Original Article

Evaluation framework for selecting wearable activity monitors for research

Kay Connelly[1], Haley Molchan[1], Rashmi Bidanta[1], Sudhanshu Siddh[1], Byron Lowens[2], Kelly Caine[2], George Demiris[3], Katie Siek[1], Blaine Reeder[4,5]

[1]Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, Indiana; [2]School of Computing, Clemson University, Clemson, South Carolina, USA; [3]School of Nursing, University of Pennsylvania, Philadelphia, Pennsylvania, USA; [4]Sinclair School of Nursing, University of Missouri, Columbia, Missouri, USA; [5]University of Missouri Institute for Data Science and Informatics, University of Missouri, Columbia, Missouri, USA

Contributions: (I) Conception and design: K Connelly, B Reeder, G Demiris, K Siek, K Caine; (II) Administrative support: H Molchan, K Connelly, B Reeder; (III) Provision of study materials or patients: K Connelly, B Reeder, G Demiris, K Siek, K Caine; (IV) Collection and assembly of data: K Connelly, R Bidanta, H Molchan, S Siddh; (V) Data analysis and interpretation: All authors; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Blaine Reeder, PhD. Sinclair School of Nursing, MU Institute for Data Science and Informatics, S305 School of Nursing, Columbia, MO 65211-6000, USA. Email: blaine.reeder@missouri.edu.

Background: Wearable devices that support activity tracking and other measurements hold great potential to increase awareness of health behaviors and support the management of chronic health conditions. There is a scarcity of guidance for researchers of all disciplines when planning new studies to evaluate and select technologies appropriate for study purpose, population, and overall context. The aim of this study was to develop and test an evaluation framework to rapidly and systematically evaluate and select consumer-grade wearable devices that serve individual study needs in preparation for evaluations with target populations.

Methods: The wearable evaluation framework was defined based on published literature and past research experiences of the research team. We tested the framework with example case studies to select devices for two different research projects focused on aging-in-place and gestational diabetes. We show how knowledge of target population and research goals help prioritize application of the criteria to inform device selection and how project requirements inform sequence of criteria application.

Results: The framework for wearable device evaluation includes 27 distinct evaluation criteria: 12 for everyday use by users, 6 on device functionality, and 9 on infrastructure for developing the research infrastructure required to obtain the data. We evaluated 10 devices from four vendors. After prioritizing the framework criteria based on the two example case studies, we selected the Withings Steele HR, Garmin Vivosmart HR+ and Garmin Forerunner 35 for further evaluation through user studies with the target populations.

Conclusions: The aim of this paper was to develop and test a framework for researchers to rapidly evaluate suitability of consumer grade wearable devices for specific research projects. The use of this evaluation framework is not intended to identify a definitive single best device, but to systematically narrow the field of potential device candidates for testing with target study populations. Future work will include application of the framework within different research projects for further refinement.

Keywords: Wearable electronic devices; mobile devices; research; chronic disease; fitness trackers

Received: 14 December 2019; Accepted: 21 May 2020; Published: 20 January 2021.
doi: 10.21037/mhealth-19-253
View this article at: http://dx.doi.org/10.21037/mhealth-19-253

Introduction

Personal technologies—such as wearable devices—that support activity tracking and other measurements hold great potential to increase awareness of health behaviors and support the management of chronic health conditions (1,2). Deploying wearable devices outside of controlled laboratory settings into everyday living, however, is fraught with unpredictable external factors that make study implementation difficult (3). Researchers designing and implementing wearable technology studies must select a wearable device that: (I) captures specific indicators of health status; (II) is usable by the population targeted for enrollment; and (III) serves other needs of the study, such as appropriate data granularity for analysis (4,5). Research-grade wearable devices, such as the Actigraph and activPal, are well-understood and have usable form factors, user communities, and copious published literature to guide device selection for their use in research (6-11). However, with the unprecedented availability of consumer-grade wearable devices in a rapidly changing technology landscape, there is little understanding of how consumer-grade wearable devices function under similar conditions (12). Indeed, there is a scarcity of guidance for researchers of all disciplines when planning new studies to evaluate and select technologies appropriate for the study purpose, population, and overall context.

This paper introduces an evaluation framework for researchers to systematically and quickly evaluate and select commercial wearable devices that serve the needs of their projects. Similar to usability inspection methods, the goal of the framework is to enable experts to narrow down a large number of potential devices to a small number of devices for further evaluation with the target population (see *Figure 1*, where the portion of the selection process circled in blue is covered by this paper). After defining the framework, we present two case studies of how we used the framework to help select devices for two very different research projects. We conclude with a discussion of the next steps for researchers once they apply the framework to make a final selection of a wearable device for their project.

Methods

We developed a research evaluation framework for wearable activity monitors which includes 27 distinct evaluation criteria: 12 for everyday use by users, 6 on device functionality, and 9 on the research infrastructure required to obtain the data. These criteria were drawn from previous research, including our own, and refined through internal team discussion. We then tested the framework to select wearable devices using cases from two different projects: one promoting aging in place and one targeting gestational diabetes. Below we describe the evaluation criteria in the areas of everyday use, functionality, and infrastructure support with a description of how the framework can be applied based on project needs and resources.

Everyday use criteria

Four of the 12 everyday use criteria were adopted from design guidelines for wearable devices by Motti and Caine (13) and denoted with a word change for standalone recognition in the item list. The four items are: ease of use for device controls (2), device wearability (6), device aesthetics (11), device customization (12). These items focus specifically on the usability and wearability of the devices in daily living conditions. The everyday use criteria are classified as having both pragmatic and hedonic qualities (13). Pragmatic qualities refer to the functionality of the device and its capability to support the accomplishment of a set of goals. These tend to be objective. Hedonic qualities refer to the product's perceived capability to support the achievement of user needs, and are subjective. These criteria are essential human factors considerations for ensuring the target user population will be able to use the devices over the course of a research project. We itemize the 12 everyday use criteria in *Table 1*.

While some criteria (items 7, 8, 12) are objective and independent of the population (e.g., the device specifications indicate if it is water resistant or waterproof), others are subjective (items 1–6, 9–11) and may depend on the population. For example, a device's "Display Viewability" may differ between populations with differing visual acuity. Each item is ranked on a 5-point scale (with 1 being "not at all/never" and 5 being "all the time/very"), with the exception of item 12, and includes qualitative descriptions where appropriate. Items 1–5 relate to ease of use and understanding the information, items 6–10 deal with wearing the device every day, and items 11–12 are related to the aesthetics of the device.

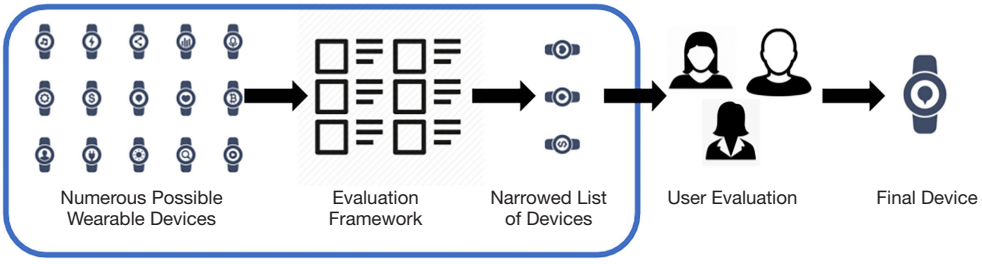

Figure 1 Process for selecting wearable device for research study. This paper describes the evaluation framework, used in the early stages of devices selection to move from many devices to a few devices.

Table 1 Everyday use criteria

Criteria	Definition	Description
1. Ease of setup[c]	Deals with the amount of effort invested to start using the device	Includes items such as pairing with mobile device, account setup, and finding the app from the app store
2. Ease of use for device controls[a]*	Deals with effort needed to use the physical device controls	Includes the comfort of using the controls, ease of accessing different screens using control buttons, and ease of navigating on the wearable device
3. Wearable display viewability[a]	Deals with the overall viewability of the wearable display	Includes simplicity of the display, size of the text or infographics, clarity with which data is displayed on the wearable device, and visibility of display under different light conditions
4. Wearable display interpretability[a]	Deals with cognitive load of interpreting the wearable display	Includes ease of accessing the different functions and data on the wearable device, as well as the granularity of the data displayed
5. Ease of use for mobile app[b]	Deals with cognitive load of accessing and interpreting the data on the paired mobile app	Includes ease of accessing the different data on the paired mobile app, as well as the granularity of the data displayed
6. Device wearability[a]*	Deals with parameters related to the user experience wearing the device/placing the wearable device upon the user's body	Includes the size of the device, size of display, and comfort while wearing the device
7. Device water resistance[a]	Measures the water resistance of the wearable device, rated based on the scale shown in description	1: Not waterproof or resistant 3: Water-resistant/splash proof 5: Waterproof/submersible
8. Wearable device battery[a]	Deals with the battery life of wearable device (T) and is rated using a scale of 1–5 available in description based on evaluator experience with device	1: T ≤2 days 2: 2 days < T ≤1 week 3: 1 week < T <1 month 4: 1 month < T <6 months 5: 6 months < T
9. Device effect on mobile battery[b]	Deals with any effect on the mobile phone battery life	Includes information derived from typical use case scenarios for the target population, such as using GPS or not
10. Syncing performance[c]	Deals with the syncing performance of the device	Includes the max duration without syncing, number of day's data the device can store, effect of not syncing on the device, ease of syncing device
11. Device aesthetics[a]*	Deals with overall look and feel of the device, such as modern, or classic, and bulky or slim	Includes subjective assessment of overall aesthetics, in which preferences will likely vary between individual users
12. Device customization[a]*	Deals with the different customization options available	Includes available choices for color options, belt clip and strap options. Binary rating

*, items 2, 6, 11 and 12 were adopted from Motti & Caine (2014); [a], applies to wearable device only; [b], applies to mobile phone only; [c], applies to the interaction of wearable device and mobile phone.

Table 2 Functionality criteria

Criteria	Definition	Description
1. Parameter measures	Deals with overall functionality, availability and relative accuracy	Includes steps, sleep, elevation, intensity, activity recognition, heart rate, oxygen level, and GPS
2. Motivational features	Deals with availability of motivational features to encourage usage	Includes vibration, sounds or app notifications to encourage meeting goals, such as step count or sleep duration
3. Notifications	Deals with availability of notification support	Includes mobile notifications to wearable device and do not disturb mode
4. Clock	Deals with availability of clock function on wearable device	Includes glanceable *vs.* manually initiated (e.g., button press), and analog *vs.* digital
5. Availability of personal data inputs/ reminders	Deals with availability of adding personal data inputs and setting reminders	Includes if reminder can be set on wearable and/or mobile device
6. Connectivity to other apps	Deals with option to pair device or app with other 3rd party applications	Includes popular fitness apps and social media

Functionality criteria

Functionality criteria were informed by technical specification sheets provided for each device (www. garmin. com, withings. com, www.fitbit.com, striiv.com) since these devices have wide adoption, high consumer reviews, and incorporate common features in the device marketplace. Many of these items have a binary rating in that the device either has the specified software feature or does not. We itemize these six functionality criteria with descriptions in *Table 2*.

When choosing a device, it is important to ensure it collects the data necessary and has the required functions for the proposed study. For example, an observational study may not want any motivational features or notifications since those may alter participant behavior; whereas an intervention study trying to increase the physical activity of participants with congestive heart failure may require these features (1).

Infrastructure criteria

Table 3 lists the criteria for evaluating the infrastructure support provided by device vendors. Most vendors provide a free application programming interface (API) for software developers to access user data, and some provide an additional API to access more features (e.g., computation of gait) for a charge. In addition, vendors may provide developer and runtime support, which is critical for projects that need to automatically download user data. The outcome of the infrastructure evaluation is primarily based on vendor documentation and requires a technical member of the research team to perform the evaluation.

Contextualizing framework application by project

While the evaluation can be initiated independent of a specific research project, the target population and research goals should be known or estimated to most effectively prioritize the various criteria and make a final selection. For example, a project targeting a small number of participants that are at risk of developing a chronic health condition, such as congestive heart failure, may place a high priority on physiological measures like heart rate (1,12). Whereas a project targeting a large number (1,000+) of participants newly diagnosed with type 1 diabetes, may place a higher priority on the inclusion of notifications or aesthetics of the device because these factors may be a priority to young adults (14). Similarly, devices meant for users with functional limitations such as blindness or a loss of dexterity would require different usability parameters than those meant for the general population.

Some criteria lend themselves more readily to straightforward quantitative measurements for comparison of evaluation metrics across devices and evaluators, while other items in the framework require a qualitative approach that relies on independent rating by multiple evaluators followed by a reconciliation process that establishes inter-rater reliability (15). A quantitative measurement example is battery life which can be determined measuring the number of days or a calculation of the battery (mAh), components

Table 3 Infrastructure criteria

Criteria	Definition
1. Ability to access device data	Describes the availability of vendor API for accessing data in a secured and non-intrusive manner
2. API: cost to access data	Describes the cost/fees or any special permission associated with using the Vendor APIs
3. API: data type availability	Describes the type and the granularity of the data available through the APIs
4. API: scalability and rate limiting	Describes the limitations on the number of API calls that can be made
5. API: data availability notification	Describes the ability to receive real-time data pushed from the vendor, as opposed to periodically pulling data from the vendor
6. API: maturity of access path	Describes the historic stability of major versions and the frequency of API changes. Newer technology and start-ups more likely to push frequent changes that can impact study infrastructure that uses APIs
7. API resources	Describes the availability and quality of API reference material for developers
8. Developer support	Describes the availability and quality of dedicated forums and teams to offer support for the developers. Some vendors have active and helpful developer groups, others have employees to respond to developer questions, and others have a combination of the two
9. API: system health check	Describes the provision of broadcasting information to partners of any possible downtime

API, application programming interface.

(e.g., GPS, 3-axis accelerometer), and chipset used in a controlled test setting that allows comparison of function for multiple devices.

A qualitative measurement example, such as ease of physical controls or aesthetics, can vary widely based on user physical abilities (e.g., dexterity), and technology preferences (e.g., affinity for a traditional analog watch versus a digital device). For these types of qualitative measurements, we provide a priori codes that form the basis for qualitative evaluation items. For this approach, a minimum of two evaluators, with an understanding of the targeted population of users, should familiarize themselves with the evaluation framework and discuss the application of the criteria. Then, each should independently apply the framework criteria in a sample evaluation of one device, followed by a discussion of evaluation results, to establish an understanding of the application of the framework criteria within the evaluation team and reconcile disagreements. How framework criteria are applied within a given project should be documented in a codebook that can be referred to during successive device evaluations. To increase rigor, a third evaluator can be used to reconcile disagreements. Additional device evaluations are conducted iteratively, with successive reconciliations informing and refining the codebook for application of framework criteria. This evaluation process is informed by approaches from ubiquitous computing to "debug the viability of the systems in everyday use, using ourselves and a few colleagues as guinea pigs" (16) and usability inspection methods, such as heuristic evaluations (17), that rely on small evaluation teams to uncover usability issues (17).

After narrowing device candidates based on evaluations with team members, the next step is to perform testing with the targeted population of users to verify feasibility and usability of the device before moving to field studies that test the device in everyday living settings. This progression of testing includes needs assessment, design validation, usability testing, laboratory function testing, laboratory user effect through field tests and ultimately assessment of broader population impacts (18). Furthermore, this evaluation strategy follows recommendations to iteratively test using a series of small evaluation studies that progress to community-based studies (19).

Test application of framework using two cases

To test the evaluation framework and demonstrate how it may be used, we present an evaluation of ten wearable activity monitors available in the United States in early 2017. We use the framework with case studies of the following two projects, each of which had IRB approval in place, separate from this investigation:

Table 4 Ten devices evaluated using the framework. Price from summer 2017

Device	Price	Heart rate	GPS
Garmin Vivofit 2	$69.95	–	–
Withings Activite	$129.95	–	–
Withings Go	$49.95	–	–
Striiv Fusion	$79.99	–	–
Withings Pulse O2	$99.95	√	–
Fitbit Charge HR	$149.95	√	–
Withings Steel HR	$179.95	√	–
Garmin Vivosmart HR	$129.95	√	–
Garmin Vivosmart HR+	$179.95	√	√
Garmin Forerunner 35	$199.99	√	√

HomeSHARE

The HomeSHARE project (20) provides wearable activity monitors to approximately 30 older adults (>65 years old) to be used for at least 2 years. This project was primarily observational, to better understand older adult activities and independent living. It created a data set that combined data from wearables, home-based sensors, surveys, and interviews.

Precision Health Initiative

The Precision Health Initiative project (in progress, results unpublished) provides ~500 pregnant women a wearable activity monitor to wear for at least 12 months. This project is prospective in nature, seeking to build a model of women who had gestational diabetes that later develop type 2 diabetes. As such, it is primarily observational, creating a data set that combined data from wearables, electronic health records, genetic mapping, surveys, and interviews.

Both projects require that the wearable device has some measure of physical activity and sleep. In addition, the investigators were interested in evaluating whether monitoring Heart Rate and GPS were feasible given their respective budgets. We evaluated ten commercially available activity trackers based on cost (≤$200 per device) and available features. The ten devices, listed in *Table 4*, collected sleep and activity information, and a subset of devices collected heart rate and/or GPS.

Two researchers wore each of the ten devices for 1 week, removing the device only as needed for charging or if submerging in water for non-waterproof devices. A total of eight researchers from the same research lab participated in the evaluation, assisting with two devices each. Evaluators independently filled out a form for criteria #1–18 for each device they wore (the final version of the form available online: http://cdn.amegroups.cn/static/application/80aaeab616cf1f7a5419149b997d5368/mhealth-19-253-1.pdf). On average, filling out the evaluation form took 42 minutes to complete for each device.

Because not all of the framework criteria are objective, prior to using the average ratings from the two evaluators, we analyzed the ratings and comments for inconsistencies. Any ratings that differed by a point or more, or comments that were contradictory between evaluators, were flagged for further discussion. These evaluators were gathered together, and each noted discrepancy was discussed and resolved. Most discrepancies were due to confusion about the definitions of the evaluation items, which led us to refine the definitions and improve the framework. Some discrepancies were due to different use cases encountered by the different evaluators for their regular routines. For example, one evaluator used a device more in an outdoor setting, helping them to notice a viewability issue in sunlight that another evaluator did not notice. In such cases, the evaluators discussed the differing experiences and adjusted their scores accordingly.

Results

Results for everyday use criteria

Table 5 presents the average results for the criteria which could be mapped to ratings. Devices with heart rate or GPS

Table 5 Everyday use ratings for wearable devices

Criteria	Withings				Garmin				Striiv	Fitbit
	GO	Activite	Pulse O2[1]	Steel HR[1]	Vivofit 2	Vivosmart HR[1]	Vivosmart HR+[1,2]	Forerunner 35[1,2]	Fusion 2	Charge HR[1]
Ease of setup	3	3.75	4	4.5	4	5	4.5	4.5	2.75	4.5
Ease of use for device controls	1	4	2	4.75	4	5	5	4.5	2.75	4.5
Wearable display viewability	4.5	3.75	2	4.5	5	5	5	5	2.25	4.5
Wearable display interpretability	4.5	5	4	4.75	4.5	5	5	4.5	2	3.5
Ease of use for mobile app	4	3	4	4.5	4	4.5	5	4.5	2.5	5
Device wearability	2	4.5	3	4.5	4	4.75	4.75	5	3.5	4
Device water resistance	3.5	5	1	5	5	5	5	5	3	5
Wearable device battery	5	5	3	4	5	2	2.25	3	2	3
Device effect on mobile battery	4	4.5	4	4.5	4	5	4	4.5	4	3.5
Syncing performance	4	4	4	4	4	5	5	5	2.5	4.5
Device aesthetics	3	5	4	4.5	4	5	4.5	4	4	3.5

[1], denotes device with Heart rate function; [2], denotes device with GPS function.

are identified by a superscript 1 or 2, respectively.

For the non-HR devices, the Withings Activite and Garmin Vivofit 2 easily outperformed the Withings Go and Striiv Fusion. The latter both scored low on physical controls, aesthetics, and wearability because they were found to have unmanageable buttons/screens, problematic band clasps and a clunky design. The Fusion scored low on water resistance because it offered minimal splash resistance. Thus, the Go and Fusion were eliminated from further contention.

For the HR devices, the Withings Pulse O2 and Fitbit Charge HR underperformed compared to the other four evaluated devices, receiving the poorest scores for water resistance, display viewability, ease of physical controls, and device battery (Fitbit Charge HR). Because of their low scores in these categories, the O2 and Charge HR was eliminated from further contention. The Withings Steele HR was considerably better than the others for battery life; otherwise, it compared comparably to Garmin Vivosmart HR, Garmin Vivosmart HR+ and Garmin Forerunner 35. The latter two are the only two to include GPS as part of their functionality. For the Ease of Setup criteria, the Vivosmart HR was the only device to receive a perfect score. The other devices received lower ratings because they suffered from multiple failed Bluetooth pairing attempts and/or difficulty finding the correct device among the many similarly named devices. These issues were documented in the notes of the evaluator worksheets.

Table 6 presents the binary items that could not be mapped to ratings. For these criteria, evaluators checked if devices supported these features or not. In the everyday use category, there is only one criteria (customization) rated this way.

Results for functionality criteria

Table 7 presents the results for the binary items that were evaluated in the functionality category. *Table* 7 includes the example features/functions that the research team chose to evaluate the devices for, however depending on the study, these examples can be decided upon based on the research goals and target population. Based on the research goals of a given project, the inclusion of these criteria could also be used to narrow the device pool.

Results for infrastructure criteria

We provide infrastructure inspection results for Garmin, Withings (Nokia) and Fitbit in qualitative terms. We did not consider infrastructure for the Striiv device, as it was ruled out from the further study based on the results of applying the everyday use criteria. After reviewing API features,

Table 6 Everyday use qualitative criteria

Customization	Withings				Garmin				Striiv	Fitbit
	GO	Activite	Pulse O2	Steel HR	Vivofit 2	Vivosmart HR	Vivosmart HR+	Forerunner 35	Fusion 2	Charge HR
Colored bands	√	√	√	√	√	√	√	√	√	√
Belt clips	√	–	√	–	√	–	–	–	–	–
Different band materials	–	√	–	√	–	–	–	–	–	–

Table 7 Functionality qualitative ratings for wearable devices

Category 2: functionality	Withings				Garmin				Striiv	Fitbit
	GO	Activite	Pulse O2	Steel HR	Vivofit 2	Vivosmart HR	Vivosmart HR+	Forerunner 35	Fusion 2	Charge HR
Physiological measures										
Steps, sleep, HR, calories	–	–	√	√	–	√	√	√	–	√
Motivation										
App badges	√	√	√	√	√	√	√	√	√	√
Device motivation*	√	–	√	–	–	–	–	–	–	–
Notifications (text, calls, etc.)	–	–	√	√	–	√	√	√	√	√
Clock	√	√	√	√	√	√	√	√	√	√
Manual inputs/reminders										
Weight	√	√	√	√	√	√	√	√	√	√
Food intake	√	√	√	√	√	√	√	√	√	√
Connectivity to other apps										
myFitnessPal	√	√	√	√	√	√	√	√	–	√
Strava	–	–	–	–	√	√	√	√	–	√

*, text encouragement and "move" bar that encourages walking after long periods of inactivity.

members of the development teams for HomeSHARE and the Precision Health Initiative agreed that all three vendors provided support, to varying degrees, in each of the nine areas in *Table 3*. We provide a summary of this subjective review below.

All three vendors provide basic access to data through their free APIs. All used RESTful APIs, a popular architecture for developing client-server applications. REST stands for Representational State Transfer and provides interoperability between software systems. Any system that complies with REST architectural style can provide services that can be accessed through world wide web (www) URLs (21). This architecture provides developers with an ease way to access API features by consuming URLs in the programming language of their choice. Software applications, including mobile as web-based applications, can use the APIs to retrieve health, fitness, and wellness data from vendor data repositories. No vendors provided access to raw sensor data. Garmin provides access to their "Garmin Health API" by request, after submission and approval of a required application. Withings/Nokia and Fitbit provide access to intraday minute and second-level data, upon request with justification of how the data will be used.

With regard to scalability, all three vendors enforce rate limits for frequency of data access. This limitation should be considered when developing third party applications based on target number of users and vendor policies for data

access. Each vendor employs a Subscribe-to-Notification mechanism for third party apps to receive notifications whenever a wearable device is synced to its corresponding native application. For example, a Withings device syncs with the Nokia HealthMate mobile app over Bluetooth and a Garmin device syncs with the Garmin Connect mobile app. When a device syncs to a native mobile app, the third-party app is notified of new data and the third-party app must retrieve data from the vendor system via the API within the constraints of the vendor data access policy.

At the time of evaluation, Garmin and Fitbit had a more mature and stable APIs than Withings. During the 2 years prior to our evaluation period, they both had regular major releases to their APIs to incorporate new features, such as Pulse Ox. Withings/Nokia, on the other hand, had virtually no releases in 2017–2018, the timeframe in which they were sold/rebranded two times. API reference documentation was comprehensive for Garmin and Fitbit, whereas Withings documentation was incomplete with regard to supported API features that were known to exist, requiring a team member to obtain the information from other developers. Garmin and Fitbit were responsive with regard to technical support. The Garmin technical support had a turn-around time of within a day for most of technical difficulties. Fitbit has an extremely active developer's forum, with most questions being responded to within a day, often by Fitbit engineers, but also by other developers. Withings support was not responsive to our emails. Data such as heart rate on Withings HR Steel were only accessible through the intraday API. Withings did not respond to multiple requests for access to the intraday API which precluded us from exploring the intraday data capabilities. All vendors provide some support for a system health check with Fitbit providing status of each API service through a dedicated URL, and the others providing maintenance notifications in a developer portal website.

Results for contextualizing framework application by project

Each research project has different requirements for their target population and the infrastructure. Below, we give two examples of such projects in which the authors with significant experience with that population used the evaluation framework to narrow down the devices. We then briefly describe two follow-up user studies which resulted in a final device for each project.

HomeSHARE

Based on our experience in design of technologies for older adults, we prioritized seven criteria from the framework:

- 1-Ease of Setup: as adults age, they may lose some cognitive abilities. Thus, the selected device should be easy to setup and learn.
- 2-Ease of Physical Controls: older adults may have challenges with dexterity and muscle grip. The physical controls on the device should be easy to use, including their size and pressure required to work them. Touchscreens can be particularly challenging for older adults, so should be evaluated with care.
- 3-Display Viewability: older adults may have deteriorating eyesight, which makes it difficult to read small fonts on screens. When choosing a device, screen and text size must be considered.
- 4-Display Interpretability: older adults can have difficulty with displays that are too distracting or complex to navigate. The display should be simple to interpret.
- 6-Wearability: the device should be comfortable for short term and long-term use. The strap/band shouldn't irritate the skin of older adults and should feel natural to them. The device should neither be too big nor too small. The device should be easy to put on/off and should avoid complicated/difficult clasps.
- 8-Device Battery: while many populations struggle with keeping devices charged, older adults are particularly challenged due to not being as familiar with such devices, age-related memory loss, difficulty taking the device on and off, and difficulty connecting the device to its charger. Battery life is critical for this population. Having a display that provides feedback when the battery is getting low will also help older adults remember to charge their device.
- 11-Aesthetics: older adults, much like their younger counterparts, want a fashionable device that looks good. They may find devices that look more like a watch more pleasant than more modern designs. Additionally, older adults in various studies have indicated that devices should not be "stigmatizing" them as frail or in need of assistance with a "look" that indicates a medical (22,23).

Precision Health Initiative

Based on our experience of design for pregnant women and

women with infants and toddlers, we prioritized different criteria:

- 4-Display Interpretability: women with infants and young children have many demands on their attention, making an easily glanceable interface a high priority.
- 6-Wearability: the device should be comfortable for short term and long-term use and should not problematic if young children pull on it.
- 7-Water Resistance: mothers must often do a variety of household and childcare activities which involve water and will not want to have to remove the device frequently throughout their day.
- 8-Device Battery: the battery life should be as long as possible, so women do not have to frequently remember to take the device off and charge it.
- 9-Smartphone Battery: the drain on the phone battery should not require charging during the day.
- 11-Aesthetics: women will want a device that fits in with their style.
- 12-Customization: pregnant women may require different sized straps over the course of the pregnancy. They may also prefer to wear the device in different locations, depending on their activities (e.g., a belt clip if the device does not pick up on steps when pushing a stroller since the hands will not sway naturally).

Selecting devices for further investigation

Using our framework enabled us to quickly and easily narrow down the list of devices for the HomeSHARE and Precision Health Initiative projects. To narrow down and select a device, researchers selected items from the prioritized criteria for each project. From this, researchers highlighted any criteria for each device that scored ≥4 and tallied how many categories each wearable scored this in. The five devices with the lowest tallied scores in the prioritized criteria categories were discounted from the list of possible devices. For the HomeSHARE project, researchers narrowed the list of devices down to five: the Withings Steel HR, Garmin Vivofit 2, Garmin Vivosmart HR, Garmin Vivosmart HR+ and Garmin Forerunner 35. For the Precision Health Initiative project, multiple devices tied for the number of categories they received high scores, so researchers narrowed it down to six devices: the Withings Activite, Withings Steel HR, Garmin Vivofit 2, Garmin Vivosmart HR, Garmin Vivosmart HR+ and Garmin Forerunner 35. Both projects, were interested in collecting HR and potentially GPS data, so the next step for researchers was to cut any devices that didn't have at least one of these functions. Even though the prioritized criteria were different for the two projects, the final list of potential devices were the same: Withings Steel HR, Garmin Vivosmart HR, Garmin Vivosmart HR+ and Garmin Forerunner 35. These devices were then narrowed down to three devices after eliminating the Vivosmart HR. Researchers chose to eliminate this device because of its similarities to HR+ and lack of GPS.

Once a smaller set of promising devices are identified with the framework, we suggest a user evaluation with the target population to make a final selection. Here, we summarize two such studies for our demonstration projects. Both studies were approved by the Indiana University IRB. For the HomeSHARE project, we conducted an in-lab usability study with older adults (n=9) recruited locally using three potential devices (Withings Steel HR, Garmin ForeRunner 35, Garmin Vivosmart HR+). These semi-structured interviews uncovered that most participants did not like the Withings Steel HR due to difficulties in screen readability and the overall weight of the device. Two of the participants preferred the Vivosmart HR+ because it had a slightly smaller form factor and was similar to what they already owned. However, seven of the participants preferred the Garmin Forerunner 35 over the Vivosmart and Withings devices because they felt the display was easy to see and intuitive, and they preferred the shape/size. Based on these results, we selected the Garmin Forerunner 35 for the HomeSHARE study.

For the Precision Health Initiative, we conducted an identical study with a population of pregnant and recently (<2 years) postpartum participants. Recruited locally, participants (n=9) explored the same three wearables over the course of a semi-structured interviews. Above all, participants focused on the aesthetic of the devices. Many found the Forerunner 35 to be "outdated" and clunky when compared to the other devices. The Vivosmart HR+ and Steel HR both had mostly positive comments, however some participants preferred the Steel HR due to its more un-athletic appearance and perceived appropriateness in a workplace environment.

These two studies highlight the importance of direct evaluation with targeted populations, as seen through the different aspect of the devices focused on the differing populations. Through the user studies, we were able to further narrow down preferred devices and identify the

focal features different populations prefer. Based on the results of this user study, we selected the Vivosmart HR+ for the Precision Health Initiative study.

Finally, while an in-lab user evaluation can help select the final device, it cannot fully replace in the wild testing (3). For both the Precision Health Initiative and HomeSHARE studies, we proceeded with *in-situ* pilot studies to ensure the target populations could, and would, use the devices in their everyday lives.

Discussion

The primary goal of this paper is to provide a framework for researchers to quickly evaluate the suitability of current commercial activity monitors for their research projects. The use of this framework is not intended to result in the identification of a single device that should clearly be selected. Instead, it allows researchers to narrow down prospective devices to a shortlist that can be tested further with the target population. The next steps should be to perform tests that evaluate the devices with the target population, including participants' ability to use the devices and preferences. These tests could involve focus groups about technology perceptions and acceptability, lab-based usability tests, or small "in the wild" field studies to understand real world implementation factors. There are some considerations that the evaluation framework does not address including device compatibility (e.g., some devices are only compatible with certain versions of iOS) and device availability. Research grade devices may prioritize longer term availability, where as those on the commercial market may become unavailable in a relative short period of time. Finally, there are two important additional considerations that the evaluation framework does not address, accuracy and privacy. We describe these in greater detail below.

Accuracy

The evaluation framework does not itself include items for accuracy of the data that are collected. While this information is important, it would take an entire field study with an accurate baseline [e.g., comparison to an Actigraph (10)] to provide the accuracy of each device. The proposed framework is meant to support a streamlined evaluation that can occur in the lab as an initial screening and assessment of devices. For some projects, relative accuracy may be all that is required, allowing researchers to note trends for individual participants. If clinical accuracy is an important criterion for a research study, then a separate evaluation must occur. Research requiring clinical accuracy would need to compare prospective commercial wearable devices with a baseline device.

Privacy

Continuous collection of data collected by wearable activity monitors can be useful but comes with issues (24). Data gathered by wearables can reveal sensitive information about the individual user and their surroundings which can raise privacy-related concerns (25). Some specific concerns are user knowledge and consent about personal health data collected from wearable devices. If data are transferred to external entities (e.g., device vendor or third party) outside of the control of the user who is producing the data this may result in a privacy invasion (26). These instances can lead to misuse of personal data to adjust healthcare insurance policies and premiums based on daily behaviors, activities and pre-existing medical conditions (27,28). In some instances, users may synchronize data collected by their wearable device with social media sites which present additional privacy-related threats and risks (28).

Privacy is a key concern for users in their adoption of pervasive computing technologies (13,25,29), and has been cited as a particularly important issue to address in ubiquitous and wearable computing (13,25,29). To ensure privacy protection, it is necessary for users to understand the type of data the wearable devices collects, store, and share (30). In addition, it is important that users have control over their health-related information (31). The variability of useful data generated by wearable devices, along with an even more diverse collection of interfaces through which users interact necessitates a practical approach that ensures privacy protection for users (32). Given the many different factors involved in perceptions and need for privacy, privacy concerns should be evaluated by project and participant requirements and preferences.

Conclusions

We present the development and testing of a wearable evaluation framework for selecting devices for research. The evaluation framework includes 27 distinct evaluation criteria: twelve for everyday use by users, six on device functionality, nine on infrastructure for developing the research infrastructure required to obtain the data. We presented two case studies of how to use the framework to

determine a shortlist of devices that may be appropriate for a particular research project. We chose the initial set of devices for evaluation based on features and price point. Some researchers may consider existing published evaluation studies as a component for the initial device pool, if comparability to an existing population is desired. Others may prioritize a combination of other criteria, such as battery life, privacy, or cost. Given the ever-changing nature of the commercial activity monitor market, the aim of this framework is to enable researchers to quickly and systematically evaluate the state-of-the-art commercial devices for their particular research needs.

Acknowledgments

Funding: This work was supported in part by National Science Foundation Award numbers: 1629468, 1629202, 1625451, 1629437, 1619950, 1405834, 1405873, 1405951 and 1405682 and the Indiana University Precision Health Initiative.

Footnote

Provenance and Peer Review: This article was commissioned by the Guest Editor (Mei R Fu) for the series "Real-Time Detection and Management of Chronic Illnesses" published in *mHealth*. The article was sent for external peer review organized by the Guest Editor and the editorial office.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx. doi. org/10. 21037/mhealth-19-253). The series "Real-Time Detection and Management of Chronic Illnesses" was commissioned by the editorial office without any funding or sponsorship. KC discloses National Science Foundation grant award 1405723 during the conduct of the study. KS reports grants from National Science Foundation, grants from National Science Foundation, grants from Indiana University, during the conduct of the study. The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Phillips SM, Cadmus-Bertram L, Rosenberg D, et al. Wearable Technology and Physical Activity in Chronic Disease: Opportunities and Challenges. Am J Prev Med 2018;54:144-50.
2. Reeder B, Richard A, Crosby ME. Technology-Supported Health Measures and Goal-Tracking for Older Adults in Everyday Living. In: Schmorrow D, Fidopiastis CM. Foundations of Augmented Cognition: 9th International Conference, AC 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings. 2015:796-806
3. Hazlewood WR, Connelly K, Caine KE, et al. Property Damage, Purchasing Orders, and Power Outages, Oh My!: Suggestions for Planning Your Next In-The-Wild Deployment. In: van Hoof J, Demiris G, Wouters E. editors. Handbook of Smart Homes, Health Care and Well-Being. Cham: Springer, 2014:1-14.
4. Rosenberger ME, Buman MP, Haskell WL, et al. Twenty-four Hours of Sleep, Sedentary Behavior, and Physical Activity with Nine Wearable Devices. Med Sci Sports Exerc 2016;48:457-65.
5. Rodgers MM, Cohen ZA, Joseph L, et al. Workshop on personal motion technologies for healthy independent living: executive summary. Arch Phys Med Rehabil 2012;93:935-9.
6. Gorman E, Hanson HM, Yang PH, et al. Accelerometry analysis of physical activity and sedentary behavior in older adults: a systematic review and data analysis. Eur Rev Aging Phys Act 2014;11:35-49.
7. Burton C, McKinstry B, Szentagotai Tatar A, et al. Activity monitoring in patients with depression: a systematic review. J Affect Disord 2013;145:21-8.
8. Kim Y, Beets MW, Welk GJ. Everything you wanted to know about selecting the "right" Actigraph accelerometer cut-points for youth, but…: a systematic review. J Sci Med Sport 2012;15:311-21.
9. van Nassau F, Chau JY, Lakerveld J, et al. Validity and responsiveness of four measures of occupational sitting and

standing. Int J Behav Nutr Phys Act 2015;12:144.
10. Júdice PB, Santos DA, Hamilton MT, et al. Validity of GT3X and Actiheart to estimate sedentary time and breaks using ActivPAL as the reference in free-living conditions. Gait Posture 2015;41:917-22.
11. Bassett DR Jr, John D, Conger SA, et al. Detection of lying down, sitting, standing, and stepping using two activPAL monitors. Med Sci Sports Exerc 2014;46:2025-9.
12. Reeder B, David A. Health at hand: A systematic review of smart watch uses for health and wellness. J Biomed Inform 2016;63:269-76.
13. Motti VG, Caine K. Human Factors Considerations in the Design of Wearable Devices. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2014;58:1820-4.
14. Maahs DM, West NA, Lawrence JM, et al. Epidemiology of type 1 diabetes. Endocrinol Metab Clin North Am 2010;39:481-97.
15. Boyatzis RE. Transforming qualitative information: Thematic analysis and code development. Thousand Oaks, CA, US: Sage Publications, Inc., 1998:184, xvi.
16. Weiser M. Some computer science issues in ubiquitous computing. Communications of the ACM 1993;36:75-84.
17. Nielsen J. Finding usability problems through heuristic evaluation. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Monterey, California, USA. 142834: ACM, 1992:373-80.
18. Clifford GD. Review of "Biomedical Informatics; Computer Applications in Health Care and Biomedicine" by Edward H. Shortliffe and James J. Cimino. Biomed Eng Online 2006;5:61.
19. Friedman CP. "Smallball" evaluation: a prescription for studying community-based information interventions. J Med Libr Assoc 2005;93:S43-8.
20. Reeder B, Molchan H, Gutierrez E, et al. HomeSHARE: Implementing Multi-Site Smart Technology Infrastructure. AMIA Annual Symposium 2019; November 16-20; Washington, DC, 2019.
21. Fielding RT. Architectural styles and the design of network-based software architectures: University of California, Irvine; 2000.
22. Demiris G, Chaudhuri S, Thompson HJ. Older Adults' Experience with a Novel Fall Detection Device. Telemed J E Health 2016;22:726-32.
23. Courtney KL, Demiris G, Hensel BK. Obtrusiveness of information-based assistive technologies as perceived by older adults in residential care facilities: a secondary analysis. Med Inform Internet Med 2007;32:241-9.
24. Hagen L. Overcoming the Privacy Challenges of Wearable Devices: A Study on the Role of Digital Literacy. Proceedings of the 18th Annual International Conference on Digital Government Research; Staten Island, NY, USA. 3085254: ACM; 2017:598-9.
25. Motti VG, Caine K, editors. Users' privacy concerns about wearables. International Conference on Financial Cryptography and Data Security. Springer, 2015.
26. Peppet SR. Regulating the Internet of Things: First Steps Toward Managing Discrimination, Privacy, Security & Consent. Texas Law Review, Forthcoming. 2014. Available online: https://ssrn.com/abstract=2409074
27. Spann S. Wearable fitness devices: personal health data privacy in Washington State. Seattle University Law Review 2015;39:1411.
28. Hallam C, Zanella G. Wearable device data and privacy: A study of perception and behavior. World Journal of Management 2016;7:82-91.
29. Lee L, Lee J, Egelman S, et al. editors. Information disclosure concerns in the age of wearable computing. NDSS Workshop on Usable Security (USEC), 2016.
30. Schaub F, Balebako R, Cranor LF. Designing Effective Privacy Notices and Controls. IEEE Internet Computing. 2017;21:70-7.
31. Caine K, Hanania R. Patients want granular privacy control over health information in electronic medical records. J Am Med Inform Assoc 2013;20:7-15.
32. Wolf B, Polonetsky J, Finch K. A practical privacy paradigm for wearables. 2015. Retrieved January. 2016. Available online: https://fpf.org/wp-content/uploads/FPF-principles-for-wearables-Jan-2015.pdf;

doi: 10.21037/mhealth-19-253
Cite this article as: Connelly K, Molchan H, Bidanta R, Siddh S, Lowens B, Caine K, Demiris G, Siek K, Reeder B. Evaluation framework for selecting wearable activity monitors for research. mHealth 2021;7:6.