EVIDENCE Publication Checklist for Studies Evaluating Connected Sensor Technologies: Explanation and Elaboration

Christine Manta^{a, b} Nikhil Mahadevan^{a, c} Jessie Bakker^{a, d} Simal Ozen Irmak^e
Elena Izmailova^{a, f} Siyeon Park^g Jiat-Ling Poon^h Santosh Shevadeⁱ
Sarah Valentine^h Benjamin Vandendriessche^{j, k} Courtney Webster^l
Jennifer C. Goldsack^a

^aDigital Medicine Society, Boston, MA, USA; ^bElektra Labs, Boston, MA, USA; ^cPfizer Inc., Cambridge, MA, USA;
^dPhilips, Monroeville, PA, USA; ^eTibi Health Inc., San Francisco, CA, USA; ^fKoneksa Health Inc., New York, NY, USA;
^gGeisinger Health System, Danville, PA, USA; ^hEli Lilly and Company, Indianapolis, IN, USA; ⁱIndependent Consultant,
Mumbai, India; ^jByteflies, Antwerp, Belgium; ^kDepartment of Electrical, Computer and Systems Engineering, Case
Western Reserve University, Cleveland, OH, USA; ^lNymbly.work, Seattle, WA, USA

Keywords

Connected sensor · Validation · Digital measures

Abstract

The EVIDENCE (EVALuating connecteD sENsor teChnologiEs) checklist was developed by a multidisciplinary group of content experts convened by the Digital Medicine Society, representing the clinical sciences, data management, technology development, and biostatistics. The aim of EVIDENCE is to promote high quality reporting in studies where the primary objective is an evaluation of a digital measurement product or its constituent parts. Here we use the terms digital measurement product and connected sensor technology interchangeably to refer to tools that process data captured by mobile sensors using algorithms to generate measures of behavioral and/or physiological function. EVIDENCE is applicable to 5 types of evaluations: (1) proof of concept; (2) verification, (3) analytical validation, and (4) clinical validation as defined by the V3 framework; and (5) utility and usability assessments. Using EVIDENCE, those preparing, reading, or reviewing studies evaluating digital measurement products will be better equipped to distinguish necessary reporting

requirements to drive high-quality research. With broad adoption, the EVIDENCE checklist will serve as a much-needed guide to raise the bar for quality reporting in published literature evaluating digital measurements products.

© 2021 The Author(s)
Published by S. Karger AG, Basel

Introduction

Digital measurement products are becoming increasingly prevalent for remote monitoring in clinical research and patient care. As described elsewhere, there are multiple factors that determine whether a remote monitoring tool can be considered fit for purpose in a stated context of use [1]. To determine whether a digital measurement product – or its component parts – is fit-for-purpose for use by participants or patients in a research study or clinical care, decision makers must rely on published peer-reviewed literature or complete the evaluations themselves.

Unfortunately, interpreting results from the current corpus of published work is challenging. Depending on the technology's maturity, studies evaluating it may be con-

ducted using a variety of study designs, data collection procedures, and analytic methodologies. Additionally, the quality of reporting across these studies is highly variable and often “characterized by irrational exuberance and excessive hype” [2, 3]. Inconsistencies in essential metadata reported and variability in evaluation protocols can lead to low confidence in study results [4, 5]. For example, a systematic review of studies evaluating digital measurement products conducted by the Clinical Trials Transformation Initiative found gaps in reporting such as: only 73% of studies reported the software used in the analysis, nearly 10% did not report the make and model of the technology, and there was substantial variation in documenting sensor modalities (e.g., “motion sensor,” “accelerometer,” “tri-axial accelerometer,” or “pedometer” without specifying the actual sensors contained within the product) [4]. Consequently, developments in the field of digital medicine may be slowed as evaluations are unnecessarily repeated, which is inefficient, expensive, and in some cases unethical. To speed the development and deployment of digital measurement products worthy of our trust, the quality of reporting of evaluation studies must improve [1].

This paper presents the EVIDENCE checklist (Evaluating connected sENsor teChnologiEs) intended for researchers, journal editors, and stakeholders who perform, publish, review, and/or analyze publications where the primary study objective is an evaluation of a digital measurement product or its constituent parts. Here, we define the types of evaluations to which EVIDENCE should be applied, with the goal of clarifying report requirements for each evaluation type. To align with broadly adopted research standards, EVIDENCE is structured similarly to existing publication checklists such as PRISMA for systematic reviews and meta-analyses, CONSORT for randomized clinical trials, STARD for diagnostic accuracy studies, and STROBE for observational studies in epidemiology [6–9]. We believe the EVIDENCE checklist will serve as a much-needed guide to raise the bar for quality reporting in published literature evaluating digital measurement products.

Scope of EVIDENCE

The EVIDENCE checklist displayed in Table 1 is intended for publications where the primary objective is an evaluation of a digital measurement product or its constituent parts. Digital measurement products, also referred to as connected sensor technologies or biometric monitoring technologies (BioMeTs), process data cap-

tured by mobile sensors using algorithms to generate measures of behavioral and/or physiological function [10]. Here, we use “digital measurement product” and “connected sensor technology” interchangeably. Although many of these tools can be considered “wearables,” digital measurement products encompass many form factors, such as portable monitors or under-mattress sleep trackers. We intentionally do not use the term “device” as not all digital measurement products are classified as medical “devices” per the FDA and other regulators [11, 12].

As defined in Table 2, there are 5 types of evaluations to which EVIDENCE can be applied: (1) proof of concept; the V3 framework consisting of (2) verification, (3) analytical validation, and (4) clinical validation; and (5) utility and usability assessments. These evaluations may occur in a variety of settings from the bench to free-living conditions, but the intended use of these digital measurement products should be remote monitoring outside of the clinic.

A proof-of-concept study is one that conducts initial testing intended to indicate whether the use of a technology or the development of a digital measure may be feasible in a given context of use [13]. In many cases, proof-of-concept studies are conducted to determine whether pursuing a full analytical or clinical validation study is worthwhile. Many evaluation studies will not meet criteria for the V3 framework as predefined protocols and acceptance criteria for many measures from connected sensor technologies have not been established. For example, sensor-based measures of forgetfulness, smartphone-based measures of eye tracking or gaze, and actigraphy to predict mood do not have defined evaluation protocols or acceptance criteria [14–16]. When performed to a rigorous standard, proof-of-concept studies can characterize measurement properties to inform power calculations in subsequent V3 evaluations. Therefore, it is appropriate to use the EVIDENCE checklist to guide reporting for proof-of-concept studies to support decision making about whether to conduct a full validation of the digital measurement product.

An evaluation within the V3 framework includes verification, analytical validation, or clinical validation. Verification assesses the accuracy of sample level sensor data compared to a bench standard. Analytical validation assesses the ability of a sensor and accompanying algorithm(s) to capture the behavioral or physiological concept accurately in an intended context of use. Clinical validation determines whether the digital clinical measurement is meaningful to answer a specific clinical question in a specific population [10]. Table 2 identifies examples of each. For more information on V3 classification with additional examples, reference Table 8 in the

Table 1. EVIDENCE (Evaluating connected sENsor teChnologiEs) Checklist

Section/topic	No.	Importance	Checklist item	Proof of concept	Verification	Analytical validation	Clinical validation	Utility and usability	Page No.
<i>Title</i>									
Title	1	Preferred	Explicitly identify the study as proof of concept, verification, analytical validation, clinical validation, and/or utility and usability. If limited by journal-specified word length, it is recommended to include the evaluation type as key words.						
<i>Abstract</i>									
Structured summary	2	Required, individual elements as applicable	Provide a structured summary including the following items, as applicable to the study: evaluation type (proof of concept, verification, analytical validation, clinical validation, and/or utility and usability), study objectives, concept of interest and outcomes measured, description of patient population, digital measurement products used, wear location, reference standard, sample size, and key results.						
<i>Introduction</i>									
Rationale	3	Required	Define study rationale in the context of what is already known and any existing gaps in the field.						
Objectives	4	Required	Clearly state the research question and study aims.						
<i>Methods</i>									
Ethics and informed consent	5	Required	Include a statement that IRB approval or ethics committee review of the study documentation was completed. Indicate whether written consent was obtained from the study participants.						
Protocol and registration	6	Preferred	When evaluation studies are conducted as part of an interventional clinical trial, document the clinical trial's registration number and whether or not the protocol can be accessed.						
Participants	7	Required	Define the recruitment strategy and inclusion and exclusion criteria for study participants.						
Sample size	8	Required	Indicate how the sample size was determined. In cases of N-of-1 studies, authors may describe the sample size based on number of measurements rather than the number of participants.						
<i>Connected sensor technology</i>									
Make and model	9a	Required	State the make and model of the connected sensor technology used.						
Selection rationale	9b	Preferred	Describe why the connected sensor technology was chosen for the study.						
Product availability/maturity	9c	Preferred	Describe if the connected sensor technology is a custom prototype or a product that is currently on the market, available for purchase.						
Sensor characteristics	9d	Required	Describe the sensor modality(ies) and sample level data characteristics (e.g., units, sampling rate, etc.) used for data collection in as much detail as possible.						
Form factor and wear location	9e	Required	Describe the form factor (physical shape) and wear location (precise anatomic position of sensor).						
<i>Software</i>									
Algorithm description	10a	Required	Describe in as much detail as possible the algorithm used for data analysis in the study. If a new algorithm is being created, describe in as much detail as possible the procedure for building the algorithm. Procedures used for validating the algorithm can be included in the statistical analysis section.						
Version number and manufacturer	10b	Required	State the version number and manufacturer of any software used for data collection and analysis where possible.						
Outcome assessed	11	Required	Clearly identify the outcomes to be measured.						
Data collection protocol	12	Required	Describe experimental procedures to collect data.						
Wear time	13	Preferred	Determine the minimum wear time for sufficient data capture and a meaningful data set used in the analysis.						
Reference standard	14	Required	Describe the standard to which the performance of the connected sensor technology is being compared.						

Table 1 (continued)

Section/topic	No.	Importance	Checklist item	Proof of concept	Verification	Analytical validation	Clinical validation	Utility and usability	Page No.
Statistical analysis	15	Required	Describe relevant statistical analyses to perform verification, analytical, and/or clinical validation of the solution utilized in research.						
Training for staff and/or participants	16	Preferred	Describe any training given to study participants and/or staff for how to properly use the connected sensor technology.						
<i>Results</i>									
Participant flow	17	Required	A diagram similar to a CONSORT flowchart is strongly recommended to show numbers for participant recruitment to study completion.						
Participant demographics	18	Required	Describe the participant demographics that are minimally necessary for the study.						
Numbers analyzed/findings	19	Required	Describe the study's findings, including missing data.						
<i>Utility and usability</i>	20								
Technical problems	20a	Preferred	Describe any technical problems that impacted the study results.						
Adverse events	20b	Required	Describe unintended effects of technology causing physical or psychological harms.						
Feedback from participants and study staff	20c	Preferred	Describe any feedback from participants and study staff and/or findings from satisfaction surveys.						
<i>Discussion</i>									
Summary of findings	21	Required	Summarize the main findings and relevance for the patient population and its clinical application as appropriate.						
Comparison to existing literature	22	Required	Compare results to similar studies and describe potential reasons for any major differences observed.						
Limitations	23	Required	Discuss limitations of study methods and/or the connected sensor technology used.						
Conclusions	24	Required	Provide interpretation of findings and implications for future research.						
<i>Other</i>									
Funding and competing interests	25	Required	Describe sources of funding or other support received for work.						
This checklist is intended for studies of which the primary outcomes are the evaluation of a digital measurement product. White/blank, required; grey, preferred; black, not required. Identify the evaluation type: _____.									

study of Goldsack et al. [10]. While analytical and clinical validation studies are performed in human subjects, verification testing is performed at the bench. Thus, in Table 1 there are items identified as not applicable for verification studies. The V3 framework has been steadily gaining traction in the field [1, 10, 17, 18]. With the EVIDENCE checklist, we will further clarify reporting requirements for each step in the V3 process to further its adoption.

Utility and usability assessments evaluate the practical considerations of using the technology in an individual's daily life [19]. Utility refers to whether a product has the features that users need, and usability is how easy and pleasant those features are to use. For example, comfort, ease of set-up, adverse effects, or technical failures could be assessed [19]. This information may be collected through satisfaction surveys or inferred from participant willingness to wear or use the technology for the duration of the study. Utility and usability measures may be a secondary aim of an analytical or clinical validation study. Understanding expectations from study staff, participants, and caregivers is essential for reducing the likelihood of missing data. Even though a connected sensor technology has met V3 criteria, it may be uncomfortable or difficult to use. If these difficulties significantly limit data collection in a pivot clinical trial, the study failure will be costly. By including utility and usability, as an evaluation type applicable for EVIDENCE, we aim to elevate the importance of these assessments.

The following are out of scope for the EVIDENCE checklist:

- Studies evaluating the performance of electronic patient-reported outcomes or digital therapeutics, although some components may be applicable to those technologies
- Studies evaluating performance of digital measurement products that measure adherence to an intervention such as smart pill boxes
- Studies using animals, tissues or other biological specimens
- Systematic reviews and meta-analyses of studies evaluating connected sensor technologies
- Studies evaluating security, data privacy or operational considerations of digital measurement products

Development of the EVIDENCE Checklist

The EVIDENCE checklist was developed by an interdisciplinary group of experts convened by the Digital Medicine Society (DiMe). The DiMe is a nonprofit orga-

nization dedicated to advancing the safe, effective, ethical, and equitable use of digital technologies to optimize health through research, communication and education activities, and community building. Using the PRISMA and CONSORT checklists as guides [6, 20], an initial draft of the checklist items was developed by the first, second, and senior authors (C.M., N.M., and J.C.G.) in July 2020. A virtual 1-day workshop was held in August 2020 to solicit feedback from the DiMe community. Twenty-one colleagues attended the workshop from different organizations including pharma, clinical care, technology developer, and regulatory. Many of these colleagues hold senior leadership positions within their respective organizations, have extensive experience developing, deploying, and/or evaluating these technologies, and have made significant contributions to connected sensor technology research as authors and peer reviewers. Following the workshop, individuals were invited to provide written feedback, with 12 colleagues participating. The first, second, and senior authors (C.M., N.M., and J.C.G.) consolidated feedback to develop a second version of the checklist and manuscript, which was circulated for feedback in November 2020. This process of asynchronous expert review and feedback was repeated 4 times before final approval of the checklist from the group.

We present each checklist item with examples from the literature. Examples may have been edited to remove citations, spell out abbreviations, and make certain words bold for emphasis. Some examples will include terminology or phrasing that is not aligned with the checklist recommendations. We explain the inclusion rationale for each item with additional evidence from the literature. The items are presented in order from 1 to 25; however authors do not need to include the items in this specific order in their publications.

The EVIDENCE Checklist

Title

Item 1 – Title – Preferred

Explicitly identify the study as proof of concept, verification, analytical validation, clinical validation, and/or utility and usability. If limited by journal-specified word length, it is recommended to include the evaluation type as key words.

Example

“Vital Signs Monitoring with Wearable Sensors in High-risk Surgical Patients: A Clinical Validation Study” [21].

Table 2. EVIDENCE checklist applies to 5 types of studies

Type of study	Definition	Common distinguishing characteristics	Examples
Proof of concept	Conducts initial testing intended to indicate whether the use of a technology or the development of a digital measure may be feasible in a given context of use	Described as a pilot study with a small sample size and a short duration Evaluating a novel measure that does not have predefined protocols and acceptance criteria	Sensor-based measures of forgetfulness [14] Smartphone-based measures of eye tracking or gaze [15] Actigraphy to predict mood [16]
V3 Framework Verification	Measures the accuracy of sample level sensor data compared to a bench standard	No human subjects	Raw data from the ECG sensor is accurate, precise, and consistent [58]
Analytical validation	Determines the ability of a sensor and accompanying algorithm(s) to capture the behavioral or physiological concept accurately in an intended context of use	Comparison to a reference standard Measure has a defined protocol and acceptance criteria	Accuracy of heart rate variability compared with a traditional ECG and Kubios clinical grade software [59]
Clinical validation	Determines whether the digital clinical measurement is meaningful to answer a specific clinical question in a specific population	Measurement performance in healthy controls compared to those with the disease Identifies or predicts a meaningful change	Heart rate variability identifies the presence of autism [82]
Utility and usability	Evaluates the practical considerations of using the technology in an individual's daily life	Assesses whether all of the necessary features exist and how pleasant these features are to use	Assessing technical problems and the comfort of wearable devices [83]

Explanation. Identifying the evaluation type in the title may improve indexing and streamline identification of appropriate studies for individuals conducting literature reviews.

There are certain terms that should be avoided in the title and throughout the manuscript in order to build a foundation for standardized terminology. For example, “feasibility” is a term that is widely used, even in some of the examples provided in this checklist. “Feasibility” should be avoided as the term could reflect a number of performance metrics and requires more context to be meaningful [10]. For the same reason, “valid,” “validity,” “verify,” and “validation” without designating analytical validation or clinical validation should be avoided [10].

Abstract

Item 2 – Structured Summary – Required

Provide a structured summary including the following items, as applicable to the study: evaluation type (proof of concept, verification, analytical validation, clinical validation, and/or utility and usability), study objectives, concept of interest, outcomes measured, description of the patient population, digital measurement products used, wear location, reference standard, sample size, and key results.

Example

Aims. “Early detection of atrial fibrillation (AF) is essential for stroke prevention. Emerging technologies such as smartphone cameras using photoplethysmography (PPG) and mobile, internet-enabled electrocardiography (iECG) are effective for AF screening. This study compared a PPG-based algorithm against a cardiologist’s iECG diagnosis to distinguish between AF and sinus rhythm (SR).”

Methods and Results. “In this prospective, two-centre, international, clinical validation study, we recruited in-house patients with presumed AF and matched controls in SR at 2 university hospitals in Switzerland and Germany. In each patient, a PPG recording on the index fingertip using a regular smartphone camera followed by iECG was obtained. Photoplethysmography recordings were analysed using an automated algorithm and compared with the blinded cardiologist’s iECG diagnosis. Of 672 patients recruited, 80 were excluded mainly due to insufficient PPG/iECG quality, leaving 592 patients (SR: $n = 344$, AF: $n = 248$). Based on 5 min of PPG heart rhythm analysis, the algorithm detected AF with a sensitivity of 91.5% (95% CI 85.9–95.4) and specificity of 99.6% (97.8–100). By reducing analysis time to 1 min, sensitivity was reduced to 89.9% (85.5–93.4) and specificity to 99.1% (97.5–99.8). Correctly classified rate was 88.8% for 1-min PPG analysis and dropped to 60.9% when the threshold for the analysed file was set to 5 min of good signal quality.”

Conclusion. “This is the first prospective clinical two-centre study to demonstrate that detection of AF by using a smartphone camera alone is feasible, with high specificity and sensi-

tivity. Photoplethysmography signal analysis appears to be suitable for extended AF screening” [22].

Explanation. Since abstracts are often used as a screening tool, including metadata about the technology and patient population is important. Authors are encouraged to provide comprehensive details so that those who may not have access to the full text can draw appropriate conclusions.

Introduction

Item 3 – Rationale – Required

Define the study rationale in the context of what is already known and any existing gaps in the field.

Example

“The use of subjective, episodic, and insensitive clinical assessment tools, which provide sparse data and poor ecological validity, can be an impediment to the development of new therapies... Clinical assessments performed using the MDS-UPDRS are time-consuming, require the presence of a trained clinician, are inherently subjective and lack the necessary resolution to track fine grained changes... A home diary completed for a few days preceding clinic visits by the patient or caregiver is another instrument that is commonly used in clinical trials for evaluating treatment efficacy based on a report of motor symptoms experienced outside the clinic. However, issues such as lack of compliance, recall bias and diary fatigue limit the accuracy of information that can be collected with this approach. The limitations of these tools contribute to the need for large sample sizes and long durations of clinical trials for new therapies, and increase the risk of failures” [23].

Explanation. A clearly stated rationale helps readers and reviewers understand the importance of conducting the study. In many cases it will be beneficial to outline limitations of current clinical assessments and describe how the digital clinical measurement will benefit a patient population. If there are existing connected sensor technologies for the study’s use case, they should be described.

Item 4 – Objectives – Required

Clearly state the research question and study aims.

Example

“Here, we present the development and validation of a method for continuous, objective assessment of resting tremor and bradykinesia based on data from a single wrist-worn accelerometer” [23].

“The aim of this study was to evaluate feasibility of physical activity measurement by accelerometry in colorectal cancer patients under free-living conditions at 6, 12 and 24 months after surgery, to evaluate the appropriate wear time and to compare results to pedometry” [24].

Explanation. Objectives are the questions which the study is designed to answer. It is critical that the study objectives be written clearly so readers and reviewers understand the scope. For clarity and uniformity in the research field more broadly, we suggest following the PICOS approach, as described in Box 2 of the PRISMA checklist [6].

Methods

Item 5 – Ethics and Informed Consent – Required (Excluding Verification Studies)

Include a statement that institutional review board (IRB) approval or ethics committee review of the study documentation was completed. Indicate whether written consent was obtained from the study participants.

Example

“The study had approval from the Tufts Medical Center and Tufts University Health Sciences Institutional Review Board. All participants in the study gave written informed consent prior to enrollment” [23].

Explanation. The IRB or ethics committee oversees that the study meets criteria to ensure the safety, privacy, and data protection of participants. In the manuscript, authors are encouraged to include the name of the IRB, the protocol ID, and the date of approval. For IRB there are 3 types of review pathways, depending on the risk level (e.g., minimal or greater than minimal risk of harm) and type (e.g., psychological, physical, or economic) [11]. If authors are unsure whether their study requires an IRB or ethics committee review, we encourage them to check regulations appropriate to their geography. In the USA, the Office for Human Research Protections provides detailed information concerning decisions on when IRB oversight is required [25]. As indicated in Table 1, this item is not applicable to verification studies.

Item 6 – Registration and Protocol – Preferred

When evaluation studies are conducted as part of an interventional clinical trial, document the clinical trial’s registration number and whether or not the protocol can be accessed.

Example

“The phase 2b, interventional clinical trial (ClinicalTrials.gov identifier: NCT02333331) recruited 217 patients...” [26].

Explanation. Analytical validation, clinical validation, or utility and usability evaluations may be conducted as part of a clinical trial of a medical product. Including the registration number can help create links between pub-

lished peer-reviewed literature and ClinicalTrials.gov data. For more information on whether a study should be registered, see Applicable Clinical Trial (ACT) requirements in the USA [27]. If the protocol can be accessed, explain how and where to find it.

Item 7 – Participants – Required (Excluding Verification Studies)

Define the recruitment strategy and inclusion and exclusion criteria for study participants.

Example

“Participants were included if they: (1) had multiple sclerosis (MS) as defined by 2010 International Panel criteria confirmed by a MS neurologist; (2) were ≥ 18 years of age; (3) were able to walk for at least 2 min with or without an assistive device; (4) had no clinical MS relapse within 30 days of cohort entry; and (5) had access to Wi-Fi Internet at home or in their community. Exclusion criteria included: (1) major musculoskeletal, cardiovascular or respiratory comorbidities that, in the opinion of the study investigators, could substantially impair physical activity and/or confound results; and (2) a clinical relapse within 30 days of cohort entry. Relapsing and progressive phenotypes were defined according to the 2014 Advisory Committee on Clinical Trials in MS Committee definitions. We recruited in blocks to a target goal based on EDSS: no disability (0–1.5), mild disability (2–3.5), mild ambulatory disability (4), moderate ambulatory disability (4.5–5.5), unilateral support needed for ambulation (6), and bilateral support needed for ambulation (6.5)” [28].

Explanation. It is best practice to include a figure or state in text the following items: inclusion and exclusion criteria, how many people were contacted, how many declined, how many were excluded because of exclusion criteria, how many enrolled, how were randomized, how many dropped out and why, and how many completed the study. As indicated in Table 1, this item is not applicable to verification studies.

Authors should be clear if the study enrolled both healthy participants and those with a disease or condition. When describing the patient population, authors should be specific about symptom severity and/or treatments to clarify the disease phenotype for which the study outcomes will be most relevant. As shown in the example above, symptom severity should be classified using current clinical assessment criteria rather than subjective categorizations of mild or severe. If not already covered in the Item 3 – Rationale, authors should define the reasoning behind the defined inclusion/exclusion criteria. For example, authors should describe why only a subset of the total available population is included in the study. This information is especially important for clinical validation studies assessing if the digital clinical measure

meaningfully answers a specific clinical question in a specific population [10].

If public datasets are utilized, authors should describe the dataset as well as the rationale for use. Rationale for use could be that the database contains labeled data sets for specific activities of interest and contains the same sensing modalities (e.g., accelerometer) and similar sensor characteristics (e.g., appropriate dynamic range, sampling rate) as the technology chosen for the study [29]. It is also suggested that authors describe any data cleaning efforts performed (e.g., excluding subjects due to missing data or unusable data), if applicable.

Item 8 – Sample Size – Required (Excluding Verification Studies)

Indicate how the sample size was determined. In cases of N-of-1 studies, authors may describe the sample size based on number of measurements rather than the number of participants.

Example

“A priori sample size of 23 participants was calculated based on the most conservative findings (correlation of 0.5), α level = 0.05, and a power of 0.80” [30].

Explanation. Authors should describe: (1) how many participants were recruited, (2) how many participants went into the final analysis, and/or (3) how many data collection periods were recorded and (4) how many data collection periods were utilized in the final analysis. It is strongly recommended that this information be presented as a participant attrition table. It is best practice to include a power calculation that justifies the sample size chosen and that it can support the results intended to be observed. In this section, authors should document whether or not a formal power calculation was performed. If it was performed, the authors should state which assumptions and data set were used. For analytical or clinical validation studies, authors should document whether or not a formal power calculation was performed a priori. If it was performed, the authors should state which assumptions and data set were used. Authors should include a reference to the methodology used for the sample size calculation. As shown in Table 1, this item is not applicable for verification studies.

Connected Sensor Technology

Item 9a – Make and Model – Required

State the make and model of the connected sensor technology used.

Example

“Each participant was asked to wear a single tri-axial accelerometer-based BWM (Axivity AX3; York, UK; dimensions: 23.0 × 32.5 × 7.6 mm; weight: 9 g; accuracy: 20 parts per million) which has been validated for its suitability in capturing high-resolution data akin to human movement” [31].

Explanation. Stating the make and model of the connected sensor technology is vital. This is especially important for manufacturers that have multiple product lines in different form factors. For example, the Fitbit Zip, a clip worn on the hip was recently discontinued in March 2019 and studies using this product were still being published in 2020 [32]. Without identifying the Zip, readers may incorrectly assume the results apply to currently available wrist-worn Fitbit products. Authors may consider including a picture or diagram of the technology, especially if the product is not widely used or known.

Item 9b – Selection Rationale – Preferred

Describe why the connected sensor technology was chosen for the study.

Example

“The recorded data is uploaded online to a user-friendly personalized account, and is easily searchable by date and time with a resolution of 15-min time intervals. FitBit is considered one of the leaders in the market of wearable activity sensors and, at a cost of under USD 60, the Zip model is far more affordable than comparable devices” [33].

“Each participant was asked to wear a single tri-axial accelerometer-based BWM (Axivity AX3; York, UK; dimensions: 23.0 × 32.5 × 7.6 mm; weight: 9 g; accuracy: 20 parts per million) which has been validated for its suitability in capturing high-resolution data akin to human movement... The BWM was located on the fifth lumbar vertebra... attached directly to the skin with double sided tape...” [31].

Explanation. Understanding why a particular connected sensor technology was chosen over other alternatives, if any are available, can be helpful for readers looking to reproduce the work. Example rationales can include: operationalization advantages (e.g., low burden for purchasing/procurement), meeting the minimum recording duration requirements (e.g., battery life and memory storage allowing for desired multi-day recording), or optimizing the subject/clinical site experience while using the technology (e.g., low burden on product setup and data extraction).

If the study was bring-your-own-device (BYOD), it is recommended that authors provide rationale along with details regarding safeguards implemented to ensure consistent data collection and improve data quality. For example, if the primary method of collecting data is via a

mobile application installed on a smartphone, rationale for leveraging a BYOD model could be increased recruitment and easier access to subjects located in different geographical locations. Example safeguards to ensure consistency and improve data quality could include specifying smartphone characteristics in the inclusion criteria (e.g., operating system: Apple iOS/Android; smartphone model: Apple iPhone 7 and up).

Item 9c – Product Availability/Maturity – Preferred

Describe if the connected sensor technology is a custom prototype or a product that is currently on the market, available for purchase.

Example

“Besides the aTUG system, a wearable system was utilized, which is also commercially available” [34]

Explanation. When considering replicating or deploying sensors described in research and assessing generalizability of results, it is helpful for the readers to know if the sensor is readily available for purchase or if it is a prototype in development. This information can be especially important for clinical trial sponsors who may be looking to deploy the solution into a multi-site clinical trial. Including the sensor release date, if known, is preferred to indicate if the sensor is still available for purchase. Authors should refrain from classifying sensors as “medical grade” or “consumer devices” as these terms do not provide insight into quality of sample level data [19]. Products from traditionally consumer-facing companies have been shown to take accurate measurements and a medical device designation does not render a product “fit for purpose” by default [35, 36]. If the sensor has regulatory clearance (e.g., FDA 510k cleared), citing reference documents outlining clearance is suggested.

Item 9d – Sensor Characteristics – Required

Describe the sensor modality(ies) and sample level data characteristics (e.g., units, sampling rate, etc.) used for data collection in as much detail as possible.

Example

“All participants were equipped with the OPAL system, sample rate 128 samples/s, 3DOF accelerometer (range ±16 g) and 3DOF gyroscope (range ±2,000°/s) (APDM, Inc., Portland, OR, USA)” [37]

“Data were collected with an inertial sensor measurement system consisting of 2 sensor units (Shimmer Sensing, Dublin, Ireland), including: (1) a tri-axial accelerometer (Freescale Semiconductors MMA7361, range ±6 g, sensitivity of 200 mV/g) and a (2) tri-axial gyroscope (InvenSense 500 series, range ±500°/s, sensitivity ±2 mV/°/s)” [38]

Explanation. Clearly describing the sensor characteristics used for data collection will enable reproducibility and facilitate readers' understanding on the applicability of the sensor to measure the intended activity of interest. Authors should elaborate on what sensing modalities were used in the study. For example, if the measurement is taken with an inertial measurement unit (IMU), indicate whether the sensor is a 3-axis or a 6-axis IMU. Authors are also encouraged to describe all the sensing modalities included in the product, as sometimes features can be added or removed in the product's lifecycle.

Many terms (preprocessed, raw) are used to describe the data coming off a sensor. We recommend using "sample level" to be consistent with language proposed in the V3 framework [10]. Reporting appropriate characteristics of sample level data is important as it is related to the ability of the chosen sensing modality to measure the use case of interest. For example, if accelerometers are used, the sampling rate and dynamic range of the data collected should be presented to better inform if the measurements collected adequately capture the motion of interest. If higher-intensity activities such as playing a sport or running are measured with accelerometers, low sampling rates and dynamic range settings would not be appropriate [39]. If sample level data is resampled, authors should describe this process, such as "the raw 3-D accelerometer data from both wrists in units of g sampled at 100 Hz were read into Matlab... synchronized with one another and down-sampled to 20 Hz" [40].

Item 9e – Form Factor and Wear Location – Required

Describe the form factor (physical shape) and wear location (precise anatomic position of sensor).

Example

"All participants were equipped with the OPAL system, sample rate 128 samples/s, 3DOF accelerometer (range ± 16 g) and 3DOF gyroscope (range $\pm 2,000^\circ/\text{s}$; APDM, Inc., Portland, OR, USA). Data obtained from the IMU at the lower back were used for this analysis" [37].

"Participants were asked to wear the device on their nondominant wrist as much as possible except while swimming and instructed to continue with their normal daily lives" [28].

"Each participant was asked to wear a single tri-axial accelerometer-based BWM (Axivity AX3; York, UK; dimensions: $23.0 \times 32.5 \times 7.6$ mm; weight: 9 g; accuracy: 20 parts per million) which has been validated for its suitability in capturing high-resolution data akin to human movement... The BWM was located on the fifth lumbar vertebra... attached directly to the skin with double sided tape..." [31].

Explanation. Authors should provide as much detail as possible on the form factor of sensors utilized and the body location that the sensor is affixed to. Form factor details can inform applicability for long-term monitoring as well as impact on patient experience. For example, large and noticeable products may be burdensome for patients to wear and can reduce compliance during extended periods of wear time compared to flexible, patch-based products [23]. Details about body location may be driven by the clinical concept being measured. For example, if measuring parkinsonian tremor in the arm, authors may choose to affix the technology to the most affected side [23]. Form factor constraints may require sensor technologies to take measurements from locations that differ from reference standards, which may impact the accuracy and reliability of measurements (e.g., optical heart rate sensing on the wrist compared to traditional ECG measurements). If applicable, authors should also describe the protocol for proper placement of the technology as indicated by the manufacturer, especially if that differs from how the technology was worn in the study. For technologies not worn on the body, authors should describe the placement, such as on a bedside table or under a mattress, that is required for high-quality measurement. Lastly, providing a picture of the sensor or diagram of where the sensor(s) is placed on the body is encouraged.

Software

Item 10a – Algorithm Description – Required (Excludes Verification Studies)

Describe in as much detail as possible the algorithm used for data analysis in the study. If a new algorithm is being created, describe in as much detail as possible the procedure for building the algorithm. Procedures used for validating the algorithm can be included in the statistical analysis section.

Example

Utilizing Previously Published Algorithms. "Accelerometer signals were transformed to a horizontal-vertical coordinate system, and filtered with a 4th order Butterworth filter at 20 Hz. The calculation of the 14 gait characteristics representative of 5 domains (pace, variability, rhythm, asymmetry and postural control)... the same methodology was applied to both the groups. Briefly: the initial contact (IC, heel strike) and final contact (FC, toe-off) events within the gait cycle were identified from the Gaussian continuous wavelet transform of the vertical acceleration. ICs and FCs detection allowed the estimation of step, stance and swing time. The IC events were also used to estimate step length using the inverted pendulum model. To estimate a value for step velocity we utilized the simple ratio between step distance (length) and step time" [31].

New Algorithm Development. “We trained a binary machine learning (ML) classifier to detect periods of gait from the raw accelerometer data. Observations of the positive class (gait) were derived from 2 gait tasks (2.5- and 10-m walk) whereas the remaining tasks (excluding the ADL tasks that included walking) from each visit were used to derive observations of the negative class (not gait). All data from the available HC and PD subjects were used for training the gait classifier model.”

“The pipeline for training the gait classifier, included steps for preprocessing, feature extraction, feature selection, and model training/evaluation. The raw acceleration data was band-pass filtered using a first-order Butterworth IIR filter with cutoff frequencies of 0.25–3.0 Hz to attenuate high-frequency movements associated with tremor. We then projected the band-pass filtered 3-axis accelerometer signals along the first principal component derived using principal component analysis (PCA) to generate a processed signal that is independent of device orientation. These preprocessing steps yielded 4 processed time series of acceleration signals (3 band-pass filtered signals and 1 PCA projection). The signals were then segmented into 3-s nonoverlapping windows and a total of 47 time and frequency domain features (listed in supplementary Table 3) were extracted from each window. The number of observations was then randomly sampled to balance both the positive and negative classes prior to the feature selection step. Feature selection was performed using recursive feature elimination with cross-validated selection of the optimal features using a decision tree classifier. We then trained a random forest classifier using the selected features. A leave-one-subject-out approach was used to assess the performance (accuracy, precision, recall, and F1 score) of the gait detection model” [23].

Explanation. When utilizing a product with a proprietary algorithm(s), we recognize that details may be difficult or impossible to obtain. Stating that details could not be obtained from the manufacturer of interest may be sufficient for this section. However, the algorithm is a core component when performing analytical validation and any details that can be obtained should be included.

For studies developing new algorithms, authors should provide relevant details about the data used to build and validate the algorithm, relevant algorithm parameters and training routines (if machine learning is used), as well as performance, and limitations of the proposed approach. Details about the dataset used for building the algorithm should include any partitioning that was performed (e.g., training, validation, and testing sets), any manipulations performed on the sample-level data (e.g., preprocessing routines such as filtering the sample level signal), and any details on reference data used for validation (further explained in Item 14), if applicable. If reference data is used, authors should specify details about the reference device (e.g., lead setup in a polysomnography device used to obtain reference measurements of sleep). If human reviewers are used to annotate data, authors should provide descriptions of all

guidelines and instructional templates used by reviewers. Details about algorithm development and relevant parameters used should be explained. For example, if a machine learning approach is used, authors should describe the model type, relevant model parameters, any hyperparameter tuning performed, and training routines utilized (e.g. cross-fold validation, leave-one-subject-out validation, etc.). Details about algorithm performance and limitations of the approach can be included in the results and discussion section, respectively. All methods used to perform analytical and clinical validation should be included in the statistical analysis section. Further details on good practices can be seen here [41]. To increase transparency and enable reproducibility, authors are encouraged to share their work on public code repositories, if applicable [23, 42–44].

Item 10b – Version Number and Manufacturer – Required (Excludes Verification Studies)

State the version number and manufacturer of any software used for data collection and analysis where possible.

Example

Device	Firmware version	Software version
Microsoft band (v.1)	10.3.3304.0 09 R	1.3.10901.1
Basis peak (v.2)	Build 14.6.0	1.15.0
Fitbit surge	16.31.6.3	2.10.1
Apple watch	1.0.1	123632
Mio	01.01	2.4.5
PulseOn	2.36	1.1.9541JP
Samsung Gear		R720XXU2AOIL

([45] – supplementary file 1)

Explanation. Authors should provide names of all technology manufacturers and software version numbers used in the study. This can help readers and reviewers backtrack to identify various firmware versions used with the product that may be relevant if prior research is not able to be replicated within a reasonable margin of error.

Item 11 – Outcome Assessed – Required

Clearly identify the outcomes to be measured.

Example

“The primary outcome was bias and precision (95% limits of agreement [LoA]) of heart rate and respiratory rate of the wireless sensor compared with the bedside monitor... A secondary endpoint was the reliability of detecting true critical clinical conditions such as bradycardia (HR <50 beats/min), tachycar-

dia (HR >100 beats/min), bradypnoea (RR <12 breaths/min), and tachypnea (RR >20 breaths/min). Another secondary outcome was the reliability defined as time until the first occurrence of data loss (defined as a duration of a gap within the data of 2 min, 15 min, 1 h, or 4 h) and the overall amount of data loss from various causes” [46].

Explanation. For the EVIDENCE checklist to apply, the primary outcomes of the study should be related to an evaluation of a digital measurement product or its constituent parts as a proof of concept study, a study within the V3 framework or a utility and usability assessment. Outcomes should be identified as primary, secondary, or exploratory and the performance targets adopted should be stated. Outcomes are preferred in a table format with measurements and units [47]. If an analytical validation study is undertaken in order to assess performance of a tool that identifies the presence/absence of a behavioral/physiological status, the authors should provide the definition they used to determine whether the condition is present (positive diagnostic result) or absent (negative diagnostic result). For more on how to select outcomes of interest that matter most to patients, see prior work [47].

Item 12 – Data Collection Protocol – Required Describe experimental procedures to collect data.

Example

“All caregivers were mailed a package containing a Philips Actiwatch 2 (Philips Respironics, Bend, Oregon) and a sleep diary to record their child’s sleep. The actigraph was programmed to collect the data in 30 s epochs day and night for 7 consecutive days. Caregivers were instructed to place the device snugly on their child’s wrist. The watch was placed on the ankle for participant 1, due to recommendations for children under 2 years old. Although hand stereotypy is present in many of those with Rett syndrome (RTT), it does not occur during sleep. Thus, the watch was placed on the wrist consistent with other actigraphy studies.”

Questionnaires. Ad hoc questionnaires (described below) were included with the sleep watch to gather more information about each participant’s overall and daily health and mood. This included items related to each participant’s alertness, additional medications taken, pain experienced, and seizure activity for each day of the collection period. Due to the addition of new questionnaires during the study period, not all families completed all questions (completion rates described below). The CSHQ is a parent-completed questionnaire aimed at gathering information about different dimensions of children’s sleep. The questionnaire includes items about sleep onset and bedtime behavior, sleep duration, morning and night wakings, sleep anxiety, behavior during sleep, daytime sleepiness, parasomnias, and breathing of school-aged children. Items are scored on a 3-point scale based on how often they occurred in the previous week (1 = rarely or 0–1 times, 2 = sometimes or 2–4 times, and 3 = usually or 5–7 times), and higher scores in-

dicate more sleep-related problems. Of the 45 items on the questionnaire, 33 are scored for a score range of 33–99, and a score of 41 or more indicating a need for further evaluation of a potential sleep disorder. Eleven of 13 families received the CHSQ (participants 2 and 3 did not due to changes in study protocol). We evaluated internal consistency of the questionnaire using Cronbach’s α .

“A sleep diary is a tool that caregivers complete daily in the home environment to indicate the time their child was put to bed, the time their child fell asleep, any night wakings, and the time their child woke up in the morning, as well as any daytime sleep. Sleep diary tools are often included in actigraphy studies for verification of times and identifying artifact. Sleep diaries were completed by parents for each day of actigraphy recording and used to verify actigraph data during the editing process. Twelve of 13 families completed the sleep diary for a total of 78 of 91 nights (85.7%). Participant 9 did not return the sleep diary, and thus daytime sleep and parent-reported TNS, and total sleep time (TST), could not be calculated” [47].

Explanation. Authors should identify if the study is using secondary/retrospective data analysis or prospective data collection. In addition to the connected sensor technology description as defined in Item 9, authors should describe all measurement methods used. This is especially important if the tool or assessment was used to score or interpret the data obtained from the connected sensor technology. These may include clinical outcome assessments such as PRO or electronic patient-reported outcomes, participant diaries, or traditional clinical assessments. When outlining the protocol, include the frequency of measurement (e.g., once a day, once an hour), the location where the measurement was collected (e.g., in the lab, in the patient home) and duration of testing (e.g., 1 h, multiple months, during the daytime or only nighttime hours).

If the study includes a utility and usability assessment the methods and timing of feedback should be described [49]. For example, the method of soliciting feedback could be based on quantitative surveys, qualitative anecdotes, or testimonials from participants.

If applicable, for verification studies, describe if the sensors were tested under conditions (e.g., temperature and pressure) different from the conditions described by the manufacturer.

Item 13 – Wear Time – Preferred

Determine the minimum wear time for sufficient data capture and a meaningful data set used in analysis.

Example

“A day of recording was defined from noon to noon. Each day of recording was evaluated for quality. Any day with >4 h of missing data or >2 min of missing data during sleep in a main

rest interval was considered invalid. Data could be missing due to off-wrist detection or a technical failure of the device. In the entire Sueño study, 208 out of 15,719 days (1.3%) were discarded due to missing data. Only studies with ≥ 5 valid days were considered adequate for analysis” [50].

“Data were considered valid if the devices were worn for at least 4 days and for at least 6 h per day. Nonwear time was defined as at least 60 min of consecutive zero counts with a 2 min interruption tolerance” [24].

Explanation. Stating data quality thresholds for wear time is helpful for readers and reviews to understand how the data was cleaned. If applicable, describe the sensor/s and algorithm/s used to define wear time. For example, a temperature sensor or skin capacitance sensor could be used to determine wear time. The item may be especially important in clinical validation studies to determine what minimum wear time is required to capture meaningful information rather than simply report it. For example, when measuring gait speed, only 2 or 3 purposeful bouts of walking per day may be needed to get a daily average ([51] is suppl. Fig. S1). This item is also important in usability and utility studies to determine if minimum wear times set out by clinical validation can be met in practice. For example, the study cited in the above example found that 3 valid days of physical activity assessments in their study population of colon cancer patients was sufficient to achieve an intraclass correlation coefficient of 0.84–0.93 when comparing the first 3 days with the entire 10 days at all 3 follow-up time points [24].

Item 14 – Reference Standard – Required

Describe the standard to which the performance of the connected sensor technology is being compared.

Example

Proof of Concept. “Criterion standard: two researchers (B.T., E.B.) observed patients during each session with a physical therapist. Similar to methods used in previous studies, the gold standard for the actual number of steps was the average of the 2 values counted by each researcher using a mobile counting app” [33].

Verification. “In the test ($n = 35$) devices were mounted to a single-axis shaker table (manufactured by Instron) and subjected to 14 sets of sinusoidal oscillations. Each set had a different stroke length and amplitude and each was run for a period of 100 s. Sensors were mounted so all the forces affected the z-axis of the AX3. This axis was chosen as it has the most margin for error according to the manufacturer’s data sheet. Each AX3 was set to record with a range of ± 8 g and a sample frequency of 100 Hz” [52].

Analytical Validation. “23 male volunteers performed an exercise stress test on a cycle ergometer. Subjects wore a Polar RS800 device while ECG was also recorded simultaneously to extract the reference RR intervals. A time-frequency spectral analysis was performed to extract the instantaneous mean heart

rate (HRM), and the power of low-frequency (PLF) and high-frequency (PHF) components, the latter centred on the respiratory frequency. Analysis was done in intervals of different exercise intensity based on oxygen consumption. Linear correlation, reliability, and agreement were computed in each interval” [53].

Clinical Validation. “In the present study we have set out to test (if subjective accounts of disease are key components of measures of disease severity and quality of life) using visual analogue scales (VAS) for itch, as a subjective measure, and actigraphy as an objective measure” [54].

Explanation. The reference standard used will vary depending on the type of evaluation. For a verification evaluation, the sensor will be compared to a ground truth reference standard, such as a shaker table for an accelerometer as described in the example. In analytical or clinical validation, there may be multiple reference standard options available for a single metric and not all will be sensor based. For example to demonstrate analytical validation, sleep measures might be compared to polysomnography, heart rate measures from a patch with an ECG sensor could be compared to an ECG monitor previously analytically validated, gait measures might be compared to a motion capture system, and respiratory rate could be compared to manual counting of chest rise and fall [Table 3 in 1]. To demonstrate clinical validation, the digital clinical measure may be compared to an existing clinical outcome assessment (COA) or clinical instrument on its ability to distinguish health and sick population or moderate or severe presentations of a disease. In some cases, the field needs agreement on rigorous and quantitative reference standards that should be used [1]. This checklist is not advocating for particular standards for particular tools but rather for the importance of using a reference standard with a justification.

Although the term is used in the proof-of-concept example above, authors are encouraged to avoid the term “gold standard” as some may be suboptimal and only deemed the best available by consensus [10]. For example, in Duchenne muscular dystrophy (DMD) the 6-min walk test is often used in clinical trials of medical products to treat the disease. However, approximately 60% of DMD patients are nonambulatory or cannot walk well enough to adequately perform the test [55]. If authors are performing a clinical validation study of total arm movement measured with a wrist-worn accelerometer in DMD, comparing the performance to the 6-min walk test is not an equivalent comparison [55].

If the reference standard is another connected sensor technology or medical device (e.g., motion capture or ECG) with accompanying software, authors are encour-

aged to include make/model and associated metadata described in Items 9–10. Additionally, authors should include details about how the connected sensor streams are aligned with the reference standard. For example, the data may need time alignment between the reference product and sensor to ensure comparing the data from the same timer periods. If comparison to the reference standard requires any manual data processing, it is recommended to include a statement on whether this was undertaken blinded to other study data and undertaken independently from other analysts. Ideally, an auto-scoring algorithm would be validated against multiple human scorers rather than just one, as there is known variability across human scorers.

Item 15 – Statistical Analysis – Required

Describe relevant statistical analyses to perform verification, analytical and/or clinical validation of the solution utilized in research.

Example

Overview of Statistical Methods, Aggregation, and Software Used. “Statistical analysis was performed in R, version 3.4.1 (The R Project for Statistical Computing)... using the following packages: psych for intraclass correlation coefficient (ICC), BlandAltmanLeh for Bland-Altman plots, nlme for linear mixed-effects model, car for type 3 analysis of variance, and MASS for stepwise model selection” [56].

“For in-lab walk test, for each digital device and algorithm aforementioned, the median of gait metrics across all steps for each lap was computed. Then, the median values across all laps per visit were used for statistical analysis” [57].

Verification. “Depending on the distribution of data either a Student paired *t* test, or a Wilcoxon matched pairs test, was used to determine the differences between the data obtained from the ECG and HRM for both the RR intervals and the calculated HRV parameters” [58].

Analytical Validation. “To analyze the performance of the walking speed estimations for normal and impaired subjects, we report the root-mean-squared-error (RMSE), the Bland-Altman limits of agreement (LOA), and the slope (*m*) and intercept (*b*) of the following linear model: $y = my^{\wedge} + b$, where *y* corresponds to the truth values, and $y^{\wedge}$ corresponds to the associated estimates (median speed from each walking test)” [59]. “Test-retest reliability of gait features was assessed by calculating the ICC on data collected from healthy volunteers during visit 1 and visit 2” [56].

Clinical Validation. “Variation of features with the live rater’s item score was quantified by the Kruskal-Wallis test” [23].

“Finally, we establish concurrent validity in the context of MS, by examining the relationship between estimated and ground truth walking speeds sampled from the comfortable 6MWT of Protocol B and indicators of mobility impairment and fall risk. Specifically, the Pearson product moment correlation coefficient is used to characterize the relationship between walking speed and MSWS and EDSSSR scores, and the Mann-Whitney

U test is used to test for a significant difference in walking speed between subjects who reported a fall in the 6 months prior to the test and those who did not. For all statistical analyses, significance is assessed at the $\alpha = 0.05$ level” [59].

Explanation. Authors should describe all statistical analyses used to perform verification of sensor technologies and analytical and/or clinical validation of algorithm systems used in the solution [10]. Statistical analysis performed to verify sensor technologies may include assessments on intersensor reliability (reliability of measurements from multiple sensors from a given manufacturer), intrasensor reliability (reliability of measurements from a single sensor over time), or agreement of preprocessed outputs with a relevant reference standard. Statistical analyses performed for analytical validation of an algorithm system may include comparisons of algorithm outputs with respective reference standard measurements. For example, sensor-derived measures of sleep quantity compared to polysomnography readings. These could also include test-retest reliability of the algorithm outputs. Statistical analyses performed to demonstrate clinical utility (e.g., criterion validity: association between sensor measures and clinical ratings; discriminative validity: ability of sensor measures to discriminate between different disease states) of a given solution may include relevant comparisons of algorithm outputs with currently used clinical assessment tools or patient-reported outcomes. It is recommended to include confidence intervals as well as statistical significance where applicable. It also is suggested to include descriptions of any data cleaning or aggregation of data performed for analysis, along with motivation for doing so. Lastly, authors are encouraged to provide a statement highlighting statistical software and software versions used in analysis.

Item 16 – Training for Staff and Participants – Preferred

Describe any training given to study participants and/or staff for how to properly use the connected sensor technology.

Example

“Written instructions on operating the Fitbit software were provided to each participant” [60].

Explanation. Training for staff, study participants and caregivers will be most relevant when data collection is done within the participant’s home. A recent study found that study coordinators may desire “hands-on” experience with products and software to increase comfort level so this element should not be overlooked [61]. Simi-

larly, when surveyed on training preferences, clinical trial participants reported the highest comfort with in-person training followed by written instruction and a short video [62]. Describing training procedures is important as it may impact adherence to wearing the product and using the product properly to possibly reduce the number of technical errors. For more information on best practices for training, see *The Playbook: Digital Clinical Measures* and work from the Clinical Trials Transformation Initiative (CTTI) [17, 63].

Results

Item 17 – Participant Flow – Required (Excludes Verification Studies)

A diagram similar to a CONSORT flowchart is strongly recommended to show numbers for participant recruitment to study completion.

Example

Participant selection is shown in Table 1 in Perez et al. [64].

Explanation. It is important for readers and reviewers to know how many participants were recruited versus how many participants' data was used for analysis. Authors should describe reasons for any study exits, including those lost to follow-ups. If it is a prospective study, authors should include recruitment dates. A diagram is strongly preferred.

Item 18 – Participant Demographics- Required (Excludes Verification Studies)

Describe the participant demographics that are minimally necessary for the study.

Example

Characteristics of participants enrolled in the Apple Heart Study at baseline [Table 1 in 64].

Explanation. Presenting demographic information for participants contributing data to the study is critical to draw conclusions on the generalizability and/or applicability of a digital tool to a different population than that studied. Recognizing that demographic reporting requirements will likely vary by study context of use, authors could consider the following as examples of minimally necessary elements: age, sex/gender, race and/or ethnicity, and relevant comorbidities. This information can be displayed in a table, in the text, and/or in a supplementary table, depending on journal requirements.

Item 19 – Numbers Analyzed/Findings – Required Describe the study's findings, including missing data.

Example

"The mean difference and limits of agreement derived from the mixed effects models for RR of the SensiumVitals, EarlySense, and Masimo Radius-7 were all within the predefined accepted range as shown in Table 2. The HealthPatch overestimated RR, with a mean difference of 4.4 breaths/min and with wide levels of agreement of -4.4 to 13.3 breaths/min. The 95% limits of agreement calculated from the Bland and Altman method showed wider limits of agreement for all sensors. EarlySense showed the narrowest limits of agreement for RR. Figure 3a-d illustrates the Bland and Altman plots" [21].

"Data loss of HR measurements was 12.9% (83 of 633 h), 12.3% (79 of 640 h), 27.5% (182 of 664 h), and 6.5% (47 of 727 h) for SensiumVitals, HealthPatch, EarlySense, and Masimo Radius-7, respectively" [21].

Explanation. Clearly describing the data collected and study findings is a hallmark of a high quality study. It is suggested that authors include whether or not adjustments were made for multiplicity and hypothesis testing to enable the interpretation of *p* values. For analytical validation, authors should include results from a direct comparison between the calculated metric and reference standard, including the statistical analysis methods. If appropriate, utility and usability evaluations should include whether or not patients met wear time requirements set out by clinical validation [60]. Compliance with the protocol, such as hours a day the product was actually in use compared to what was expected, are important to report.

Utility and Usability

Item 20a – Technical Problems – Preferred

Describe any technical problems that impacted the study results.

Example

"There were no serious adverse events observed during the study. Five adverse events were recorded, including 3 upper respiratory tract infections and 2 technical difficulties in operating the device, which were not related to device malfunction" [65].

Explanation. This area is important to note as it may impact participant adherence in using the product, the amount of missing data at the study conclusion, and decisions to use the technology in future studies. While this item is not required, authors are highly encouraged to report if there are significant deterrents to the study from technological issues such as frequent Bluetooth connection failures.

Item 20b – Adverse Events – Required

Describe unintended effects of technology causing physical or psychological harms.

Example

“Eighty-nine percent (63/71) agreed that they did not experience any adverse effects related to using the device (median = 7, interquartile range = 6–7). Four patients developed a rash or skin irritation from the wristwatch, and 2 users found that the device disturbed the function of other home appliances” [66].

Explanation. Adverse events are critical considerations when evaluating the benefits of a technology. The Office of Human Research Protections defines adverse events as “any untoward or unfavorable medical occurrence in a human subject... associated with the subject’s participation in the research” [65]. This item is required on the checklist because IRB and ethics committees require adverse event reporting. While physical harm may be unlikely with connected sensor technologies, researchers should be mindful that self-monitoring can have a psychological burden [67, 68]. For studies deemed exempt from the common rule by an IRB based on minimal-risk-of-harm or proof-of-concept studies where monitoring may occur for short durations, there will likely be no serious adverse events to report [69]. In that case, we strongly encourage researchers to collect and report on Item 20a and Item 20c as these are valuable sources of information driving decisions to use the technology in future studies.

Item 20c – Feedback from Participants and/or Staff on Technology – Preferred

Describe any feedback from participants and study staff and/or findings from satisfaction surveys.

Example

“Approximately, 85% of subjects were either likely or very likely to wear the sensors for an extended period of time (Fig. 6). Of the subjects that were very likely to wear the devices for an extended period of time and reported them very comfortable, there was a marked preference (54.3 vs. 40%) for the flexible patch form factor. While both types of devices were rated highly by subjects for comfort, 7 out of 8 subjects reported sternum as the most uncomfortable location for devices with a rigid form factor, whereas 3 out of 4 subjects reported flexible patches placed on the lower extremity (thigh and ankle) as uncomfortable. We observed a high level of acceptance for the wrist location for either device types” [23].

“Most patients evaluated the device as good or very good at enrollment (89%, $n = 65$) and at the end of the study (87%, $n = 63$)” [65].

Explanation. Results of utility and usability assessments are important as they may impact decisions to use the technology in future studies. Reporting on negative results is a known challenge in the scientific community [70]. To build a foundation of transparent results, we en-

courage reporting of all feedback so researchers do not select only the positive findings.

Discussion

Item 21 – Summary of Findings – Required

Summarize the main findings and relevance for the patient population and its clinical application as appropriate.

Example

“In this prospective study, we demonstrate that physical activity monitors (PAMs) are a feasible tool for assessing long term physical activity in patients with cancer who are undergoing therapy. PAM-derived data also accurately correlated with clinician assessments and QOL measures using standardized tools. The number of steps per day separated patients with different clinician-assessed ECOG PS with extreme sensitivity and also correlated with multiple functional and QOL tools such as FACT-G, QIDS-SR16, and BFI” [71].

Explanation. Authors should give a balanced summary of the study results. There should be a clear statement as to whether a connected sensor technology meets expectations for verification, analytical validation and/or clinical validation. Especially in clinical validation, it is recommended to focus on clinical relevance rather than an overemphasis on p values.

Item 22 – Comparison to Existing Literature – Required

Compare results to similar studies and describe potential reasons for any major differences observed.

Example

“This is consistent with data in persons with musculoskeletal and neuromuscular conditions in an inpatient rehabilitation facility where consumer-grade activity trackers were less accurate under conditions in which stride lengths were shorter” [72].

“Our results are consistent with the findings in people with traumatic brain injury and stroke, which revealed greater accuracy in waist-worn trackers as compared to wrist-worn in the 2-min walk test” [72].

Explanation. Compare and contrast the findings of the study to others in a similar context of use that used either the same or different connected sensor technology. This is helpful for readers and reviewers to understand what value has been added to the field through this study. In some cases, authors may be publishing the first study evaluating a particular connected sensor technology or the first study in a unique patient population. If there no comparable studies in the literature, authors should restate the study rationale and articulate

how the study is essential to filling a gap in the literature.

Item 23 – Limitations – Required

Discuss limitations of study methods and/or the connected sensor technology used.

Example

“Study limitations include the relatively short duration of walking that occurred among the various tasks. In this study, participants engaged in 2-minute walk tests that ranged from 231 to 260 steps and simulated household and obstacle negotiation courses in which step counts ranged from 56 to 72 steps. Although 2-minute walk tests have been used to study the accuracy of activity trackers in other studies, longer duration walking tests may result in reduced variability and higher levels of accuracy – particularly given the higher rates of accuracy we observed with more continuous walking” [72].

“Finally, our sample population included patients with mild to moderate PD with an average walking speed of 1.26 m/s who are able to ambulate without the use of an assisted device. Results may not generalize to individuals with greater disease severity” [72].

“It should be recognized that the current study was based on a convenience sample, and it is worth pointing out the limitations created by such an approach. Generalizing to the RTT population is not warranted given no random sampling, and the results are best considered specific to the sample” [47].

“Further, the different devices used were not attached in the same location on the body. While this helped to minimize interference between devices, there might be some error due to the different attachment locations” [73].

Explanation. Authors should include limitations of the study design, technical limitations of the connected sensor technology, or generalizability of results from the study sample to the target patient populations and/or other patient populations. Operational limitations, such as scalability of technology for use in multisite or international trials, are especially important to note for custom or multicomponent products.

Item 24 – Conclusions – Required

Provide interpretation of findings and implications for future research.

Example

“This study provides evidence on the feasibility of using actigraphy, an objective, in-home recording system, to characterize sleep in Rett syndrome (RTT). Overall, some participants had age-appropriate levels of total sleep time and sleep onset within the recommended guidelines. On the other hand, the results indicate the presence of dysfunction for some sleep parameters in this RTT sample, specifically the continuance of daytime sleep across adolescence, low sleep efficiency, a lack of age-re-

lated changes in total night sleep, and clinically significant scores on the CSHQ. Future work should investigate the validity of using actigraphy to measure sleep in RTT, to establish an objective, in-home method to assess sleep in this population” [47].

Explanation. Conclusions should be closely linked to the study objectives. Authors should avoid drawing conclusions that go beyond the data presented. If no conclusions can be drawn due to limitations of the data collected, this is still an important finding for the field. A strong conclusion should use the summary of findings described in Item 21 to make recommendations for future research to build upon their work.

Other

Item 25 – Funding and Competing Interests – Required

Describe sources of funding or other support received for work.

Example

“This study was supported, in part, by the Mayday Foundation and NICHD grant No. 73126 and 44763” [47].

“This work did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. All authors disclose being share-holders of Empatica and having received salary or consulting fees from Empatica” [74].

Explanation. Authors should be transparent about sources of funding. Given that study findings could directly impact product sales, there is potential for studies funded by product manufacturers to unintentionally introduce biases.

Discussion

The EVIDENCE (EValuatIng connecteD sENsor teCh-nologiEs) checklist was developed by a multidisciplinary group of content experts from the Digital Medicine Society, representing the clinical sciences, data management, technology development, and biostatistics. The aim of EVIDENCE is to promote high quality reporting in studies where the primary objective is an evaluation of a digital measurement product or its constituent parts. Here we use the terms digital measurement product and connected sensor technology interchangeably to refer to tools that process data captured by mobile sensors using algorithms to generate measures of behavioral and/or physiological function. EVIDENCE is applicable to the following 5 types of evaluations: (1) proof of concept; the V3 framework, consisting of (2) verification, (3) analytical validation, and

(4) clinical validation; and (5) utility and usability assessments. Using EVIDENCE, those preparing, reading, or reviewing studies evaluating digital measurement products will be better equipped to distinguish necessary reporting requirements to drive high quality research.

EVIDENCE was developed to prompt consistency in reporting essential metadata for connected sensor technologies and their software. The intent is that this will drive a higher-quality body of literature evaluating digital measurement products, making it easier for decision-makers selecting digital tools to rely on existing studies rather than repeating them. Including appropriate metadata for connected sensor technologies and their software is important given: (1) the variability in specifications and (2) the potential time lag between study conduction and publication while technology updates happen quickly. As outlined in checklist Items 9a to 10b, describing the make and model, software version number, sensor modality, form factor, and wear location will enable readers to evaluate the relevance of a study years after completion to build a body of evidence for a specific methodology. Even in technical papers describing algorithm development, readers should be able to find the key information necessary for adequate interpretation. Ultimately, by including the consistent set of metadata described in EVIDENCE, direct comparisons across studies results can be more readily made.

By highlighting 5 applicable study types, EVIDENCE is intended to guide thoughtful evaluation of digital tool performance. Researchers, readers, and reviewers should be able to clearly discern the study objectives that align with one or more of the 5 evaluation types. Moreover, researchers should be able to identify with the appropriate study type while planning their evaluation, driving more focused assessments of digital measurement product. Validation studies within the V3 framework should be characterized by predefined protocols and acceptance criteria for measurement performance characteristics. For example, blood pressure monitors have well-established validation protocols set by professional societies [75]. Currently, many measurements collected with connected sensor technologies lack this maturity. As such, most evaluation studies to date should be considered proof of concept. It is out of scope for EVIDENCE to define the protocols and acceptance standards for each measurement – there is a lot of variability across sensor types [76]. For example, Item 14 in the checklist will not tell authors which specific reference standard to use in every conceivable context of use. Rather, the intention for EVIDENCE is to highlight required items for reporting. By bringing consistency to reporting, EVIDENCE will allow

for stronger synthesis of proof-of-concept studies to drive development of such standards.

If a study that includes a connected sensor technology is not readily identifiable as a proof of concept, verification, analytical validation, clinical validation, or utility and usability evaluation, then authors, readers, or reviewers should reevaluate the study's objectives. If the study is not focused on security or data rights factors, then likely a proof of concept, V3, and utility and usability objective should be considered. For example, if the product is used in a cross-sectional or observational study where the objective is to assess a disease state (e.g., correlations between physical activity and multiple sclerosis), authors should consider refining the objectives to a proof-of-concept investigation of clinical validation or assessing an element of utility and usability. If we do not build a strong body of evidence around these 5 evaluation types, we will be unable to draw conclusions on the tool's performance.

EVIDENCE has similarities and differences to existing publication checklists in terms of content and methodologies of development. Many checklist items that may seem obvious to experienced researchers, such as title, abstract, rationale, objectives, limitations, and conclusions, were adapted from CONSORT and PRISMA items. With 25 items, EVIDENCE is well in line with the length of other checklist, which range from 22 to 29 items [6–9]. To keep pace with this rapidly evolving field and the proliferation of publications on digital sensor technologies, EVIDENCE was developed with less people on a shorter timeline than other checklists. For example, PRISMA, STARD, and STROBE were developed over multiday workshops consisting of 23–85 people, with 8–11 subsequent meetings and revisions [6, 8, 9]. However, the 21 participating experts for EVIDENCE allowed for a process of rapid iteration and focused development. The group was agile and included representatives from a variety of technical, clinical, and regulatory backgrounds, all with a deep and applied knowledge of connected sensor technologies, as well as scientific best practice.

As the professional home for all who serve in digital medicine, the The DiMe is uniquely positioned to drive adoption and take ownership of the revision process. The DiMe will take a similar approach as existing checklists by establishing a publicly available website (<https://www.dimesociety.org/tours-of-duty/EVIDENCE/>) and partnering with academic journals publishing applicable studies to endorse and adhere to the EVIDENCE checklist [77]. The EVIDENCE checklist website will provide a version of the checklist that can be downloaded and used in the journal submissions. There will be an open submission

form on the website for update requests. Update requests will be responded to as needed by the first and senior authors of this paper. A workshop will be convened annually by the DiMe Research Committee to review update requests and proposed revisions. Using the DiMe community and given the rapid evolution of technologies, we intended for updates to EVIDENCE occur more regularly compared to the 5- to 10-year span for CONSORT, PRIMSA, and STARD [6–8]. Additionally, similar to PRISMA, on the website will be a public listing of journals who have endorsed the checklist [77]. Using direct outreach to prominent journals in the field and using connections through the DiMe community, the authors of this paper hope to build recognition and secure at least 3 endorsements in the coming year.

Finally, finding high-quality examples in the literature that met EVIDENCE requirements for terminology was difficult. Terms we encourage authors to avoid, such as “gold standard,” “feasibility,” and “validation,” were prevalent. Through EVIDENCE adoption, we hope to establish uniformity to the terminology used in peer reviewed literature. Both CONSORT and STARD were found to have an impact on improving reporting accuracy and quality in the years following their release [78–81]. As a part of the monitoring/revision process, the Research Committee at the DiMe intends to evaluate the impact of EVIDENCE on the body of literature in a similar fashion.

Conclusion

Interpreting results of studies evaluating the performance of connected sensor technologies is challenging. Publication checklists have historically been used to im-

prove quality and consistency in reporting. The EVIDENCE checklist, developed by experts in the DiMe community, is intended to raise the quality of publications leading to stronger protocols and more meaningful results to identify products worthy of our trust in a given context of use. DiMe is uniquely positioned to engage stakeholders, drive adoption, own the revision process, and assess the impact in the years to come.

Conflict of Interest Statement

C.M. is a full-time employee of Elektra Labs. J.B. is a full-time employee and shareholder of Philips. E.I. is an employee of Koneksa Health and may own company stock. S.S. has nothing to disclose. J.-L.P. and S.V. are employees and shareholders of Eli Lilly and Company. N.M. is a full-time employee and shareholder of Pfizer Inc. S.O.I. is a full-time employee and cofounder of Tibi Health Inc. B.V. is an employee and shareholder of Byteflies.

Funding Sources

No funding was received for this work. This publication is a result of collaborative research performed under the auspices of the DiMe.

Author Contributions

C.M., N.M., and J.C.G. contributed to the conception and design of the checklist and drafting of this paper. J.B., S.O.I., E.I., S.P., J.-L.P., S.V., B.V., and C.W. contributed to development and content of checklist items and substantial revisions of this paper.

References

- Godfrey A, Vandendriessche B, Bakker JP, Fitzer-Attas C, Gujar N, Hobbs M, et al. Fit-for-purpose biometric monitoring technologies: leveraging the laboratory biomarker experience. *Clin Transl Sci*. 2021 Jan;14(1):62–74.
- Leenen JP, Leentveld C, van Dijk JD, van Westreenen HL, Schoonhoven L, Patijn GA. Current evidence for continuous vital signs monitoring by wearable wireless devices in hospitalized adults: Systematic review. *J Med Internet Res*. 2020 Jun;22(6):e18636.
- Steinhubl SR, Topol EJ. Digital medicine, on its way to being just plain medicine. *NPJ Digit Med*. 2018 Jan;1(1):20175.
- Bakker JP, Goldsack JC, Clarke M, Coravos A, Geoghegan C, Godfrey A, et al. A systematic review of feasibility studies promoting the use of mobile technologies in clinical research. *NPJ Digit Med*. 2019 Jun;2(1):47.
- Badawy R, Hameed F, Bataille L, Little MA, Claes K, Saria S, et al. Metadata concepts for advancing the use of digital health technologies in clinical research. *Digit Biomark*. 2019 Oct;3(3):116–32.
- Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med*. 2009 Jul;6(7):e1000100.
- Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010 Mar;340:c869.

- 8 Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al.; STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015 Oct;351:h5527.
- 9 Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al.; STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *PLoS Med*. 2007 Oct;4(10):e297.
- 10 Goldsack JC, Coravos A, Bakker JP, Bent B, Dowling AV, Fitzer-Attas C, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). *NPJ Digit Med*. 2020 Apr;3(1):55.
- 11 Coravos A, Goldsack JC, Karlin DR, Nebeker C, Perakslis E, Zimmerman N, et al. Digital medicine: A primer on measurement. *Digit Biomark*. 2019 May;3(2):31–71.
- 12 US Food and Drug Administration. *Medical device overview*. 2018. [cited 2020 Nov 23]. Available from: <https://www.fda.gov/industry/regulated-products/medical-device-overview#What%20is%20a%20medical%20device>.
- 13 Goldsack JC, Dowling AV, Samuelson D, Patrick-Lake B, Clay I. Evaluation, acceptance, and qualification of digital measures: from proof of concept to endpoint. *Digit Biomark*. 2021;5(1):53–64.
- 14 Rawtaer I, Mahendran R, Kua EH, Tan HP, Tan HX, Lee TS, et al. Early Detection of Mild Cognitive Impairment With In-Home Sensors to Monitor Behavior Patterns in Community-Dwelling Senior Citizens in Singapore: Cross-Sectional Feasibility Study. *J Med Internet Res*. 2020 May;22(5):e16854.
- 15 Strobl MA, Lipsmeier F, Demenescu LR, Gossens C, Lindemann M, De Vos M. Look me in the eye: evaluating the accuracy of smartphone-based eye tracking for potential application in autism spectrum disorder research. *Biomed Eng Online*. 2019 May;18(1):51.
- 16 Jacobson NC, Weingarden H, Wilhelm S. Digital biomarkers of mood disorders and symptom change. *NPJ Digit Med*. 2019 Feb;2(1):3.
- 17 DiMe. *The Playbook: Digital Clinical Measures*. Available from: <https://playbook.dime-society.org/>.
- 18 Robin J, Harrison JE, Kaufman LD, Rudzicz F, Simpson W, Yancheva M. Evaluation of Speech-Based Digital Biomarkers: review and Recommendations. *Digit Biomark*. 2020 Oct;4(3):99–108.
- 19 Coravos A, Doerr M, Goldsack J, Manta C, Shervey M, Woods B, et al. Modernizing and designing evaluation frameworks for connected sensor technologies in medicine. *NPJ Digit Med*. 2020 Mar;3(1):37.
- 20 Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010 Mar;340(340):c332.
- 21 Breteler MJ, KleinJan EJ, Dohmen DA, Leenen LP, van Hillegersberg R, Ruurda JP, et al. Vital signs monitoring with wearable sensors in high-risk surgical patients: A clinical validation study. *Anesthesiology*. 2020 Mar;132(3):424–39.
- 22 Brasier N, Raichle CJ, Dörr M, Becke A, Nothdurft V, Weber S, et al. Detection of atrial fibrillation with a smartphone camera: first prospective, international, two-centre, clinical validation study (DETECT AF PRO). *Europace*. 2019 Jan;21(1):41–7.
- 23 Mahadevan N, Demanuele C, Zhang H, Volfson D, Ho B, Erb MK, et al. Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device. *NPJ Digit Med*. 2020 Jan;3(1):5.
- 24 Skender S, Schrotz-King P, Böhm J, Abbenhardt C, Gigic B, Chang-Claude J, et al. Repeat physical activity measurement by accelerometry among colorectal cancer patients—feasibility and minimal number of days of monitoring. *BMC Res Notes*. 2015 Jun;8(1):222.
- 25 HHS.gov. *Human subject regulations decision charts*. Available from: <https://www.hhs.gov/ohrp/regulations-and-policy/decision-charts>.
- 26 Mueller A, Hoefling HA, Muaremi A, Praestgaard J, Walsh LC, Bunte O, et al. Continuous Digital Monitoring of Walking Speed in Frail Elderly Patients: Noninterventonal Validation Study and Longitudinal Clinical Trial. *JMIR Mhealth Uhealth*. 2019 Nov;7(11):e15191.
- 27 ClinicalTrials.gov. Checklist for evaluating whether a clinical trial or study is an Applicable Clinical Trial (ACT) under 42 CFR 11.22(b) for clinical trials initiated on or after January 18, 2017. [cited 2020 Nov 23]. Available from: https://prsinformo.clinicaltrials.gov/ACT_Checklist.pdf.
- 28 Block VJ, Lizée A, Crabtree-Hartman E, Bevan CJ, Graves JS, Bove R, et al. Continuous daily assessment of multiple sclerosis disability using remote step count monitoring. *J Neurol*. 2017 Feb;264(2):316–26.
- 29 Ihlen EA, Weiss A, Helbostad JL, Hausdorff JM. The discriminant value of phase-dependent local dynamic stability of daily life walking in older adult community-dwelling fallers and nonfallers. *BioMed Res Int*. 2015;2015:402596.
- 30 Alharbi M, Bauman A, Neubeck L, Gallagher R. Validation of Fitbit-Flex as a measure of free-living physical activity in a community-based phase III cardiac rehabilitation population. *Eur J Prev Cardiol*. 2016 Sep;23(14):1476–85.
- 31 Del Din S, Godfrey A, Galna B, Lord S, Rochester L. Free-living gait characteristics in ageing and Parkinson's disease: impact of environment and ambulatory bout length. *J Neuroeng Rehabil*. 2016 May;13(1):46.
- 32 Dagenais M, Salbach NM, Brooks D, O'Brien KK. Assessing the measurement properties of the Fitbit Zip® among adults living with HIV. *J Phys Act Health*. 2020 Mar;17(3):293–305.
- 33 Appelboom G, Taylor BE, Bruce E, Bassile CC, Malakidis C, Yang A, et al. Mobile Phone-connected wearable motion sensors to assess postoperative mobilization. *JMIR Mhealth Uhealth*. 2015 Jul;3(3):e78.
- 34 Hellmers S, Izadpanah B, Dasenbrock L, Diekmann R, Bauer JM, Hein A, et al. Towards an automated unsupervised mobility assessment for older people based on inertial TUG measurements. *Sensors (Basel)*. 2018 Oct;18(10):3310.
- 35 Bent B, Goldstein BA, Kibbe WA, Dunn JP. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ Digit Med*. 2020 Feb;3(1):18.
- 36 Izmailova E, Bloofield D, Homsy J, Liu Q, Wood W, Zipunnikov V, et al. *Remote cardiac monitoring for clinical trials*. Remote Digital Monitoring Workshop, February 18–19, 2020. [cited 2017 Feb 9]. Available from: https://fnih.org/sites/default/files/final/pdf/CS2_Cardiac%20Monitoringv2.pdf.
- 37 Pham MH, Elshehabi M, Haertner L, Del Din S, Srujijes K, Heger T, et al. Validation of a step detection algorithm during straight walking and turning in patients with Parkinson's disease and older adults using an inertial measurement unit at the lower back. *Front Neurol*. 2017 Sep;8:457.
- 38 Schlachetzki JC, Barth J, Marxreiter F, Gossler J, Kohl Z, Reinfelder S, et al. Wearable sensors objectively measure gait parameters in Parkinson's disease. *PLoS One*. 2017 Oct;12(10):e0183989.
- 39 Cheresnev R, Kertész-Farkas A. HuGaDB: Human Gait Database for Activity Recognition from Wearable Inertial Sensor Networks. In: van der Aalst W, et al., editors. *Analysis of images, social networks and texts*. AIST 2017. Lecture Notes in Computer Science. Volume 10716. Cham: Springer; 2018. Available from: https://doi.org/10.1007/978-3-319-73013-4_12.
- 40 Moreau A, Anderer P, Ross M, Cerny A, Almazan TH, Peterson B. Detection of nocturnal scratching movements in patients with atopic dermatitis using accelerometers and recurrent neural networks. *IEEE J Biomed Health Inform*. 2018 Jul;22(4):1011–8.
- 41 Kakarmath S, Esteve A, Arnaout R, Harvey H, Kumar S, Muse E, et al. Best practices for authors of healthcare-related artificial intelligence manuscripts. *NPJ Digit Med*. 2020 Oct;3(1):134.
- 42 Christakis Y, Mahadevan N, Patel S. SleepPy: A python package for sleep analysis from accelerometer data. *J Open Source Softw*. 2019;4(44):1663.
- 43 Christakis Y. *SleepPy*. GitHub. [cited 2020 Nov 23]. Available from: <https://github.com/elyiorgos/sleeppy>.
- 44 Mahadevan N. *Analyze-tremor-bradykinesia-PD*. GitHub. [cited 2020 Nov 23]. Available from: <https://github.com/NikhilMahadevan/analyze-tremor-bradykinesia-PD>.
- 45 Shcherbina A, Mattsson CM, Waggott D, Salisbury H, Christle JW, Hastie T, et al. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *J Pers Med*. 2017 May;7(2):3.

- 46 Breteler MJ, Huizinga E, van Loon K, Leenen LP, Dohmen DA, Kalkman CJ, et al. Reliability of wireless monitoring using a wearable patch sensor in high-risk surgical patients at a step-down unit in the Netherlands: a clinical validation study. *BMJ Open*. 2018 Feb;8(2):e020162.
- 47 Merbler AM, Byiers BJ, Garcia JJ, Feyma TJ, Symons FJ. The feasibility of using actigraphy to characterize sleep in Rett syndrome. *J Neurodev Disord*. 2018 Feb;10(1):8.
- 48 Manta C, Patrick-Lake B, Goldsack JC. Digital measures that matter to patients: A framework to guide the selection and development of digital measures of health. *Digit Biomark*. 2020 Sep;4(3):69–77.
- 49 US Food and Drug Administration. *Human factors and medical devices*. 2018 [cited 2020 Nov 23]. Available from: <https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/human-factors-and-medical-devices>.
- 50 Patel SR, Weng J, Rueschman M, Dudley KA, Loreda JS, Mossavar-Rahmani Y, et al. Reproducibility of a standardized actigraphy scoring algorithm for sleep in a US Hispanic/Latino population. *Sleep (Basel)*. 2015 Sep;38(9):1497–503.
- 51 Doherty A, Jackson D, Hammerla N, Plötz T, Olivier P, Granat MH, et al. Large scale population assessment of physical activity using wrist worn accelerometers: The UK Biobank Study. *PLoS One*. 2017 Feb;12(2):e0169649.
- 52 Ladha C, Jackson D, Ladha K, Nappey T, Olivier P. *Shaker table validation of OpenMovement AX3 accelerometer*. 2013. doi: <https://doi.org/10.13140/RG.2.2.31312.51208>.
- 53 Hernando D, Garatachea N, Almeida R, Casajús JA, Bailón R. Validation of Heart Rate Monitor Polar RS800 for Heart Rate Variability Analysis During Exercise. *J Strength Cond Res*. 2018 Mar;32(3):716–25.
- 54 Murray CS, Rees JL. Are subjective accounts of itch to be relied on? The lack of relation between visual analogue itch scores and actigraphic measures of scratch. *Acta Derm Venereol*. 2011 Jan;91(1):18–23.
- 55 Clinical Trials Transformation Initiative. Use case for developing novel endpoints generated using mobile technology: Duchenne muscular dystrophy. 2020 [cited 2020 Nov 23]. Available from: <https://www.ctti-clinicaltrials.org/files/usecase-duchenne.pdf>.
- 56 Czech M, Demanuele C, Erb MK, Ramos V, Zhang H, Ho B, et al. The Impact of reducing the number of wearable devices on measuring gait in Parkinson disease: Noninterventonal exploratory study. *JMIR Rehabil Assist Technol*. 2020 Oct;7(2):e17986.
- 57 Czech MD, Psaltos D, Zhang H, Adamusiak T, Calicchio M, Kelekar A, et al. Age and environment-related differences in gait in healthy adults using wearables. *NPJ Digit Med*. 2020 Sep;3(1):127.
- 58 Giles D, Draper N, Neil W. Validity of the Polar V800 heart rate monitor to measure RR intervals at rest. *Eur J Appl Physiol*. 2016 Mar;116(3):563–71.
- 59 McGinnis RS, Mahadevan N, Moon Y, Seagers K, Sheth N, Wright JA Jr, et al. A machine learning approach for gait speed estimation using skin-mounted wearable sensors: from healthy controls to individuals with multiple sclerosis. *PLoS One*. 2017 Jun;12(6):e0178366.
- 60 Deka P, Pozehl B, Norman JF, Khazanchi D. Feasibility of using the Fitbit® Charge HR in validating self-reported exercise diaries in a community setting in patients with heart failure. *Eur J Cardiovasc Nurs*. 2018 Oct;17(7):605–11.
- 61 Izmailova ES, McLean IL, Bhatia G, Hather G, Cantor M, Merberg D, et al. Evaluation of wearable digital devices in a phase I clinical trial. *Clin Transl Sci*. 2019 May;12(3):247–56.
- 62 Perry B, Geoghegan C, Lin L, McGuire FH, Nido V, Grabert B, et al. Patient preferences for using mobile technologies in clinical trials. *Contemp Clin Trials Commun*. 2019 Jun 20;15:100399.
- 63 Clinical Trials Transformation Initiative. CTTI recommendations: Advancing the use of mobile technologies for 2 data capture & improved clinical trials. 2020 [cited 2020 Nov 24]. Available from: <https://www.ctti-clinicaltrials.org/sites/www.ctti-clinicaltrials.org/files/mobile-devices-recommendations.pdf>.
- 64 Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, et al; Apple Heart Study Investigators. Large-scale assessment of a smartwatch to identify atrial fibrillation. *N Engl J Med*. 2019 Nov;381(20):1909–17.
- 65 Kupczyk M, Hofman A, Kołtowski Ł, Kuna P, Łukaszyk M, Buczyłko K, et al. Home self-monitoring in patients with asthma using a mobile spirometry system. *J Asthma*. 2020 Jan;1–7.
- 66 Meritam P, Ryvlin P, Beniczky S. User-based evaluation of applicability and usability of a wearable accelerometer device for detecting bilateral tonic-clonic seizures: A field study. *Epilepsia*. 2018 Jun;59 Suppl 1:48–52.
- 67 Franciosi M, Pellegrini F, De Berardis G, Belfiglio M, Cavaliere D, Di Nardo B, et al; QuED Study Group. The impact of blood glucose self-monitoring on metabolic control and quality of life in type 2 diabetic patients: an urgent need for better educational strategies. *Diabetes Care*. 2001 Nov;24(11):1870–7.
- 68 O’Kane MJ, Bunting B, Copeland M, Coates VE; ESMON Study Group. Efficacy of self monitoring of blood glucose in patients with newly diagnosed type 2 diabetes (ESMON study): randomised controlled trial. *BMJ*. 2008 May;336(7654):1174–7.
- 69 HHS.gov. Exempt research and research that may undergo expedited review. 2003 [cited 2021 Mar 4]. Available from: <https://www.hhs.gov/ohrp/regulations-and-policy/guidance/exempt-research-and-research-expedited-review/index.html>.
- 70 Matosin N, Frank E, Engel M, Lum JS, Newell KA. Negativity towards negative results: a discussion of the disconnect between scientific worth and scientific culture. *Dis Model Mech*. 2014 Feb;7(2):171–3.
- 71 Gupta A, Stewart T, Bhulani N, Dong Y, Rahimi Z, Crane K, et al. Feasibility of wearable physical activity monitors in patients with cancer. *JCO Clin Cancer Inform*. 2018 Dec;2(2):1–10.
- 72 Wendel N, Macpherson CE, Webber K, Hendron K, DeAngelis T, Colon-Semenza C, et al. Accuracy of activity trackers in Parkinson disease: Should we prescribe them? *Phys Ther*. 2018 Aug;98(8):705–14.
- 73 Moon Y, McGinnis RS, Seagers K, Motl RW, Sheth N, Wright JA Jr, et al. Monitoring gait in multiple sclerosis with novel wearable motion sensors. *PLoS One*. 2017 Feb;12(2):e0171346.
- 74 Regalia G, Onorati F, Lai M, Caborni C, Picard RW. Multimodal wrist-worn devices for seizure detection and advancing research: focus on the Empatica wristbands. *Epilepsy Res*. 2019 Jul;153:79–82.
- 75 O’Brien E, Atkins N, Stergiou G, Karpettas N, Parati G, Asmar R, et al; Working Group on Blood Pressure Monitoring of the European Society of Hypertension. European Society of Hypertension International Protocol revision 2010 for the validation of blood pressure measuring devices in adults. *Blood Press Monit*. 2010 Feb;15(1):23–38.
- 76 Manta C, Jain SS, Coravos A, Mendelsohn D, Izmailova ES. An evaluation of biometric monitoring technologies for vital signs in the era of COVID-19. *Clin Transl Sci*. 2020 Nov;13(6):1034–44.
- 77 PRISMA. *PRISMA endorses*. [cited 2020 Nov 23]. Available from: <http://prisma-statement.org/Endorsement/PRISMAEndorses>.
- 78 CONSORT. *Impact of CONSORT*. 2010 [cited 2021 Mar 4]. Available from: <http://www.consort-statement.org/about-consort/impact-of-consort>.
- 79 Plint AC, Moher D, Morrison A, Schulz K, Altman DG, Hill C, et al. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Med J Aust*. 2006 Sep;185(5):263–7.
- 80 Korevaar DA, Wang J, van Enst WA, Leeflang MM, Hooft L, Smidt N, et al. Reporting diagnostic accuracy studies: some improvements after 10 years of STARD. *Radiology*. 2015 Mar;274(3):781–9.
- 81 Korevaar DA, van Enst WA, Spijker R, Bossuyt PM, Hooft L. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. *Evid Based Med*. 2014 Apr;19(2):47–54.
- 82 Frasch MG, Shen C, Wu HT, Mueller A, Neuhaus E, Bernier RA, et al. Can a composite heart rate variability biomarker shed new insights about autism spectrum disorder in school-aged children? 2019. [cited 2021 Mar 4]. Available from: <https://arxiv.org/abs/1808.08306>.
- 83 Pavic M, Klaas V, Theile G, Kraft J, Tröster G, Guckenberger M. Feasibility and usability aspects of continuous remote monitoring of health status in palliative cancer patients using wearables. *Oncology*. 2020;98(6 Suppl 6):386–95.