VALUE IN HEALTH 21 (2018) 631–639

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Selection of and Evidentiary Considerations for Wearable Devices and Their Measurements for Use in Regulatory Decision Making: Recommendations from the ePRO Consortium

Bill Byrom, PhD[1,], Chris Watson, PhD[2], Helen Doll, DPhil[3], Stephen Joel Coons, PhD[4], Sonya Eremenco, MA[4], Rachel Ballinger, PhD[3], Marie Mc Carthy, MBA[5], Mabel Crescioni, DrPh[4], Paul O'Donohoe, MSc[6], Cindy Howry, MS[7], on behalf of the ePRO Consortium*

[1]*ICON Clinical Research, Marlow, Buckinghamshire, UK;* [2]*ERT, Nottingham, Nottinghamshire, UK;* [3]*ICON Clinical Research, Abingdon, Oxfordshire, UK;* [4]*Critical Path Institute, Tucson, AZ, USA;* [5]*ICON Clinical Research, Dublin, Ireland;* [6]*CRF Health, London, UK;* [7]*assisTek, Scottsdale, AZ, USA*

ABSTRACT

Background: Wearable devices offer huge potential to collect rich sources of data to provide insights into the effects of treatment interventions. Despite this, at the time of writing this report, limited regulatory guidance on the use of wearables in clinical trial programs has been published. **Objectives:** To present recommendations from the Critical Path Institute's Electronic Patient-Reported Outcome Consortium regarding the selection and evaluation of wearable devices and their measurements for use in regulatory trials and to support labeling claims. **Methods:** The evaluation group was composed of Critical Path Institute's clinical outcome assessment (COA) scientists and COA specialists from pharmaceutical trial eCOA solution providers, including COA development and validation specialists. The resulting recommendations were drawn from a broad range of backgrounds, perspectives, and expertise that enriched the development of this report. Recommendations were developed through analysis of existing regulatory guidance relating to COA development and use in clinical trials, medical device certification/clearance regulations, literature-reported best practice, and practical experience of wearable technology application in clinical trials. **Results:** We identify the essential properties of fit-for-purpose wearables and propose evidence needed to support their use. In addition, we overview the activities required to establish clinical endpoints derived from wearables data. **Conclusions:** Using this framework, we believe there is enough current understanding to promote the appropriate use of wearables in study protocols. We hope this will provide a basis for discussion among clinical trial stakeholders and catalyze the development of more robust regulatory guidance.

Keywords: clinical outcomes, clinical trial endpoints, clinical trials, performance outcomes, remote monitoring, validation, wearables.

Introduction

Miniaturization of sensors and circuitry has given rise to huge proliferation in the development and commercialization of wearables and sensors with application to health and wellness. Examples are wide-ranging and include patches for electrocardiogram monitoring, wrist-worn devices for sleep assessment, and sensors with subcutaneous probes for continuous glucose monitoring. Activity monitors with their associated mobile applications and software are increasingly popular among those wishing to improve fitness or manage weight through regular exercise regimens.

Responding to this growing market of novel and interesting wearables and sensors, the biopharmaceutical industry is actively interested in knowing how to harness these devices to enable greater information to be learnt about the effects of treatments during drug development. Despite the promise of this technology, there is uncertainty regarding the regulatory acceptability of data collected in this way—specifically in understanding what evidence should be available and considered when selecting an appropriate device for use in a clinical trial to ensure adequate precision, accuracy, and reliability of data collected and the nature of evidence required to demonstrate appropriateness and clinical relevance of new endpoints derived from the data.

The purpose of this report was to propose a set of recommendations for the biopharmaceutical industry in relation to the selection of and evidentiary considerations for wearable devices and sensors and their outcome measures for use in regulatory clinical trials. These recommendations are based on the current literature, regulatory guidance available to date, and expert consensus of the member firm representatives of the Electronic Patient-Reported Outcome Consortium, a technology industry research group with the Critical Path Institute as its managing member.

* *Address correspondence to:* Bill Byrom, ICON Clinical Research, 3rd Floor Marlow International, Parkway, Marlow, Buckinghamshire, SL7 1YL, UK.

E-mail: bill.byrom@iconplc.com

1098-3015$36.00 – see front matter

http://dx.doi.org/10.1016/j.jval.2017.09.012

Table 1 – Categories of microsensors used in health assessment.

Category	Description
External devices/external sensors	An apparatus containing one or more sensors that is either physically separate from the user or that the user interacts with periodically during a specific user operation but is not worn, implanted, or ingested. An example of a physically separate sensor might be a depth (3D) camera installed at a home location to detect movement or falls in frail or disabled patients. Examples of external sensors that a user interacts with during a specific user operation include an electronic weighing scale or a digital spirometer.
Wearable devices/wearable sensors	A small electronic device containing one or more sensors that are integrated into clothing or other accessories that can be worn on the body [2], such as on a wristband, belt, headband, adhesive patch, contact lens, or glasses. Common sensors used in wearable devices include those for measuring movement and position, such as accelerometers, gyroscopes, magnetometers, and global positioning systems, or sensors for assessing electrophysiological and chemophysiological function or other physiological properties such as body temperature [3]. Examples of wearable sensors and devices in medicine and health care include wearables for body temperature measurement, respiration monitors, heart rate monitors, devices measuring electrocardiogram or electroencephalogram, pulse oximeters, blood pressure monitors, pH monitors, continuous glucose monitors (although these contain subcutaneous components, these are minimally intrusive and we include them in this category), and galvanic skin response detectors [4].
Implantable devices/ implantable sensors	Devices that are inserted inside the human body. Examples include cardiac arrhythmia monitors and brain liquid pressure sensors [4].
Ingestible devices/ingestible sensors	Sensors that are swallowed by the user and serve to make recordings or signals on the basis of physiological stimuli. Examples include core temperature sensors, ingestible tags included in medication that emit a signal when detecting stomach acid to identify medication adherence (see [5], for example), and capsule endoscopy to visualize the esophagus, small bowel, and colon with a small, disposable capsule containing image capture sensors [6].

Before examining and making recommendations regarding device selection and evidence necessary to establish clinical trial endpoints arising from their use, it is helpful to define some of the terminology surrounding wearables and sensors and the different types of outcome measures and data they can produce. Definitions based on a review of the literature and consortium consensus opinion are detailed here.

In the broadest terms, a *sensor* is a device or device component that detects and measures physical or chemical information from a surrounding physical environment and translates this into an electrical output signal. Example measurement parameters may include light, heat, motion, moisture, pressure, chemical content, or other environmental properties. The electrical output signal is generally stored, displayed on a device or hardware, or transmitted over a wireless network such as wireless Internet, 3G, 4G, Bluetooth, or near-field communication for reading, storage, and further processing.

Microsensors are miniature sensors that have electrical and mechanical operation components, and are also termed microelectromechanical systems. These are usually produced by integrated circuit manufacturing from silicon or similar materials.

The use of reliable, high-performance microsensors in the medical field is of growing importance for patient health monitoring [1], personal wellness, and clinical research. We group sensors into four categories: external, wearable, implantable, and ingestible devices and sensors. These are defined in Table 1. Sensors, and software collecting sensor data, may collect data passively (passive data) or while the patient is performing a prescribed activity (active data).

Literature definitions of clinical outcomes, outcome assessments, endpoints, biomarkers, and performance outcomes are presented in Table 2.

Since the original definition of a performance outcome (PerfO) by the Food and Drug Administration (FDA) [10], there has been a growing trend to use wearables and sensors as a means of instrumenting performance tests to collect potentially more accurate and more informative outcomes data. Examples include using accelerometers and gyroscopes to measure outcomes in a timed-up-and-go test or a 6-minute walk test, in which the use of wearables can enable the measurement of outcomes in addition to the time taken to complete the test—such as duration of resting time and number of steps taken to complete a 180-degree turn. Given the increasing utility of wearable devices for remote monitoring and measurement, we recognize the growing ability to collect PerfO data on the basis of the conduct of an instructed performance task that the patient is requested to complete in a nonsupervised setting, such as within his or her own home (e.g., measuring the number of steps or cadence during a short walking task). Consequently, we recommend extension of the FDA's definition of a PerfO as follows: *A PerfO assessment is a measurement based on a specific task or tasks performed by a patient according to instructions provided by a qualified test administrator in either a supervised or an unsupervised setting. These include, for example, measures of gait speed (e.g., timed 25-foot walk test), measures of frailty and fall risk on the basis of a timed-up-and-go test, memory recall, or other cognitive testing (e.g., digit symbol substitution test). Performance outcomes may be assessed directly by a qualified test administrator, or instrumented using sensors and/or wearable devices such as an accelerometer to measure the time and steps taken to complete a stair-climbing test.*

Wearable technology and sensors may be classified as medical devices, and companion applications displaying or transmitting data from these devices may, in some circumstances, be considered mobile medical applications. It is useful to understand this regulatory context, although as you will read later in this article, we do not consider the clearance or certification of a wearable device a prerequisite for its use in clinical trials.

The main classification of medical devices and their associated requirements for commercial deployment in the consumer marketplace are defined by the FDA, the European Union (EU), and Health Canada, with rest-of-world territories generally accepting the EU requirements and certification as a basis for registration.

The FDA classifies medical devices as class I (e.g., elastic bandages), class II (e.g., acupuncture needles), or class III (e.g., implantable pacemakers), depending on the level of risk

Table 2 – Definitions of terms used in health assessment.

Category	Description
Clinical outcome	Measurable characteristics influenced by an individual's baseline state or an intervention [7]. Examples of clinical outcomes from sensor data include estimates of sleep patterns and quality using a wrist-worn sleep and activity monitor, or free-living activity measurements using an accelerometer.
Outcome assessment	"The measuring instrument that provides a rating or score that is intended to represent some aspect of the patient's health status" [8]. Outcome assessments are used to define efficacy endpoints when developing a therapy for a disease or condition.
Clinical endpoint	"A characteristic or variable that reflects how a patient feels, functions, or survives" [9]. Endpoint descriptions include information defining how and when they are measured, how they are calculated, rules for missing data, and how they are analyzed.
Biomarker	"A defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions. Molecular, histologic, radiographic, or physiologic characteristics are types of biomarkers. A biomarker is not an assessment of how an individual feels, functions, or survives" [7]. Although biomarkers may be surrogates for clinical outcomes, they are not clinical outcomes themselves. The term digital biomarker has been used to define a biomarker that is collected digitally using a sensor. Some sensor data may be considered to generate biomarkers, for example, digital diagnostic biomarkers might be derived from wearable heart rate monitors or continuous glucose monitors.
Performance outcome assessment	"A measurement based on a task(s) performed by a patient according to instructions that is administered by a health care professional. Performance outcomes require patient cooperation and motivation. These include measures of gait speed (e.g., timed 25 foot walk test), memory recall, or other cognitive testing (e.g., digit symbol substitution test)" [10].

associated with their use [11]. The usual route to market in the United States is via the 510(k) market clearance. In the EU, *Conformité Européene* marking is awarded to certified medical devices, which are classified as class I, IIa, IIb, or III and which follow similar approval requirements as in the United States [12]. In Canada, devices of classes II to IV require a medical device establishment license (MDEL) application [13].

Medical device classification/clearance includes standards in manufacturing, quality systems and change control process, and a statement of the "indications for use" that describes the usage for which the device has been deemed appropriate and may define certain populations, clinical settings, and usage parameters (e.g., body location for the wearable). To obtain such clearance/certification, safety data (including electrical safety) and data showing substantial equivalence to a similar marketed device (a predicate) will need to be demonstrated through provision of scientific data supporting the specific indication for use. When no predicate exists to enable concurrent construct validity to be demonstrated, further evidence is also required to obtain device clearance/certification.

Although useful, wearable device market clearance/certification is not a requirement for use in clinical trials, in particular because within a clinical trial the device is being used within a controlled group of subjects with clinical oversight. Nevertheless, devices without an MDEL may still need to be declared within the regulatory documentation for clinical trial approval, and responses to basic safety questions may need to be provided as part of this process (e.g., in Canada). Although market clearance/certification is a desirable attribute, this should not be considered a requirement for device selection for use in clinical trials, as long as the evidentiary considerations described later are satisfied.

When companion mobile applications connect with wearable devices to interface with the data collected, they may be subject to the regulations around mobile medical applications. In general, however, when an application is designed to simply report data (e.g., a medical device data system) or provide basic coaching/behavior change, it is generally under the radar (enforcement discretion) in terms of the FDA guidance on mobile medical applications and medical devices [14]. Nevertheless, independent of this, a research exemption applies for use of mobile applications in US clinical trials, and the same is true in the EU. Although Health Canada also accepts that an application that acts as a medical device data system does not need regulating, if the application itself has additional functionality when it acts upon the data being collected, then although this is still covered by the FDA and EU exemptions, Health Canada offers no such exemptions, and vendors and sponsors will be required to provide supporting information necessary to enable Health Canada to approve its use in a clinical research study. In general, however, applications used to display and transmit data from wearables and sensors are unlikely to fall within the remit of these regulations.

Evidence Needed to Support Identification of a Suitable Device

Endpoint Model

As with evaluating the suitability of a patient-reported outcome (PRO) measure to assess study outcomes, an endpoint model should be defined [15]. This model will show each study concept of interest, and the endpoints relevant to each concept, which the device will be required to measure. For example, if the concept of interest is improved physical function, the proposed endpoint could be the mean number of steps per day during week 1 versus week 12 in a 12-week treatment trial. The adequacy of the device as a measurement approach will depend on its role and relationships with other clinical trial endpoints as depicted in the endpoint model. The placement or role in the endpoint hierarchy (e.g., whether primary, secondary, or exploratory) should also be specified so that appropriate statistical methods can be planned.

If we select a wearable device to measure this concept of interest, we should ensure that the device measures the concept of interest faithfully. Identifying a device that is fit for purpose in measuring the identified concepts of interest within a clinical trial or drug development program requires the consideration of three factors: 1) Is the wearable device or sensor safe to use? 2) Are the device and vendor suitable for the trial objectives and patient population studied? and 3) Is there satisfactory evidence of data validity and reliability to confirm that the device provides

the required level of measurement accuracy and precision in measuring the concept of interest?

Safety

Determination that a device is safe for use by patients requires the manufacturer to provide evidence of testing in a number of areas including, as applicable, mechanical, electrical, and biological engineering performance, such as fatigue, wear, tensile strength, and compression; electrical safety and electromagnetic compatibility; sterility; and stability/shelf-life. It would be expected that sufficient data will be available from the device manufacturer on request. For example, if the device is to be worn in contact with the skin, the materials and metals used should be hypoallergenic and fit for purpose, and shown to not result in adverse effects such as skin abrasion or tissue inflammation when worn for the periods of time consistent with the study objectives.

In addition, proven methodologies for use in patients should be available. This should include usage instructions, including maximum wear intervals and wear locations, and instructions for the safe preparation and re-use (if appropriate) of the wearable device, such as processes for the sterile cleaning of a device before and after use.

Suitability

A number of factors are important in determining the suitability of a wearable for use in a regulatory clinical trial (Fig. 1). Device selection will be influenced by the study design. For example, the choice of a wearable device to measure heart rate may be influenced by the required measurement period because some devices may be inconvenient if worn for longer periods or may have insufficient battery or storage capacity. The choice of an activity monitor may be influenced by the study objectives; for example, devices supporting different wear locations may be more suitable for study of sedentary behavior in which posture determination is important, as opposed to the study of free-living physical activity [16].

Patient population considerations include assessing whether a device will be acceptable in that patient group. Acceptability may be influenced by how easy it is to use, where it needs to be worn and for how long, what it looks like, and aspects of its design and form factor. Certain patient populations, for example, may find the wrist straps or belts provided by some devices too short (e.g., obese patients) or too long (e.g., gaunt older adults) to wear comfortably, and young people may be unwilling to wear a visible device if they feel it is unfashionable or may draw unwanted attention to them. Ease of use may also include how the patient operates the device, whether they need to remove and replace the device, and whether they need to charge and maintain the device in any way. All these considerations form part of the usability and acceptability profile of the device in the target patient population. It is our recommendation that for many devices it is not essential to perform specific usability studies in the target population, but published studies in similar groups of patients or early use of the device in phase II may provide reassurance when acceptability, usability, and burden are considered potential barriers. In specific cases in which operation and use are considered complicated, assessing usability and training information developed for patients using a cognitive interview and usability study in a small number of patients (typically 6–10) may be recommended (see [17] for details on the conduct of cognitive interview studies).

If patients are required to wear or use a device for a specific number of days to obtain reliable estimates of the endpoint of interest, then the device should have sufficient battery length and data storage capacity (if appropriate) to support this duration. When devices are removed for periods of time, charging periods and/or data transmission may enable longer observation intervals when battery life does not cover the entire time of use. Some devices may require setup activities by site and patients, and these activities may impact site or patient burden. In addition, some devices may be re-used by other patients as the study progresses, and in this case cleaning and resetting processes should be considered. Manufacturers should be able to provide guides as to the suitable sterilizations of the devices, or state whether the devices are single-use devices only.

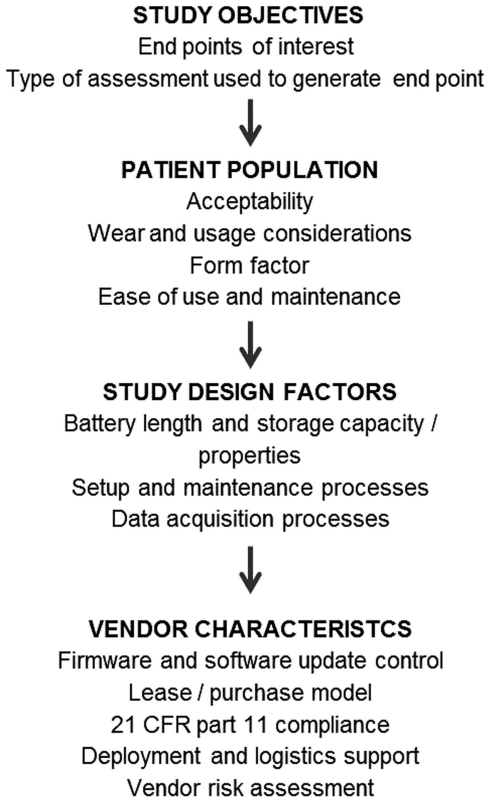

Fig. 1 – Factors influencing device suitability. CFR, Code of Federal Regulations.

It is also important to determine whether patients should have visibility of their data. Some devices display measurement output, for example, heart rate, glucose levels, or total steps taken. In a controlled clinical trial it may be important to select a device that has the ability to blind these data from the patient so that the device itself does not bias results or add to potential placebo effects by inadvertently becoming a component of the intervention studied (e.g., target setting). Some patients may expect to see their wearables data, particularly if they have had experience using consumer wearables, but in general the data from a wearable device should be shared with patients only after study completion. Exceptions to this do, however, exist especially when measurements are used to aid safety evaluation or self-management (e.g., blood glucose meter readings in diabetes). The guiding principle should always be whether it is anticipated that access to the data may affect the behavior of the patient and add bias or increase placebo effect. When wearables data may reveal information about the treatment group allocated, further steps should be taken to double-blind these data.

Data acquisition processes are a consideration if the study design requires remote access to the data collected for ongoing review by the investigator or safety monitoring committee. In

this case, devices may need to support data transmission from the patient's home setting, such as via direct wireless connectivity, Bluetooth connectivity to a mobile device or hub, or Web download through a connected charging base. Devices and vendor clouds may also need to be able to interoperate with integration middleware solutions to provide data in other eClinical systems and data stores as required. The amount of data that the device can store when connectivity is not available may be an additional consideration to avoid loss of data in situations in which connectivity is problematic. Device data storage, for example, should be nonvolatile (not lost when the device is switched off) and should not be overwritten with new data when full.

Finally, evaluation of vendor characteristics is an additional important consideration. Access and control of source data is of vital importance and can be an area that differentiates research devices from consumer devices. For example, it is vital that the data collected and stored on the device or within a vendor cloud cannot be changed or modified from its original form. Consumer devices may not enable control over the implementation of new firmware updates on the device or software updates within the vendor cloud. Importantly, such updates may apply revisions to algorithms that are used to derive outcome measures reported by the device. These changes, if implemented without control during a study, may affect the integrity of the data and the ability to compare data collected before and after the update.

A further consideration is risk assessment of the vendor, with the objective of ensuring access to the data collected on the device for the course of the clinical trial or program. The device industry is rapidly changing, which may mean that some devices available now may not exist in the future. Although this is hard to predict, a risk assessment should take particular interest in access to the data collected using the vendor solution over the period of use and, when appropriate, seek to understand how data can be provided in the event of a change of the vendor or the withdrawal of a device.

As for all technology systems used in clinical trials, vendor-supplied software to manage, initialize, and review device data, when used, should adhere to appropriate regulations regarding data security, traceability, data protection, and title 21 of the Code of Federal Regulations part 11 compliance [18].

Other considerations that may affect device selection include whether the vendor offers both purchase and lease business models, because these can impact costs for studies of different sizes and lengths, and whether the vendor also offers global device deployment and logistics management and support if needed.

Evidence Supporting the Validity and Reliability of Data Generated

Evidence that the device is generating sufficiently valid and reliable data is an essential element in the selection of a device for use in clinical trials. It is important to be able to demonstrate that the device is providing measurement to the level of accuracy and precision appropriate to the use of the data.

Although the outcomes data collected using wearable technology are objective rather than subjective, FDA's guidance on PRO measures [15] provides a context for consideration of the evidentiary expectations of wearable devices as an "instrument" to collect outcomes data.

On the basis of this, we consider the following to be the minimum evidence required to demonstrate that a device has suitable measurement properties to measure the concept of interest in the target patient population and is suitable for use in clinical trials aimed at regulatory approval of drugs and other medical products:

1. *Content validity*: Establishing that the device provides a sufficiently comprehensive assessment of a concept of interest that is meaningful to patients;
2. *Reliability assessment*: Intradevice and interdevice agreement, including calibration methods when appropriate;
3. *Concurrent (criterion-related) validity*: Assessment of measurement accuracy and concordance with an alternative accepted approach, and when appropriate sensitivity and specificity in measurement;
4. Ability to detect change.

The aforementioned evidence should be provided on the basis of data collected in subjects representative of the target population to be studied and using devices and protocols representative of the intended use. Nevertheless, it is considered not necessary to test a device on each specific population studied. Rather, we recommend that device use should be supported by evidence of its accuracy and reliability over the *range of measures expected* in the target population. For example, evidence of acceptability of a step counter would not need to be provided for specific populations unless aspects of gait specific to that population may be thought to affect measurement (e.g., Parkinson disease).

Content validity

Evidence must be generated or assembled that demonstrates that the device is measuring a meaningful aspect of the disease/condition or treatment from the patient's perspective. Qualitative research involving the target patient population or other reporters may be needed to establish the extent to which data from the device are appropriate and comprehensive relative to their intended measurement concept (i.e., concept of interest) (Table 3). For example, if the concept of interest is improved physical function and the device used is an activity monitor, then the measure derived from the activity monitor data should be regarded by the patients as both relevant to their condition (and fully reflective of it) and to an improvement in function. Patients with chronic obstructive pulmonary disease, for example, may not regard time spent in moderate to vigorous activity as relevant or important to them, but may view their ability to sustain bouts of continuous, purposeful ambulatory movement (as measured by bouts exceeding a defined cadence) as important and relevant to an improvement in function [16].

Intradevice and interdevice reliability

Reliability data may be provided by the device vendor or found in the published literature. In some circumstances, reliability data may be obtained through artificial or simulated laboratory testing (e.g., multi-axis shaking table for accelerometer testing [19]), but in all cases they should be supplemented with data from testing in human subjects in controlled conditions, with the use of an appropriate anchor to identify stability. As for other measures of outcome, intraclass correlation coefficients (ICCs) are the appropriate statistic for measuring agreement, both intradevice and interdevice. Acceptability of reliability would be indicated by ICC values being higher than a certain threshold, for example, the lower limit of the 95% confidence interval exceeding 0.7 [20]. In addition, to ensure reliability is maintained, device manufacturers must be able to demonstrate that devices are produced in adherence with a quality system to ensure equivalence of devices between batches and with the reliability data provided.

Concurrent (criterion-related) validity

This is evidence that the instrument correlates with another instrument or measure that is regarded as a more accurate

Table 3 – Recommended validation evidence required for the selection and implementation of wearable devices and sensors and their derived end points in clinical trials.

Measurement property	Type	What is assessed?	Evidence needed
		Validity of the device	
Content validity	Measuring a meaningful aspect of how patients feel, function, or survive as a result of treatment	Evidence to support the importance and relevance of the derived concept of interest, as measured by the device, to the patient, their condition, and its treatment	Concept elicitation from patients or other reporter input via qualitative research. This could be primary research or similar studies with evidence reported in peer-reviewed literature.
Reliability of outcomes data	Intradevice reliability*	Stability of measures over time when no change is expected	Laboratory and human study demonstrating test-retest reliability, as measured with an intraclass correlation coefficient and, when appropriate, sensitivity and specificity assessment. May be provided by manufacturer or reported in peer-reviewed publication(s).
	Interdevice reliability*	Agreement in measures between units/devices administered together	Evidence that devices are manufactured to appropriate quality processes to ensure continued reliability of measurement.
Validity of outcomes data	Concurrent (criterion-related) validity)*	Evidence that the device measures the concept of interest by comparison with a known measurement approach (e.g., gold standard) in subjects with similar characteristics to the intended patient population	At least one validation study of an appropriate size reported independently of the vendor and published in a peer-reviewed journal. For example, a 50-subject crossover study enabling comparison with an existing valid instrument. When possible, wearable and gold standard measures taken at the same time should be compared. Acceptable methods vary and include correlation, ROC analysis, diagnostic measures such as sensitivity and specificity, and Bland-Altman plots. Evidence provided should be based on studies using protocols, devices, and individuals representative of those to be studied.
Responsiveness of outcomes data	Ability to detect change	Evidence that a device outcome measure can identify differences in measurements over time in individuals or groups (similar to those in the clinical trials) who have changed with respect to the measurement concept	At least one controlled study involving an intervention that is understood to create a change in the measurement of interest.
Usability of device	Understanding of training instructions and ability to use appropriately (only for devices that are considered complicated to use for the patient population studied)	When usage is considered complicated for the specific patient group, a cognitive interview and usability study ensuring training materials are understood and the device can be used appropriately in the population of interest	When needed: cognitive interview report.
		Validity of clinical trial endpoints derived from wearables-collected outcomes data, for endpoints intended to support labeling claims	
Responsiveness	Ability to detect change	Evidence that a clinical endpoint derived from device outcomes data can identify differences in outcomes over time in individuals or groups (similar to those in the clinical trials) who have changed with respect to the measurement concept	At least one controlled study involving an intervention that is understood to create a change in the endpoint of interest. This may be conducted within an early nonconfirmatory study.

continued on next page

Table 3 – *continued*

Measurement property	Type	What is assessed?	Evidence needed
Interpretability	Responder definition	Evidence of clinical relevance and interpretability, in particular responder definition(s) in an appropriate patient population	This definition should be determined, *a priori*, in a population representative of those to be studied. At least one study involving an intervention that is understood to create a change in the endpoint of interest, including a number of known anchor endpoints that can identify whether a meaningful change has occurred. The responder definition is estimated from the change scores from the wearable device in those experiencing such a meaningful change. Typically, various methods are used, with triangulation of the results obtained. This may be conducted within an early nonconfirmatory study.

CE, *Conformité Européene*; MDEL, medical device establishment license; ROC, receiver operating characteristic curve.

* For devices with 510(k), CE, and/or MDEL, this evidence can be assumed as part of that clearance certification if the device is used to measure outcomes in line with the indications for use stated on the market clearance certificate.

criterion or "gold standard" measure. The report of at least one study of an appropriate size to demonstrate concurrent (criterion-related) validity performed and reported independently of the vendor, or published in scientific journals subject to independent peer review, is recommended. Depending on the outcomes measured, a typical study might include, for example, approximately 50 subjects and follow a crossover design enabling comparison with an existing gold standard approach via within-subject comparison. When possible, measurements using the gold standard and wearable device should be taken at the same time. Ideally, the primary analysis of equivalence should be determined by calculation of the ICC between the device and comparison instrument derived from an analysis of variance using a mixed-effects model with subject considered as a random effect and methodology as a fixed effect (see [21] for examples and formulae). Acceptability of concordance should be concluded for ICC values that are higher than a certain threshold, for example, the lower limit of the 95% confidence interval exceeding 0.7 [20]. Methodology for such equivalence studies has been described in detail elsewhere for PRO measures and can be used here [15,20,21]. When a device is used to predict a specific state, such as whether a subject is asleep or awake, assessment should include the measurement of the sensitivity and specificity of predictions (see [22], for example). Receiver operating characteristic curves with calculation of the area under the curve may also be useful in assessing the degree of agreement.

As for reliability assessment, we recommend that concurrent validation studies should be performed in a group of subjects with similar characteristics, such as their expected score distribution, to the target patient population intended to be studied, but it is not necessary to repeat such studies in each individual patient population.

When medical device certification/clearance has been granted (e.g., 510(k), *Conformité Européene*, or MDEL), the regulatory review associated with this certification may be sufficient to cover the concurrent construct validity and reliability evidence described earlier if the device is manufactured to a quality system and is used within the certified indications for use. Although many consumer devices are not intended to provide the level of accuracy and reliability of medical devices, those for which appropriate evidence exists, as described earlier, may be considered suitable for use, subject to the other safety and suitability considerations discussed earlier.

Ability to detect change

Outcome assessments provided by wearable devices or sensors should, when used in a clinical trial, be seen to be sensitive enough to detect change when a change exists. This is normally demonstrated by controlled studies involving an intervention that is understood to create a change in the outcome of interest. Ideally such studies should include additional measures to enable the identification of a true change in the outcome to be determined. It is recommended that this evidence be provided through at least one published study in a peer-reviewed journal.

Evidence Needed to Establish Clinical Trial Endpoints Derived from Wearable Device Data

After the selection of a suitable device, it is important to consider the need to provide evidence supporting the appropriateness and validity of the relevant trial endpoints derived from the outcomes data it provides. This is in common with the requirements associated with the use of any clinical trial endpoint used to support labeling claims, and this is not specific to endpoints derived from wearables data and has been well documented elsewhere [15]. In brief, evidence needed to support clinical trial endpoints derived from wearables data should address assessment of responsiveness (ability to detect change) and interpretability (understanding meaningful change) of the proposed endpoint (Table 3).

Meaningful change may be represented by the minimal important difference (or minimally clinically important difference) or the minimal individual change that distinguishes a responder from a nonresponder. The FDA in its final guidance on PRO measures [15] focuses only on the definition of a responder at the level of the individual patient.

Endpoints that are well understood, characterized, and interpretable may need little additional evidence to characterize meaningful change and clinical interpretation, such as minimum

daily oxygen saturation (SpO_2) derived from a pulse oximeter or peak daily heart rate derived from a wearable heart rate monitor. In other cases, additional evidence to demonstrate the clinical relevance of change on an outcome measure should be generated, using established approaches such as consensus-based, anchor-based, and distribution-based methods [23]. Anchor-based methods, when possible, are recommended and compare measures obtained to an anchor that is itself interpretable in having known relevance to patients [24], with distribution-based methods considered as providing supportive evidence [15,25]. Examples of estimation of meaningful change for endpoints derived from wearables include minimally clinically important difference estimation for the total daily steps measured using an accelerometer for patients with chronic obstructive pulmonary disease [26] and multiple sclerosis [27]. Estimation of responder definitions based on anchor definition typically uses receiver operating characteristic curves to determine the optimal cutoff point for the target measure to define a responder, on the basis of minimizing responder misclassification (see [28] and [29], for example).

Primary, Secondary, and Exploratory Endpoints

The evidence needed to support a device and its endpoint depends on the ultimate use of the endpoint. When endpoints are used in labeling claims, they should be included in the study protocol's endpoint hierarchy, detailed in the statistical plan, and significance tests adjusted for multiplicity when appropriate. In general, evidence for validity and reliability as described earlier should be provided for all primary and secondary endpoints intended for inclusion in product labeling. Although not essential, available evidence supporting the measurement properties of the wearable device used to measure exploratory endpoints should also be assembled. Early nonconfirmatory studies may provide an ideal opportunity to implement devices and collect data required for endpoint validation and usability in preparation for later confirmatory studies.

Discussion

At the time of writing, there is no published regulatory guidance specifically addressing the implementation of wearables in clinical trial protocols. In the making of recommendations on the selection and evaluation of wearables and their measurements, we have, however, drawn substantially from parallels with existing guidance for the use of PRO measures to support medical product labeling claims [15]. Although the data and measurement methodologies are different, commonalities have enabled our recommendations to be drawn.

Importantly, we have avoided making generalizations about the suitability of consumer devices compared with certified/cleared medical devices. It is our view that any wearable, and the endpoints derived from its data, has the potential to be considered appropriate for use if it adheres to a basic set of properties important to clinical trials (such as source data control, traceability, and security) and if evidence can be provided to support the reliability, validity, and interpretability of the data it generates. In some cases, much of this groundwork, in addition to early understanding of the potential of wearable-derived endpoints, can be assessed in early nonconfirmatory studies.

Although there remains a lack of specific guidance from regulatory bodies, we believe this work provides a robust framework for the adoption of wearables in regulatory trials. Although including wearable devices in clinical trial protocols may require some additional evidence gathering or generation, we believe the reward for this effort will be substantial. Wearable technology enables us to gather more information about treatment effects during our clinical development programs, and this provides greater insights and important evidence supporting the findings of other study endpoints. It also enables us to collect data that are perhaps more relevant to the patient, for example, free-living activity compared with in-clinic functional performance tests. As we develop more patient-centric trials, these kinds of measures may enable us to get closer to the measurement of the things that really matter to the patient, in addition to facilitating more remotely conducted studies.

Overriding all, however, is the principle that what we measure should support the study objectives. We do not endorse using wearables because we can, but rather we promote an approach whereby the endpoints of interest are determined from the study objectives, and these endpoints determine the suitability of a wearable device or sensor as an assessment approach.

Conclusions

There is a growing use of wearable technology in the personal health and wellness arena. This has been helpful in showing the potential offered by wearable sensors in the measurement of treatment intervention effects. Although there is more work required to demonstrate the suitability of specific wearable devices and establish endpoints derived from their data, there is enough understanding of how to do this to enable their inclusion in study protocols. This article provides a recommended framework to adopt when selecting and implementing wearable devices in clinical development programs, which we hope will, at least, provide a basis for discussion and become a catalyst for the development of robust regulatory guidance.

Acknowledgements

Critical Path Institute (C-Path) is supported, in part, by Critical Path Public-Private Partnerships Grant Number U18 FD005320 (effective 2015-2020) from the U.S. Food and Drug Administration. Financial support for C-Path's ePRO Consortium comes from membership fees paid by the ePRO Consortium's members (https://c-path.org/programs/epro/).

REFERENCES

[1] Tsoukalas D, Chatzandroulis S, Goustouridis D. Capacitive microsensors for biomedical applications. In: Webster JG, ed., Encyclopedia of Medical Devices and Instrumentation. Hoboken, NJ: John Wiley & Sons., Inc, 2006.
[2] Wright R, Keith L. Wearable technology: if the tech fits, wear it. J Electron Resour Med Libr 2014;11:204–16.
[3] Redmond SJ, Lovell NH, Yang GZ, et al. What does big data mean for wearable sensor systems? Contribution of the IMIA Wearable Sensors in Healthcare WG. Yearb Med Inform 2014;9:135–42.
[4] Al Ameen M, Lui J, Kwak K. Security and privacy issues in wireless sensor networks for healthcare applications. J Med Syst 2012;36:93–101.
[5] Belknap R, Weis S, Brookens A, et al. Feasibility of an ingestible sensor-based system for monitoring adherence to tuberculosis therapy. PLoS One 2013;8:e53373.
[6] Yuce MR, Dissanayake T. Easy-to-swallow wireless telemetry. IEEE Microw Mag 2012;13:90–101.
[7] US Food and Drug Administration-National Institutes of Health Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) resource. 2016. Available from: http://www.ncbi.nlm.nih.gov/books/NBK326791/toc/?report=reader. [Accessed May 15, 2017].
[8] Walton MK, Powers JH, Hobart J, et al; International Society for Pharmacoeconomics and Outcomes Research Task Force for Clinical Outcomes Assessment. Clinical outcome assessments: conceptual foundation—report of the ISPOR Clinical Outcomes Assessment – Emerging Good Practices for Outcomes Research Task Force. Value Health 2015;18:741–52.

[9] Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 2001;69:89–95.
[10] US Food and Drug Administration. Clinical outcome assessment (COA): glossary of terms. Available from: http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm370262.htm#ObsRO. [Accessed May 15, 2017].
[11] US Food and Drug Administration. The 510(k) program: evaluating substantial equivalence in premarket notifications [510(k)]: guidance for industry and Food and Drug Administration staff. 2014. Available from: http://www.fda.gov/downloads/MedicalDevices/.../UCM284443.pdf. [Accessed May 15, 2017].
[12] The Council of European Communities. Medical Device Directive 93/42/EEC. 1993. Available from: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:1993L0042:20071011:en:PDF. [Accessed May 15, 2017].
[13] Health Canada. Medical device establishment license application: form and instructions. 2012. Available from: http://www.hc-sc.gc.ca/dhp-mps/alt_formats/pdf/compli-conform/licences/form/FRM-0292-eng.pdf. [Accessed May 15, 2017].
[14] US Food and Drug Administration. Mobile medical applications: guidance for industry and Food and Drug Administration staff. 2015. Available from: http://www.fda.gov/downloads/MedicalDevices/.../UCM263366.pdf. [Accessed May 15, 2017].
[15] US Food and Drug Administration. Guidance for industry: patient-reported outcome measures—use in medical product development to support labeling claims. 2009. Available from: http://www.fda.gov/downloads/Drugs/.../Guidances/UCM193282.pdf. [Accessed May 15, 2017].
[16] Byrom B, Rowe DA. Measuring free-living physical activity in COPD patients: deriving methodology standards for clinical trials through a review of research studies. Contemp Clin Trials 2016;47:172–84.
[17] Beatty PC. Cognitive interviewing: the use of cognitive interviews to evaluate PRO instruments. In: Byrom B, Tiplady B, eds, ePRO: Electronic Solutions for Patient Reported Data. Farnham, England: Gower Publishing Ltd., 2010.
[18] US Food and Drug Administration. General principles of software validation: final guidance for industry and FDA staff. 2002. Available from: http://www.fda.gov/cdrh/comp/guidance/938.html. [Accessed May 15, 2017].
[19] Eslinger D, Rowlands AV, Hurt TL, et al. Validation of the GENEA accelerometer. Med Sci Sport Exerc 2011;43:1085–93.
[20] Coons SJ, Gwaltney CJ, Hays RD, et al. Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO Good Practices Task Force Report. Value Health 2009;12:419–29.
[21] McEntegart DJ. Equivalence testing: validation and supporting evidence when using modified PRO instruments. In: Byrom B, Tiplady B, eds, ePRO: Electronic Solutions for Patient Reported Data. Farnham, England: Gower Publishing Ltd., 2010.
[22] Meltzer LJ, Montgomery-Downs HE, Insana SP, Walsh CM. Use of actigraphy for assessment in pediatric sleep research. Sleep Med Rev 2012;16:463–75.
[23] McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. JAMA 2014;312:1342–3.
[24] Brożek JL, Guyatt GH, Schünemann HJ. How a well-grounded minimal important difference can enhance transparency of labeling claims and improve interpretation of a patient reported outcome measure. Health Qual Life Outcomes 2006;4:69–75.
[25] McLeod LD, Coon CD, Martin SA, et al. Interpreting patient-reported outcome results: US FDA guidance and emerging methods. Expert Rev Pharmacoecon Outcomes Res 2011;11:163–9.
[26] Demeyer H, Burtin C, Hornikx M, et al. The minimal important difference in physical activity in patients with COPD. PLoS One 2016;11:e0154587.
[27] Motl RW, Pilutti LA, Learmonth YC, et al. Clinical importance of steps taken per day among persons with multiple sclerosis. PLoS One 2013;8:e73247.
[28] Ward MM, Marx AS, Barry NN. Identification of clinically important changes in health status using receiver operating characteristic curves. J Clin Epidemiol 2000;53:279–84.
[29] Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis 1986;39:897–906.