets collected under different experimental conditions may help to increase the accuracy of predictions [18].

**Objectives**

In this study, two distinct data sets of concurrent inputs from multiple wearable devices (ie, Fitbit Charge 2, ActiGraph GT3-x, SenseWear Armband Mini, and a polar chest strap) and measured energy expenditure (indirect calorimetry) are combined to develop predictive models of minute-level energy expenditure and physical activity. We aim to evaluate classification and regression algorithms to (1) predict the rate of energy expenditure and (2) classify a single minute as sedentary activity, light physical activity, or MVPA. Algorithms were validated using leave-one-subject-out cross-validation (LOSO) and out-of-sample validation. Concurrently, we evaluated the SenseWear armband, a device that has been shown to outperform accelerometer-based monitors when classifying activity minutes [19] and is one of the most accurate wrist or arm-based monitors for estimating energy expenditure [3].

**Methods**

**Studies**

This study aggregated the data collected as part of two separate studies at the Human Appetite Research Unit, University of Leeds. Participants were recruited from the local area using word-of-mouth and recruitment emails. Participants must have been at least 18 years of age, have been able to attend the research laboratory at the required intervals, be able to ambulate without assistance, they must not have been taking medications known to alter metabolic rate, and participants must not have had any cardiovascular, metabolic, renal disorders, illness, or injury that would increase the risk of medical events during physical activity. Both studies were approved by the University of Leeds, School of Psychology Ethics Committee (PSC-407 and PSC-744 for study 1 and 2, respectively), and all participants provided informed consent before participation in the study. The participant information for the samples is shown in Table 1. Study 2 had proportionately more males, lower age, lower average percentage of fat mass (FM), and a higher resting metabolic rate (RMR) on average.
Protocols
Study 1
The details of study 1 have been published previously [15]. The protocol of study 1 consisted of 10 activities, each performed for 5 minutes in the following order: sitting, standing, treadmill walking and incline walking (4 km/h), jogging, and incline jogging (6-8 km/h). Participants then rested for 3 minutes and transitioned to a cycle ergometer for low- and moderate-intensity cycling. After another period of recovery, participants performed a folding and sweeping task. Owing to a variation in physical fitness, the jogging task (n=49), incline jogging (n=30), and moderate cycling tasks (n=58) were not performed by all participants.

Study 2
In study 2 (total energy expenditure from wearable devices study), participants visited the lab and refrained from eating or consuming caffeine for at least 4 hours. This exercise visit is the first of three visits to the laboratory conducted as part of a wider project. Weight and height were obtained from a SECA 704s stadiometer and electronic scale (SECA, Germany), and subsequently, an activity protocol was performed. All activities were performed in 5-minute increments, and the order was identical for all participants. First, resting tasks were performed where participants lay supine, sat in a backed chair, and then stood. Next, after a 2-minute unstructured transitional period, participants performed seated typing, standing ironing, and wiping surfaces while standing. After another 2-minute transition, participants walked on a treadmill at 4 km/h, walked at an incline of 5% at 4 km/h, and subsequently jogged at 7 km/h. The participants then rested for 10 minutes. After the unstructured resting period, participants performed low-intensity and moderate-intensity cycling, low-intensity and moderate-intensity rowing, and low-intensity and moderate-intensity cross-training (elliptical), with 1-minute transitions between each, and the intensity of the tasks was determined by a self-selected perceived exertion. In study 2, one participant did not perform rowing or elliptical tasks.

Body Composition Assessment
In both studies, body composition was estimated using air displacement plethysmography (BodPod, Life Measurement, Inc), n=57 in study 1 and n=30 in study 2. Study 2 is part of a wider study in which participants visited the laboratory three times, the first of which was the laboratory validation reported here. Body composition was measured at a subsequent visit to the laboratory in a fasting state.

Energy Expenditure
This study used metabolic equivalents (METs) as the outcome variable, which served to eliminate the proportion of energy expenditure attributable to RMR. We first established the RMR of each participant, which was measured in the fasting state, before any exercise. In both studies, RMR was determined from VO\(_2\) and VCO\(_2\) data collected through a ventilated hood indirect calorimeter system (gas exchange measurement; Nutren Technology Ltd). In study 1, RMR was measured before exercise testing, and in study 2, which occurred on a subsequent visit to the laboratory. After researchers explained the procedures to the participants and an initial calibration process (approximately 10 minutes), VO\(_2\) and VCO\(_2\) were measured for 30 minutes in the supine position. The RMR was established from the VO\(_2\) and VCO\(_2\) of the 5-minute block with the lowest coefficient of variation [20]. If RMR data were unavailable (n=3 across both studies), we approximated the RMR with BMI-specific equations [21]. During the activity sessions, energy expenditure was obtained from a stationary metabolic cart (Vyntus CPX, Jaeger-CareFusion), and these data were expressed relative to the measured RMR of each subject to derive METs. Definitions of METs are inconsistent [22] and we took an individualized approach to METs calculations because the standard definition of METs may have limited applicability in some subjects [23].

Devices
Accelerometer and physiological data were collected using various sensors in both protocols. The Polar H7 chest strap (Polar Electro) was used to measure the heart rate. An ActiGraph GT3-X accelerometer (ActiGraph) and a Fitbit Charge 2 (Fitbit Inc) were attached securely to the nondominant wrist. Participants also wore the SenseWear Armband Mini (BodyMedia Inc) on the upper arm.

Data Aggregation
The sensor outputs were obtained from the device-specific software and aggregated to the minute level and time matched to the criterion energy expenditure data. Data loss attributable
to device malfunction was as follows: in study 1, Fitbit data of 2 participants, ActiGraph data of 1 participant, and polar heart rate data of 1 participant were lost. In study 2, 1 SenseWear and 1 Fitbit data set were lost because of device failure. Given the slightly different data availability in each model, our results report the number of minutes used and the number of participants. All minutes in which energy expenditure data were available (ie, face mask was not removed) were included in this analysis, and the aggregation of the data sets by time was conducted in Python 3.7.6 and R version 3.6.3 (R Core Team).

For activity-specific analyses, we grouped activities into broader categories. Activities of daily living, which involved folding, sweeping, typing, ironing, and wiping surfaces. Distinct categories were assigned for cycling, elliptical, rowing, running, and walking. The sedentary activities involved all sitting, standing, and supine tasks. The transitional category refers to unstructured resting or transitional minutes.

Features

Predictive models were built for Fitbit, ActiGraph, and SenseWear, and the features used in each model are listed in Table 2. Each device used a combination of subject-level features, accelerometer features, and physiological features, which have been related to the rate of energy expenditure in previous studies [3,5,24-26]. The features varied depending on the feature availability of each device. Where small (limit of 5 minutes) heart rate gaps existed (eg, loss of signal between the respective heart rate sensor and the skin), we used linear interpolation to fill gaps. As activity in the preceding minutes influences the rate of energy expenditure at the measurement point [27], some time-lagged features were computed: for steps (Fitbit and SenseWear), vector magnitude (ActiGraph), Fitbit heart rate (Fitbit), and polar heart rate (SenseWear and ActiGraph), the change from t-1 minutes for each minute up to t-5 minutes were included as predictive features. In addition, the mean and SD of the current and last 5 minutes were used as predictive features. If time-lagged variables could not be computed due to missing data (ie, for the first minutes for each subject), we imputed backward using the next available observation.

As a constant variance is important for some of the algorithms tested in this study, all numeric features were standardized before training using the following formula:

\[ z = (x - \mu) / sd \]  

(1)

where \( \mu \) and \( sd \) refer to the variable mean and SD, respectively.
Table 2. Predictive features used in each of the models.

<table>
<thead>
<tr>
<th>Device(^a) and category</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fitbit</strong></td>
<td></td>
</tr>
<tr>
<td>Subject features</td>
<td>Gender, age, height, weight, and sitting heart rate</td>
</tr>
<tr>
<td>Acceleration features</td>
<td>Steps features: steps mean, steps difference (t-1, t-2, t-3, t-4, and t-5 minutes); steps mean and SD of last 5 minutes</td>
</tr>
<tr>
<td>Physiological features</td>
<td>Fitbit heart rate features: Fitbit heart rate above sitting heart rate, Fitbit heart rate percentage of maximum heart rate, Fitbit heart rate mean, Fitbit heart rate difference (t-1, t-2, t-3, t-4, and t-5 minutes), and Fitbit heart rate mean and SD of last 5 minutes</td>
</tr>
<tr>
<td><strong>ActiGraph</strong></td>
<td></td>
</tr>
<tr>
<td>Subject features</td>
<td>Gender, age, height, and weight</td>
</tr>
<tr>
<td>Acceleration features</td>
<td>X, Y, Z features: minimum, maximum, mean, SD; median crossings; 10th, 25th, 50th, 75th, 90th percentiles; correlations (XY, XZ, YZ); dominant frequency; dominant frequency magnitude First order differential of X, Y, Z features: minimum, maximum, mean, SD; median crossings; 10th, 25th, 50th, 75th, and 90th percentiles; correlations (XY, XZ, YZ); dominant frequency; dominant frequency magnitude Vector magnitude features: vector magnitude mean; vector magnitude difference (t-1, t-2, t-3, t-4, and t-5 minutes); vector magnitude mean and SD of last 5 minutes</td>
</tr>
<tr>
<td>Physiological features</td>
<td>Polar heart rate features: polar heart rate above sitting heart rate; polar heart rate percentage of maximum heart rate; polar heart rate mean; polar heart rate difference (t-1, t-2, t-3, t-4, and t-5 minutes); polar heart rate mean and SD of last 5 minutes</td>
</tr>
<tr>
<td><strong>SenseWear</strong></td>
<td></td>
</tr>
<tr>
<td>Subject features</td>
<td>Gender, age, height, and weight</td>
</tr>
<tr>
<td>Acceleration features</td>
<td>X, Y, Z features: peaks, mean of absolute differences, average; Steps features: steps mean; steps difference (t-1, t-2, t-3, t-4, and t-5 minutes); steps mean and SD of last 5 minutes</td>
</tr>
<tr>
<td>Physiological features</td>
<td>Polar heart rate features: polar heart rate above sitting heart rate; polar heart rate percentage of maximum heart rate; polar heart rate mean; polar heart rate difference (t-1, t-2, t-3, t-4, and t-5 minutes); polar heart rate mean and SD of last 5 minutes; and SenseWear sensors: near body temperature average, Galvanic skin response average, skin temperature average</td>
</tr>
</tbody>
</table>

\(^a\)For each device, the subject characteristics, acceleration features, and physiological features are listed.

**Algorithms**

The SenseWear outputs a MET estimate that we evaluated in this study (SenseWear manufacturer). We also tested several machine learning algorithms for regression and classification tasks, which are described below. In the regression tasks, algorithms predicted a MET value for each minute, and in the classification tasks, algorithms classified activity categories for each minute. The activity classifications were as follows: sedentary activity (≤1.5 METs), light physical activity (>1.5 and <3 METs), and MVPA (≥3.0 METs) [18,28,29]. For each algorithm, the hyperparameters were informed by a random search through a range of potential hyperparameters in the preliminary tuning experiments. Random search iterates over a grid of randomly selected combinations of hyperparameters, rather than exploring every possible combination of features, and therefore offers a significant computational advantage over a grid-search approach [30]. Each random search was conducted with the RandomizedSearchCV class in Scikit Learn [31], using three-fold cross-validation. The specific parameters for each algorithm are detailed in Multimedia Appendix 1, and except for the neural network models (explained in the following section), the scoring or loss criterion was the default loss or scoring metrics within Scikit Learn. All algorithms were trained using Keras-GPU [32] or Scikit Learn [31].

**Random Forest**

The random forest algorithm was used for regression and classification tasks [33]. Random forests involve training of multiple decision trees on data subsamples. Importantly, when splitting these decision trees, only a subsample of the potential predictors is used, which serves to decorrelate the trees. The predictions of each tree can then be combined to produce a majority vote (classification) or continuous prediction (regression). The optimal hyperparameters of the algorithm were estimated in the tuning experiments and included the number of trees, number of samples required to split a tree, number of samples per leaf, total predictors, and the depth of
trees. In regression, the quality of a split was assessed with mean square error, and in classification, Gini impurity was used. Algorithms were implemented using the RandomForestClassifier and RandomForestRegressor classes in Scikit Learn [31].

Gradient Boosting

For the regression and classification tasks, we used the gradient boosting algorithm. Similar to random forests, this algorithm is a tree-based ensemble method. However, when random forests may be considered to use a bagging approach, gradient boosting uses boosting to learn. Boosting involves the sequential growth of small (weak) decision trees. Each tree is trained using the residuals of the previous estimator and subsequently added to the fitted function to update the residuals. In the boosting phase, a learning rate parameter penalizes the contribution of each tree to the overall model, thereby slowing the learning [34]. The gradient boosting hyperparameters were tuned in the random search experiments and included the number of boosting stages, the maximum depth of the estimators, learning rate, number of samples required to split a node, the number of samples per leaf, and the maximum number of predictors. In the regression, the loss function was least squares, and in classification, deviance was used. Algorithms were implemented using the GradientBoostingClassifier and GradientBoostingRegressor classes in Scikit Learn [31].

Neural Networks

The third algorithm, used in both regression and classification tasks, was artificial neural networks. Neural networks allow complex, nonlinear functions to be modeled and comprise layers of interconnected neurons. At each neuron, inputs are subjected to a numerical activation function, and then passed through subsequent hidden layers of neurons to an output layer [34,35]. In the training process, the interneuronal weights of the network are refined relative to a loss function (ie, mean square error or cross-entropy). Neural networks in the classification studies used the sparse categorical cross-entropy loss function, and in the regression setting, the loss was the mean square error. We tuned the learning rate of each network, the number of layers, and the number of neurons. Neural networks hidden layers used the relu activation function, and classification models used a softmax activation in the output layer, both classification and regression networks used the Adam optimizer.

K-Nearest Neighbors

For classification tasks, we tested the k-nearest neighbor (KNN) algorithm. This algorithm assigns a given point to a particular class based on the majority class of the k nearest neighbors, where the neighbors of a given point are defined by a distance metric (ie, Euclidian, Minkowski, or Manhattan) [34]. Hyperparameters adjusted in the training process included the number of neighbors in each neighborhood (k), distance metrics, and the weight applied to each of the observations in a neighborhood. KNN was implemented with Scikit Learn [31], using the KNeighborsClassifier class.

Support Vector Machine

The final classification model tested was a support vector machine classifier with a radial basis function [35]. A support vector machine aims to find a separating hyperplane between classes by maximizing the distance between the points and the hyperplane. In this study, we tuned the regularization parameter (C) and gamma, which defines the magnitude of the effect of specific training examples. The support vector machine classifier was implemented with the SVC class in Scikit Learn [31].

Statistical Analyses

We conducted two validation approaches for all the analyses and algorithms. First, LOSO validations, where algorithms are trained on all but the data of 1 participant, and the participant is held back for validation. This process was repeated until all participants had served as the validation participant once. Second, we used an out-of-sample validation in which the entire data set from one study was used as training data, and the second study was used as an out-of-sample validation. Regression algorithms were evaluated by root mean square error (RMSE), mean absolute percentage error (MAPE) with the Metrics package in R and concordance correlation coefficient (CCC) with DescTools. Agreement statistics were calculated at the minute level; however, for visualization purposes, we computed the RMSE at the level of individuals and plotted these values. Equivalence tests were used to determine if the true METs and predicted METs were statistically equivalent; tests used equivalence bounds of 10%, and to be considered equivalent, the 90% CI must fall within the equivalence bounds. Finally, linear mixed models with a random intercept of subject ID were used to investigate differences in RMSE between the models. Comparisons were conducted using the Lme4 [36] package in R, with P values adjusted by the Bonferroni method in post hoc comparisons. For classification tasks, we report the χ statistic, which compares the accuracy of the predictions to that of a random system. We also report accuracy, where accuracy is the proportion of cases that were classified correctly and the F1 score. All classification statistics were calculated using the Caret [37] package in R. A P value of <.05 was used to determine statistical significance, where P values were reported.

Results

Regression

A total of 89 participant activity sessions were included in this sample, and all models could be evaluated on at least 5448 minutes of data in the LOSO validations.

The regression algorithms predicting energy expenditure are presented for minute-level data in Table 3 and are visually displayed in Figure 1. Our results demonstrate that the greatest error in METs was observed for the manufacturer-provided SenseWear estimates, with MAPE and RMSE values of 34.54 and 1.86, respectively. For ActiGraph, the RMSE was lowest for gradient boosting (0.93 METs), which also achieved the lowest MAPE of any ActiGraph model (17.88%). Of the Fitbit models, the random forest and gradient boosting had equal RMSE (1.36 METs), but a slightly lower MAPE was achieved by the random forest. For the SenseWear, the gradient boost had the lowest RMSE value (0.91 METs), and this was the lowest RMSE of all those tested. The neural network models were associated with a greater overall RMSE for the ActiGraph, Fitbit, and SenseWear models.
Activity-specific MET predictions are presented in Multimedia Appendix 2, and the RMSE is shown in Figure 2. For all activities tested, tree-based models (gradient boost or random forest) applied to ActiGraph or SenseWear data were superior, as measured by RMSE. The manufacturer estimates of SenseWear had the highest RMSE for all activities aside from sedentary activities, in which only the ActiGraph gradient boost and random forest had a lower RMSE. Notably, all Fitbit models overestimated sedentary activities and had the highest RMSE in this category. The pairwise comparisons between models are presented in Multimedia Appendix 3 for each of the comparisons shown in Figure 1 and Figure 2. An example of the model predictions for a single subject is shown in Figure 3.

Table 3 shows the statistics for the between-study predictions. Notably larger errors were observed relative to the LOSO validations, with the Fitbit gradient boost reaching a RMSE of 1.92 METs (neural network) when study 1 was used as the training data. To estimate the relative importance of each of the features used in each model, permutation importance has been reported in Multimedia Appendix 4.

### Table 3. Leave-one-subject-out cross-validation results for each of the regression models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Minutes</th>
<th>Participants, n (%)</th>
<th>Predicted (METs), mean (SD)</th>
<th>True (METs), mean (SD)</th>
<th>MAPE</th>
<th>RMSE</th>
<th>CCC (95% CI)</th>
<th>Equivalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWA manufacturer</td>
<td>5533</td>
<td>88 (99)</td>
<td>3.8 (2.49)</td>
<td>4.04 (2.59)</td>
<td>34.54</td>
<td>1.86</td>
<td>0.73 (0.72-0.74)</td>
<td>Non-equivalent</td>
</tr>
<tr>
<td>AG gradient boost</td>
<td>5517</td>
<td>87 (98)</td>
<td>4.04 (2.35)</td>
<td>4.04 (2.59)</td>
<td>17.88</td>
<td>0.93</td>
<td>0.93 (0.93-0.93)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>AG neural network</td>
<td>5517</td>
<td>87 (98)</td>
<td>4.05 (2.55)</td>
<td>4.04 (2.59)</td>
<td>21.65</td>
<td>1.14</td>
<td>0.9 (0.9-0.91)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>AG random forest</td>
<td>5517</td>
<td>87 (98)</td>
<td>4.05 (2.32)</td>
<td>4.04 (2.59)</td>
<td>18.36</td>
<td>0.94</td>
<td>0.93 (0.92-0.93)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>FB gradient boost</td>
<td>5448</td>
<td>86 (97)</td>
<td>4.03 (2.19)</td>
<td>4.01 (2.58)</td>
<td>30.22</td>
<td>1.36</td>
<td>0.84 (0.83-0.84)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>FB neural network</td>
<td>5448</td>
<td>86 (97)</td>
<td>4.02 (2.28)</td>
<td>4.01 (2.58)</td>
<td>32.27</td>
<td>1.45</td>
<td>0.82 (0.82-0.83)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>FB random forest</td>
<td>5448</td>
<td>86 (97)</td>
<td>4.03 (2.14)</td>
<td>4.01 (2.58)</td>
<td>30.10</td>
<td>1.36</td>
<td>0.84 (0.83-0.84)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>SWA gradient boost</td>
<td>5492</td>
<td>87 (98)</td>
<td>4.04 (2.39)</td>
<td>4.04 (2.6)</td>
<td>17.83</td>
<td>0.91</td>
<td>0.93 (0.93-0.94)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>SWA neural network</td>
<td>5492</td>
<td>87 (98)</td>
<td>4.05 (2.47)</td>
<td>4.04 (2.6)</td>
<td>19.56</td>
<td>0.96</td>
<td>0.93 (0.92-0.93)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>SWA random forest</td>
<td>5492</td>
<td>87 (98)</td>
<td>4.05 (2.35)</td>
<td>4.04 (2.6)</td>
<td>18.25</td>
<td>0.92</td>
<td>0.93 (0.93-0.93)</td>
<td>Equivalent</td>
</tr>
</tbody>
</table>

**Notes:**
- Minutes refers to the number of minutes the algorithms are validated on.
- METs: metabolic equivalents.
- MAPE: mean absolute percentage error.
- RMSE: root mean square error.
- CCC: concordance correlation coefficient CCC is presented with 95% CIs.
- SWA: SenseWear.
- AG: ActiGraph.
- FB: Fitbit.

### Footnotes:
- a: Minutes refers to the number of minutes the algorithms are validated on.
- b: METs: metabolic equivalents.
- c: MAPE: mean absolute percentage error.
- d: RMSE: root mean square error.
- e: CCC: concordance correlation coefficient CCC is presented with 95% CIs.
- f: SWA: SenseWear.
- g: The model is not statistically equivalent to the criterion.
- h: AG: ActiGraph.
- i: Equivalent implies that the model is statistically equivalent to the criterion.
- j: FB: Fitbit.
Figure 1. Boxplots demonstrating the root mean square error overall for each of the tested models. AG: ActiGraph; FB: Fitbit; RMSE: root mean square error; SWA: SenseWear.

Figure 2. Boxplots demonstrating the root mean square error for each of the tested models in specific activity categories. ADL: activities of daily living; AG: ActiGraph; FB: Fitbit; RMSE: root mean square error; SWA: SenseWear.
Figure 3. A time series plot showing metabolic equivalents predicted by the models tested in this study (colored solid line) and indirect calorimeter (black dashed line), for a single subject in study 2. The x-axis represents the measurement time. Minutes 1-15=sedentary; minutes 16-17=transitional/unstructured; minutes 18-32=activities of daily living (typing, wiping surfaces, and ironing); minutes 33-34=transitional/unstructured; minutes 35-44=walking; minutes 45-49=running; minutes 50-59=transitional/unstructured; minutes 60-69=cycling; minutes 71-80=rowing; and minutes 82-91=elliptical. Participants performed cycling, rowing, and elliptical tasks at self-selected low and moderate intensity for 5 minutes each. AG: ActiGraph; FB: Fitbit; METs: metabolic equivalents; SWA: SenseWear.
Table 4. Out-of-sample results for each of the regression models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Training data</th>
<th>Minutes&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Predicted (METs&lt;sup&gt;b&lt;/sup&gt;), mean (SD)</th>
<th>True (METs), mean (SD)</th>
<th>MAPE&lt;sup&gt;c&lt;/sup&gt;</th>
<th>RMSE&lt;sup&gt;d&lt;/sup&gt;</th>
<th>CCC&lt;sup&gt;e&lt;/sup&gt; (95% CI)</th>
<th>Equivalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>AG&lt;sup&gt;f&lt;/sup&gt; gradient boost</td>
<td>Study 1</td>
<td>2690</td>
<td>4.03 (1.9)</td>
<td>3.93 (2.66)</td>
<td>36.35</td>
<td>1.37</td>
<td>0.82 (0.81-0.83)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>AG neural network</td>
<td>Study 1</td>
<td>2690</td>
<td>4.07 (2.48)</td>
<td>3.93 (2.66)</td>
<td>29.75</td>
<td>1.33</td>
<td>0.87 (0.86-0.88)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>AG random forest</td>
<td>Study 1</td>
<td>2690</td>
<td>3.97 (1.79)</td>
<td>3.93 (2.66)</td>
<td>39.50</td>
<td>1.51</td>
<td>0.78 (0.77-0.79)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>FB&lt;sup&gt;h&lt;/sup&gt; gradient boost</td>
<td>Study 1</td>
<td>2630</td>
<td>3.76 (1.7)</td>
<td>3.88 (2.65)</td>
<td>47.55</td>
<td>1.89</td>
<td>0.64 (0.62-0.66)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>FB neural network</td>
<td>Study 1</td>
<td>2630</td>
<td>3.65 (1.86)</td>
<td>3.88 (2.65)</td>
<td>47.40</td>
<td>1.92</td>
<td>0.65 (0.63-0.67)</td>
<td>__</td>
</tr>
<tr>
<td>FB random forest</td>
<td>Study 1</td>
<td>2630</td>
<td>3.76 (1.66)</td>
<td>3.88 (2.65)</td>
<td>47.45</td>
<td>1.87</td>
<td>0.64 (0.63-0.66)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>SWA&lt;sup&gt;i&lt;/sup&gt; gradient boost</td>
<td>Study 1</td>
<td>2633</td>
<td>3.92 (2.13)</td>
<td>3.94 (2.68)</td>
<td>27.35</td>
<td>1.23</td>
<td>0.87 (0.86-0.88)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>SWA neural network</td>
<td>Study 1</td>
<td>2633</td>
<td>3.88 (2.26)</td>
<td>3.94 (2.68)</td>
<td>27.07</td>
<td>1.22</td>
<td>0.88 (0.87-0.89)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>SWA random forest</td>
<td>Study 1</td>
<td>2633</td>
<td>3.91 (2.07)</td>
<td>3.94 (2.68)</td>
<td>29.54</td>
<td>1.28</td>
<td>0.86 (0.85-0.87)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>AG gradient boost</td>
<td>Study 2</td>
<td>2827</td>
<td>4.46 (2.14)</td>
<td>4.15 (2.52)</td>
<td>31.49</td>
<td>1.36</td>
<td>0.83 (0.82-0.84)</td>
<td>__</td>
</tr>
<tr>
<td>AG neural network</td>
<td>Study 2</td>
<td>2827</td>
<td>4.24 (2.56)</td>
<td>4.15 (2.52)</td>
<td>29.00</td>
<td>1.42</td>
<td>0.84 (0.83-0.85)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>AG random forest</td>
<td>Study 2</td>
<td>2827</td>
<td>4.45 (2.1)</td>
<td>4.15 (2.52)</td>
<td>31.47</td>
<td>1.38</td>
<td>0.82 (0.81-0.84)</td>
<td>__</td>
</tr>
<tr>
<td>FB gradient boost</td>
<td>Study 2</td>
<td>2818</td>
<td>4.11 (2.06)</td>
<td>4.13 (2.51)</td>
<td>34.38</td>
<td>1.66</td>
<td>0.74 (0.72-0.75)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>FB neural network</td>
<td>Study 2</td>
<td>2818</td>
<td>4.01 (2.04)</td>
<td>4.13 (2.51)</td>
<td>33.10</td>
<td>1.56</td>
<td>0.77 (0.75-0.78)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>FB random forest</td>
<td>Study 2</td>
<td>2818</td>
<td>4.21 (2.04)</td>
<td>4.13 (2.51)</td>
<td>33.79</td>
<td>1.62</td>
<td>0.75 (0.73-0.77)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>SWA gradient boost</td>
<td>Study 2</td>
<td>2859</td>
<td>4.15 (2.13)</td>
<td>4.14 (2.51)</td>
<td>24.90</td>
<td>1.25</td>
<td>0.86 (0.85-0.87)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>SWA neural network</td>
<td>Study 2</td>
<td>2859</td>
<td>3.94 (2.36)</td>
<td>4.14 (2.51)</td>
<td>25.65</td>
<td>1.25</td>
<td>0.87 (0.86-0.88)</td>
<td>Equivalent</td>
</tr>
<tr>
<td>SWA random forest</td>
<td>Study 2</td>
<td>2859</td>
<td>4.2 (2.13)</td>
<td>4.14 (2.51)</td>
<td>25.72</td>
<td>1.26</td>
<td>0.85 (0.84-0.86)</td>
<td>Equivalent</td>
</tr>
</tbody>
</table>

<sup>a</sup>Minutes refers to the number of minutes the algorithms are validated on.

<sup>b</sup>METs: metabolic equivalents.

<sup>c</sup>MAPE: mean absolute percentage error.

<sup>d</sup>RMSE: root mean square error.

<sup>e</sup>CCC: concordance correlation coefficient CCC is presented with 95% CIs.

<sup>f</sup>AG: ActiGraph.

<sup>g</sup>Equivalent implies that the model is statistically equivalent to the criterion.

<sup>h</sup>FB: Fitbit.

<sup>i</sup>The model is not statistically equivalent to the criterion.

<sup>j</sup>SWA: SenseWear.

Classification

Figure 4 presents the results of the LOSO classification experiments for all classification algorithms and the SenseWear manufacturer estimates. Classes were slightly imbalanced, approximately 19.4% sedentary activity, 22.4% light physical activity, and 58.2% MVPA with small differences between devices due to data availability. The highest accuracy for Fitbit models was the random forest (78.21%), for the Actigraph models, the random forest achieved the highest accuracy (84.56%), and for the SenseWear models, the gradient boosting algorithm (85.49%) was the most accurate.

Multimedia Appendix 5 provides class-specific statistics for each model. Models tended to perform worse in light activity with F1 scores ranging from 0.20 (SenseWear neural network) to 0.66 (SenseWear gradient boost). In sedentary activities, the F1 score was improved with a range of 0.54 (Actigraph support vector machine) to 0.83 (four models). For MVPA, the F1 score ranged from 0.80 (Actigraph support vector machine) to 0.93 (three models).
Between-Study Predictions

The between-study classification accuracies are listed in Table 5. In most cases, when study 1 served as the training data, lower accuracy was observed. When study 1 served as the training data, the accuracy ranged from 0.55 (ActiGraph support vector machine) to 0.80 (two models). When study 2 served as the training data, the accuracy ranged from 0.65 (ActiGraph support vector machine) to 0.79 (three models).
Table 5. Between-study classification results for each of the classification models.

<table>
<thead>
<tr>
<th>Training data and model</th>
<th>Accuracy</th>
<th>κ</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study 1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AG(^a) gradient boost</td>
<td>0.75</td>
<td>0.55</td>
</tr>
<tr>
<td>AG k-nearest neighbors</td>
<td>0.61</td>
<td>0.35</td>
</tr>
<tr>
<td>AG neural network</td>
<td>0.72</td>
<td>0.52</td>
</tr>
<tr>
<td>AG random forest</td>
<td>0.74</td>
<td>0.53</td>
</tr>
<tr>
<td>AG support vector machine</td>
<td>0.55</td>
<td>0.06</td>
</tr>
<tr>
<td>FB(^b) gradient boost</td>
<td>0.67</td>
<td>0.43</td>
</tr>
<tr>
<td>FB k-nearest neighbors</td>
<td>0.68</td>
<td>0.47</td>
</tr>
<tr>
<td>FB neural network</td>
<td>0.67</td>
<td>0.47</td>
</tr>
<tr>
<td>FB random forest</td>
<td>0.67</td>
<td>0.41</td>
</tr>
<tr>
<td>FB support vector machine</td>
<td>0.67</td>
<td>0.45</td>
</tr>
<tr>
<td>SWA(^c) gradient boost</td>
<td>0.80</td>
<td>0.67</td>
</tr>
<tr>
<td>SWA k-nearest neighbors</td>
<td>0.74</td>
<td>0.57</td>
</tr>
<tr>
<td>SWA neural network</td>
<td>0.79</td>
<td>0.66</td>
</tr>
<tr>
<td>SWA random forest</td>
<td>0.80</td>
<td>0.66</td>
</tr>
<tr>
<td>SWA support vector machine</td>
<td>0.68</td>
<td>0.43</td>
</tr>
<tr>
<td><strong>Study 2</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AG gradient boost</td>
<td>0.79</td>
<td>0.56</td>
</tr>
<tr>
<td>AG k-nearest neighbors</td>
<td>0.72</td>
<td>0.48</td>
</tr>
<tr>
<td>AG neural network</td>
<td>0.75</td>
<td>0.51</td>
</tr>
<tr>
<td>AG random forest</td>
<td>0.79</td>
<td>0.57</td>
</tr>
<tr>
<td>AG support vector machine</td>
<td>0.65</td>
<td>0.07</td>
</tr>
<tr>
<td>FB gradient boost</td>
<td>0.73</td>
<td>0.48</td>
</tr>
<tr>
<td>FB k-nearest neighbors</td>
<td>0.72</td>
<td>0.47</td>
</tr>
<tr>
<td>FB neural network</td>
<td>0.71</td>
<td>0.44</td>
</tr>
<tr>
<td>FB random forest</td>
<td>0.73</td>
<td>0.48</td>
</tr>
<tr>
<td>FB support vector machine</td>
<td>0.73</td>
<td>0.48</td>
</tr>
<tr>
<td>SWA gradient boost</td>
<td>0.78</td>
<td>0.57</td>
</tr>
<tr>
<td>SWA k-nearest neighbors</td>
<td>0.76</td>
<td>0.55</td>
</tr>
<tr>
<td>SWA neural network</td>
<td>0.76</td>
<td>0.55</td>
</tr>
<tr>
<td>SWA random forest</td>
<td>0.79</td>
<td>0.58</td>
</tr>
<tr>
<td>SWA support vector machine</td>
<td>0.78</td>
<td>0.55</td>
</tr>
</tbody>
</table>

\(^a\)AG: ActiGraph.  
\(^b\)FB: Fitbit.  
\(^c\)SWA: SenseWear.

Discussion

Principal Findings

This study aggregated two laboratory data sets to build on previous work demonstrating the potential for machine learning algorithms to produce accurate estimates of METs and intensity classes in a diverse set of activities and participants. In both regression and classification settings, we observed the smallest errors in energy expenditure predictions when applying tree-based algorithms (ie, random forest and gradient boosting) to SenseWear and ActiGraph outputs with the RMSE and classification errors generally being higher for Fitbit models. In almost all cases, the error was smaller than the SenseWear manufacturer estimates, and in out-of-sample generalizability experiments, we observed greater error and lower accuracy when compared with the LOSO validations. We believe that

https://mhealth.jmir.org/2021/8/e23938

JMIR Mhealth Uhealth 2021 | vol. 9 | iss. 8 | e23938 | p.117
(page number not for citation purposes)
this is the first study to classify the intensity of activity using machine learning algorithms in Fitbit devices. In Fitbit models, we demonstrated accuracies up to approximately 78% (κ=0.6), with superior performance observed for sedentary activity and MVPA classifications, but these were generally less accurate than ActiGraph and SenseWear models, where up to approximately 85% accuracy (κ=0.74) was achieved. Taken together, and if these results are verified in free-living, ecologically valid examples, these findings imply that highly accurate estimates of energy expenditure, sedentary activity, and MVPA behaviors can be estimated by the wearables tested here.

Algorithm Accuracy

We used neural networks, random forests, and gradient boosting in regression tasks. In previous studies, neural networks and random forests have been shown to be effective in modeling energy expenditure [8,9], and our results confirm this to an extent. The RMSE values observed in the trained models ranged from 0.91 METs to 1.45 METs, which improve upon the SenseWear manufacturer value of approximately 1.86 METs. However, when the average METs in this study were considered (approximately 4 METs), it was evident that the energy expenditure prediction could be further improved. It is noteworthy that neural networks resulted in the highest RMSE for all 3 devices and performed particularly poorly for Fitbit models. Similarly, Kate et al [38] showed that neural networks resulted in bias significantly different from 0, compared with bagged decision trees and numerous other algorithms, which were not statistically different. Despite the utility of deep neural networks to model highly nonlinear functions, in some use cases, the *no free lunch* theorems broadly state that there will not be an optimal algorithm for all tasks [17]. Indeed, for our data sets, tree-based ensemble models are generally superior for both learning tasks. It may be that a higher RMSE can be reduced by larger training sets [39].

We generated lagged accelerometer and heart rate variables for each model because the rate of energy expenditure depends on the rate of work in preceding minutes [27], and the relative importance of these metrics is evidenced in the variable importance analyses. Including time-lagged features allows for a clearer distinction between minutes that are relatively similar in their accelerometer pattern but differ in their measured energy expenditure, that is, sitting for a prolonged period versus sitting immediately after running. Transitional minutes were on average approximately 3 METs (largely attributable to the activity in the preceding minutes), compared with sedentary minutes, which averaged approximately 1.3 METs, yet the error statistics were generally comparable with those observed in sedentary minutes, indicating that algorithms could distinguish between those minutes. More advanced neural network architectures (ie, recurrent neural networks) [40] may further the ability of models to capture the temporal dependencies of energy expenditure.

Generalization

Although many studies have reported low errors when using machine learning approaches in the estimation of energy expenditure or classification of activity, external (out-of-sample) validations are rarer and the opportunity to identify cases of overfitting has been limited. Therefore, we used an out-of-sample validation between the two data sets. In all cases, we observed performance degradation when compared with the LOSO validations. Some of this reduction in accuracy is probably attributable to differences in protocols, activities, and participants, which means that algorithms do not have similar minutes on which to train. In addition, it is possible that the algorithms overfit the data. Overfitting occurs when a complex model learns the *noise* in the training data, which does not represent the true underlying function between the inputs and the output [41]. Previous studies have used out-of-sample validation or validation in free-living environments [10,42,43], and when compared with laboratory validations, errors may increase. Concerning the classification of physical activity intensity in multiple samples, a previous study reported reductions in out-of-sample accuracy relative to the within-sample validated models, in some algorithm and data set comparisons [44]. However, the machine learning models still outperformed the Euclidean norm minus one GGIR classification method in out-of-sample testing. In another comprehensive generalizability study, five lab-based heterogeneous data sets were used to predict exercise intensity. This study found that when models were applied to a different data set than those they were generated on, model accuracy decreased from 72-95% to 41-60% [18]. These drops are notably higher than those in this study, and this is probably attributable to the greater differences in the accelerometer models, wear position, and samples across the five data sets. However, caution must be exercised in a comparison between studies, as the balance of classes is likely to differ and therefore influence some evaluation metrics.

Classification

Our LOSO validations demonstrated a relatively high predictive accuracy (75-85%). However, research-grade device models (ActiGraph and SenseWear) were superior. Fitbit devices provide estimates of time in each category (ie, sedentary, light, and MVPA), but the criteria and algorithms remain proprietary. Feehan et al [45] compared estimates of time in intensities with devices such as ActiGraph and Actical, and concluded that more than 80% of studies reported errors >10% with mean differences ranging between 44% and 632% for estimations of activity above light intensity. Importantly, the devices used for comparison in many studies have varying cut points and are not necessarily gold standards. Our results indicate that the application of machine learning to intensity classification can refine the large errors observed in previous studies. Despite the promising results, we emphasize that laboratory studies have limited ecological validity, and future research should seek to address this. Whole-room indirect calorimetry would likely allow more realistic behaviors to be studied while providing a gold standard comparator.

Strengths and Limitations

A strength of this study is the aggregation of two data sets to provide a more comprehensive and variable data set on which to train models, although the measures (sensors and indirect calorimetry) were the same between studies. The tested cohorts differed demographically, and the protocols were heterogeneous,
which provides a good estimate of the applicability of the tested models. Combining data sets also leads to a larger number of participants (n=89), which is a larger sample size than much of the previous literature [7,9,10,44,46,47]. In general, an increase in training observations is considered a mechanism for enhancing performance [41], and the results of this study provide some evidence that this is the case in both commercial and research-grade accelerometers.

Another strength of this study is the testing of numerous algorithm and device combinations. A previous study developed a multilayer neural network that was trained on a wearable system including a vest for electrocardiogram measurements and 4 accelerometers (one on each wrist and thigh) [47]. Despite the small bias, this is unlikely to be a feasible means of assessing free-living energy balance behaviors. Participant discomfort and sensor removal present additional biases (ie, missing data), which may require additional modeling approaches to address [48-50]. The threshold of practicality varies depending on the size, duration, computational resources, and specific aims of the research study. Therefore, the development of three models with varying requirements is a central advantage of this study.

Testing both classification and regression algorithms in the same devices enhances the use of the results of this study. One area of future work is to explore combined classification and regression approaches, similar to the branched models of the Actiheart [51] or stacked ensemble approaches. This may be effective in producing refined estimates of total daily energy expenditure in free-living subjects, given that most of a day comprises resting or sedentary minutes and some of our models slightly overestimate sedentary activities, although depending on the classification or regression methods, this could incur additional computational costs when applying this to larger data sets. Future work in our lab will examine the application of such models to free-living environments against a doubly labeled water criterion.

A limitation of this study is the lack of a true testing set. Rather, we attempt to develop an unbiased estimate of the true test error by (1) testing on unseen participants and (2) testing on an unseen data set. In the former, the within-subject data are generally more correlated than the between-subject data, and this method represents the closest approximation of how such a model would perform in practice [8]. In the latter, this is extended so that the training and testing sets comprised different participants and protocols. Beyond these validation approaches, the ultimate test of the results presented here is a free-living validation for energy expenditure and intensity classes. The total daily energy expenditure can be validated using the doubly labeled water method over a 7- to 14-day period [52], and the results presented in this paper are part of a wider project in which we aim to validate model predictions in free-living. Although free-living validations are critical, the resolution required to evaluate activity-specific errors can only be obtained from indirect calorimetry. Regarding activity categories, no gold standard method exists to validate time in sedentary activity, light physical activity, and MVPA outside of a controlled environment, and the generalizability of classification models to free-living studies is somewhat uncertain. The authors have highlighted the limitations of accelerometer data collected within a laboratory [53,54]; the activities performed in a free-living environment are more diverse, which further necessitates the need for more naturalistic (ie, free-living) validation studies or at least validation studies conducted over several days using diverse activity protocols in a residential facility. Next, to replicate predictions made by the present algorithms in free-living subjects, measured RMR may be required, which increases the researcher and participant burden. A suitable alternative in the absence of measured RMR would be prediction equations derived from BMI, age, height, and gender, rather than assuming a resting value of 3.5 ml O₂/kg/min [55,56]. Finally, our use of the measured RMR to calculate METs may contribute to differences between the tested algorithms and the SenseWear manufacturer.

Conclusions
This study builds on previous work from our lab and others, demonstrating that machine learning techniques can be used to learn the complexities of human movement and physiological data in the study of human energy expenditure. Classification and regression errors were greater when comparisons were made between studies. Single-sample, cross-sectional studies generating energy expenditure models show acceptable accuracy; however, it is likely that these models are overfitted to a given sample, and thus, improving generalizability is essential. To extend the utility of energy expenditure estimates beyond lab conditions, more cross testing between data sets is required, in addition to validation in free-living samples by doubly labeled water.

Acknowledgments
This research was funded by an Engineering and Physical Sciences Research Council impact acceleration grant (EP/R511717/1).

Authors’ Contributions
ROD, JT, MH, GF, CD, and RJS designed the study. ROD and JT collected the data. ROD, JT, and GWH analyzed the data. ROD, JT, MH, CD, GWH, GF, and RJS contributed to writing and reviewing the manuscript.

Conflicts of Interest
RJS consults for Slimming world UK through Consulting Leeds, which is a wholly owned subsidiary of the University of Leeds. The other authors declare no conflicts of interest.
References


47. Lu K, Yang L, Seoane F, Abtahi F, Forsman M, Lindecrantz K. Fusion of heart rate, respiration and motion measurements from a wearable sensor system to enhance energy expenditure estimation. Sensors 2018 Sep 14;18(9) [FREE Full text] [doi: 10.3390/s18093092]


**Abbreviations**

- **KNN**: k-nearest neighbor
- **LOSO**: leave-one-subject-out cross-validation
- **MAPE**: mean absolute percentage error
- **MET**: metabolic equivalent
- **MVPA**: moderate-to-vigorous physical activity
- **RMR**: resting metabolic rate
- **RMSE**: root mean square error