Digital Biomarkers

Digit Biomark 2020;4(suppl 1):3–12
DOI: 10.1159/000511586
Received: July 27, 2020
Accepted: September 14, 2020
Published online: November 26, 2020

Viewpoint – Maturation

The Path Forward for Digital Measures: Suppressing the Desire to Compare Apples and Pineapples

Carrie R. Houts[a] Bray Patrick-Lake[b] Ieuan Clay[b] R.J. Wirth[a]

[a]Vector Psychometric Group, LLC, Chapel Hill, NC, USA; [b]Evidation Health, Inc., San Mateo, CA, USA

Keywords
Digital data · Digital measures · Patient-reported outcomes · Validation

Abstract
Digital measures are becoming more prevalent in clinical development. Methods for robust evaluation are increasingly well defined, yet the primary barrier for digital measures to transition beyond exploratory usage often relies on a comparison to the existing standards. This article focuses on how researchers should approach the complex issue of comparing across assessment modalities. We discuss comparisons of subjective versus objective assessments, or performance-based versus behavioral measures, and we pay particular attention to the situation where the expected association may be poor or nonlinear. We propose that, rather than seeking to replace the standard, research should focus on a structured understanding of how the new measure augments established assessments, with the ultimate goal of developing a more complete understanding of what is meaningful to patients.

Introduction

Digital measures, derived using computational methods from at-home monitoring technologies, including wearables and smartphones [1], offer a range of benefits. These include objective, continuous insights into patient behavior and physiology that are unencumbered by recall effects observed in subjective, patient-reported tools. Digital measures also provide a way to detect intermittent/rare events or create novel measures that more sensitively assess patient experience [2, 3].

Carrie R. Houts
Psychometrics
Vector Psychometric Group, LLC
847 Emily Ln, Chapel Hill, NC 27516 (USA)
crhouts@VPGcentral.com

Digit Biomark 2020;4(suppl 1):3–12
DOI: 10.1159/000511586

Digital measures are becoming increasingly adopted into clinical trials and clinical practice [4–6]. At the same time, the field is becoming ever more aligned on best practice for evaluation of what constitutes a fit-for-purpose tool [7, 8]. Nevertheless, the most significant barrier to demonstrating fit for purpose is clinical validation, typically the barrier for a new measure to transition beyond exploratory usage. Can a digital measure replace the established measure? Should both be used in combination? What claims can be made based on the digital measure?

The critical step in evaluation of a digital measure is the determination of whether the measure is able to capture "clinically meaningful" changes. This is a highly complex question requiring examination of change on both group and individual levels [9] to ensure that the patient experience of change is accurately captured [10]. While the US Food and Drug Administration (FDA) has embraced patient-focused drug development, which includes selecting endpoints that matter to patients, there may be limited tools available to generate evidence to support claims based on these tools [11]. For example, the existence of an effective therapy can enable critical experiments in evaluating a novel measure (steroid treatment in Duchenne muscular dystrophy was central in evidence generation for qualification of a wearable-derived measure of stride velocity [12, 13]), but what if such interventions do not yet exist or are not effective? In practice, evaluation will therefore focus on comparison of established assessments to the novel digital measure, where the established assessment has been shown to support clinically valid inferences. A strong association provides validation by proxy to the new measure, but what if there is not a good association? What counts as "good"?

This is reflected in the very few health authority-accepted or -qualified digital measures [13, 14], and very few documented examples of digital measures being used as primary outcomes [15]. While many digital measures have progressed to later stages of evaluation and wider use, they are "apples-to-apples" cases which capture a remote version of an established assessment, for example, FLOODLIGHT [16], mPOWER [17], Cognition Kit [18], or PARADE [19].

This perspective piece focuses on how researchers should approach the complex issue of comparing across assessment modalities, with the ultimate goal of developing more patient-centered measures. In scope are comparisons of subjective versus objective assessments, or performance-based versus behavioral measures, and we pay particular attention to the situation where the expected association may be poor or nonlinear.

Barriers to Adoption

The rising uptake of digital technologies into clinical trials is a highly promising indication of increasing confidence in, and availability of, digital measures [4]. Nevertheless, a lack of standardization across devices (what is worn/where/when, or what is measured/derived as endpoints, etc.) [20] and a lack of transparency in validation of devices (evidence-derived values are reliable and accurate; comparability of similar values across devices, etc.) [21] still remain as barriers to widespread adoption and incorporation of these tools into clinical decision making.

Traditionally, fit-for-purpose wearables/digital data were examined using the standard patient-reported outcome (PRO) strategy and included examining wearable data against scores from PRO measures (PROMs). In recent years, several recommendations and frameworks have been proposed for the evaluation of digital measures (e.g., see: CPATH [22], CTTI [23], a joint framework proposed by members from the Drug Information Association's [DIA] Study Endpoint Community, CTTI, the ePRO Consortium, and the Digital Medicine Society [(DiMe]) [24], "V3" [7]) which has led to new guidance documents being released by health

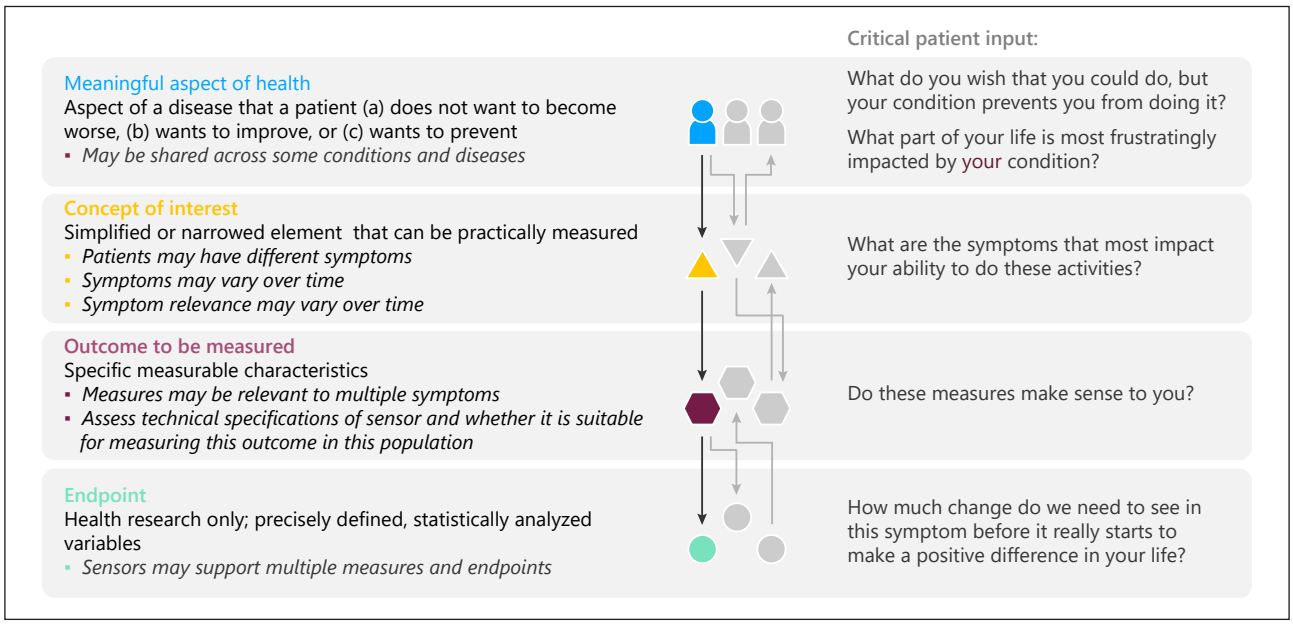

Fig. 1. On the right-hand side, the hierarchy links meaningful aspects of health (MAHs) to concepts of interest (COIs) to outcomes and endpoints. On the left-hand side, critical patient input is highlighted. Defining MAH and COI should take precedence over technical aspects of defining outcomes and endpoints. For a given individual and condition, multiple MAHs can be relevant, equally multiple COIs can inform a given MAH. These relationships may change over time and across individuals. Reproduced with permission from [68].

authorities including the European Medicines Agency Innovation Task Force [25–27] or the FDA Center for Drug Evaluation and Research [28]. These guidelines aim to improve our ability to use digital data and offer ever clearer mechanisms for early engagement [29]. Together, they should lead to greater adoption of wearable/digital data within clinical research [30].

Concepts versus Measures

How a new digital measure is established and evaluated is determined by whether what is being measured (the "concept") is novel, whether the measurement itself is novel, or both [31]. It is also crucial that measures matter to patients by assessing meaningful aspects of health (MAH).

MAH broadly defines an aspect of a disease that a patient (a) does not want to become worse, (b) wants to improve, or (c) wants to prevent [32, 33]. In the case of Duchenne muscular dystrophy, the patients report important, specific activities of their day such as not being able to navigate stairs or desires to walk longer distances [34]. Such activities can be grouped into an MAH category under ambulatory activities that can be readily assessed in a real-world setting. Once the MAH is identified, a concept of interest (COI) can be defined. The COI is a simplified or narrowed element of a MAH that can be practically measured as shown in Figure 1.

Selecting an appropriate COI that is meaningful to patients is a crucial step in narrowing the MAH into a targeted aspect for actual measurement before selecting sensors or devices that can capture specific, measurable characteristics of the disease, and before symptom to

sensor mapping occurs. In some cases, a digital measure focuses on remote capture of an already established assessment or battery of assessments. mPOWER [17], Cognition Kit [18], FLOODLIGHT [16], and PARADE [19], for example, all primarily focus on increasing sampling density and lowering patient burden for assessments which are known to be relevant for their respective indications. A direct comparison can be made between the existing assessments and their digital implementation.

More challenging is the development of novel measures which address an established concept, such as a novel measure of behavior which must be compared to established performance metrics. Measures of real-world mobility in sarcopenia [35, 36] or schizophrenia [37] (in a small case-control study), for example, can be compared to established objective measures which address the same concepts in those conditions. Similarly, a sensor-derived, objective digital measure may address a concept which is also captured by a PRO, for example, social anxiety [38] (in a small pilot study), mood disorders [39] (in a small pilot study), depression [40], or stress [41]. In all these cases, development and validation of the new measure heavily rely on comparisons to established measures that address the same or a related concept, which can often include cross-measurement modalities (e.g., behavior versus performance, objective versus subjective measures).

Where the concept itself also needs to be established is a highly challenging situation, that is, where no reference measures exist, but this is out of scope for this article as it has been covered elsewhere [33].

Apples and Pineapples

Perhaps the greatest strength of subjective assessments is that they directly answer the question of whether the patient "feels" better, specifically whether they perceive improvements in their quality of life or related MAH.

One of the most prevalent issues that has likely deterred the widespread adoption of digital/wearable outcomes in clinical trials is the discordance between the objective variables obtained from mobile devices and wearables and PROs. Associations of PROMs and digital measures have been mixed. While objective wearable data often demonstrate lower associations with PROM scores of purported similar concepts (sleep quality, scratching) than layman expectations would posit [42–45], other areas have found PROMs to be highly aligned with sensor-derived variables [46, 47]. For instance, Bahej et al. [47] used objective measures to forecast PROs and achieved accuracies of around 70–80% for predicting subjectively reported mobility. Similar analyses have found equally promising results across objective and subjective outcomes in cognition [48, 49] and stress [41].

While these results could be interpreted as evidence that objective/digital data and subjective/PROs are assessing similar constructs (and such analyses are useful for providing supportive evidence regarding the validity of new digital data outcomes), we argue that these comparisons, regardless of results (supportive or not), may often times not be appropriate, that is, we are comparing apples and pineapples and attempting to interpret results that may not be conceptually meaningful.

While values/scores from a PROM and a digital variable can be given the same name (e.g., sleep efficiency, physical functioning) that does not mean the same concept is actually being measured by different modalities/devices. For instance, a self-rated physical functioning PRO score may reflect a patient's lived experience of physical functioning due to a disease/condition. At the same time, this patient may still be able to function physically (e.g., measured by steps per day) at a high level relative to other patients with the same disease/condition due to their precondition health status. Neither the patient-reported or digital variable is

inaccurate or "wrong" in this case, it rather demonstrates one of the many possible ways in which subjective/PROM and objective/digital variables could both be "correct" and not be in agreement.

As a more concrete example, take the research area of atopic dermatitis, which is characterized by inflamed skin: there are PROMs to assess itch severity, skin pain, sleep (which can be disturbed by severe dermatitis), and patient-reported dermatitis severity, and there are also actigraphy devices and other digital methods to objectively measure scratching behaviors and sleep parameters [50]. Results across these 2 sources of patient information have found PROMs for itch or sleep and objective scratching/sleep variables to exhibit extremely limited associations. While the terms "itching" and "scratching" are often used interchangeably in everyday language, it is important to note that according to their formal definitions, they are unique [51–53]. Itch and scratch are well-defined terms referring to "an uneasy irritating sensation in the upper surface of the skin usually held to result from mild stimulation of pain receptors" and "to scrape or rub lightly (as to relieve itching)," respectively [54, 55]. So while they are related, even at their base definition one is a subjective experience (itching) and the other is an observable behavior (scratching); it seems reasonable to posit that if a person is itchy then they are also scratching but there are other possible relationships. Given the chronic condition and knowledge that scratching can exacerbate dermatitis, a patient may be extremely itchy but does not scratch. Due to the chronic nature of the condition, another patient may have developed a nonconscious habit of scratching regardless of their itch level. While some researchers have attributed the lack of a relationship between objective and subjective itch/scratch measures to limitations in the subjective ratings [43], it is also plausible that conceptually different things are being measured, and, when viewed in this light, the lack of association could be considered supportive of the discriminant validity of both types of measures.

Further examples can be found in the perception of pain and clinically measured joint function from the Osteoarthritis Initiative [56]. This large initiative has provided several examples of how perceived pain and clinically measured parameters, both of which we would argue are critical to assessing joint function, often bear very weak relationships to each other. Examples include hand pain and joint deterioration assessed by clinical imaging [57], or physical activity and knee pain [58]. While there is a very weak direct relationship (a weakly negative linear association), the importance of measuring joint function from multiple perspectives is demonstrated by perceived pain, when present over longer periods of time, being predictive of longer-term clinical outcomes [59].

In addition to the possible conceptual differences across subjective and objective measures, there are also technical data differences that make comparisons across these 2 sources of patient input questionable. One of the key benefits of digital data, as noted previously, is that it provides a continuous record of what is being measured (spatial coordinates that are translated into number of steps, etc.). While this eliminates the need for recall (e.g., "In the last 2 weeks, I have had trouble walking a city block."), which may be prone to bias or error (see, e.g., Stull et al. [60], for a review of selecting appropriate recall periods and possible sources of bias in PRO recall), it also produces an enormous amount of data. The most common approach to working with the mass of data points is to create summary scores (e.g., hourly/nightly/daily/weekly summary values from means, medians, or other statistics). While aggregation is obviously necessary at some level for analyses to be feasible and results to be interpretable (i.e., analysis of second-by-second tracking of patients' heart rate is not likely useful or interpretable), creating weekly or biweekly averages from digital data variables (to match common recall periods used by PROMs) does not guarantee that a 2-week recall PROM and a 2-week summary of contemporaneously collected data will be well aligned. Additionally, creating such summaries from digital data variables discards an enormous amount of data

(e.g., variability within a person from hour to hour, variability from day to day or night to night, or variability within a person on weekdays versus weekends). To effectively leverage the data collected via digital devices into useful and nuanced statistical results regarding patient health, researchers and analysts will likely need to move away from analyses typically used with clinical trial data and begin investigating novel analysis methods (such as n-of-1 analyses [61, 62], intensive longitudinal models [63, 64], or random effect models [65, 66]) to fully access the knowledge that is waiting to be uncovered in the wealth of information digital data collection provides, and answer a question of most importance to patients, "Based on my personal characteristics, what can I expect my outcomes to be?"

A Path Forward

Given the complex relationships among "similar" variables and analysis considerations, care and thought is needed in specifying expected relationships among objective and subjective assessments of purportedly similar constructs, particularly if a goal is to provide supportive evidence regarding the construct validity of a new digital measure. In collecting information to support the validity of inferences made from any variable, the ideal is to use a "gold standard" measure and demonstrate that the new measure results in similar conclusions regarding patient outcomes on the concept intended to be measured. Our targeted summary of relevant literature has established that PROMs of purportedly similar concepts are likely not appropriate for this use when attempting to establish the validity of digital data variables. We believe that this is primarily due to the fact that, rather than one source of data being "correct" or "incorrect," or more or less accurate, PROMs and digital devices are unique tools for addressing different questions. Rather than pitting them against one another, researchers and regulators would likely best be served by adopting the perspective that information derived from each serves to broaden and deepen our understanding of health-related concepts that are important to patients when assessing the benefits/drawbacks of interventions. This thinking also applies to outcome assessments such as the 6-minute walk test that, from the patient and clinician investigator perspectives, are considered neither to be patient-centered nor gold [67]. Where a COI as multifaceted as mobility and independence (as in the case of the 6-minute walk test) is important to patient quality of life, we would argue that it is extremely risky to only address this COI using a single assessment. Equally, if there are aspects of this COI that are not covered by existing measures and for which novel measures are developed, then requiring these novel measures to perform very similarly to an existing assessment would appear to be self-defeating.

Previous research and current recommendations [33] imply that alternate, more objective measures are likely to be the most useful in establishing that a new digital data source is assessing the concept it intends to assess (e.g., actigraphy steps confirmed by video capture; scratches per hour confirmed by multiple observer ratings); as noted in Walton et al. [33], the feature of digital data variables they term "analytical validity," encompassing "technical performance characteristics such as their accuracy, reliability, precision, consistency over time, uniformity across mobile sensor generations and/or technologies, and across different environment conditions" and content validity for some outcomes derived from digital data sources will be inextricably linked. Regardless of the source of variables being used when attempting to validate a new measure, researchers should make specific, testable hypotheses for the relationships to be tested for COI meaningful to patients prior to interacting with data and be able to justify the a priori expectations through theory and/or previous research.

Finally, given our preferred perspective that PROMs and digital data outcomes provide unique but likely complementary information, a more systematic analytical program should

be undertaken to fully understand how objective and subjective outcomes that, on the surface, are measuring similar concepts relate to one another. Rather than just correlating a PROM score and a variable derived from digital data and despairing when the association is low, systematic research examining multiple PROMs and variables from multiple devices should be investigated using methods such as longitudinal latent variable models or item response theory to construct an empirically supported understanding of how these variables fully relate to one another. Given the digital data status as the "new kid on the block," it may also be the case that until a preponderance of evidence is available that explicates a meaningful, theoretically based, and logically consistent relationship among subjective and objective measures of broader concepts that claims stemming from digital data may need to focus on the specific feature measured (e.g., steps taken, number of scratches) rather than broader, more nebulous "concepts" (e.g., physical functioning). If qualitative work with patients finds that these specific, digitally measured outcomes are meaningful and understandable to patients (i.e., they exhibit content validity), claims to broader concepts from digital data variables may not even be useful.

Regardless of whether we are using measures from subjective reports or objective, digital technologies, the most important issue is to measure what is meaningful to the people whom we are seeking to help. Where the measures fall in the hierarchy of examined endpoints should be strongly informed by qualitative research with patients, and it seems extremely likely that in most cases, the answer to "Should we use a subjective or objective measure?" will be, "Yes." While apples and pineapples are both good on their own, who does not like a nice mixed fruit salad?

Acknowledgment

The authors would like to thank Christine Manta for permission to reproduce Figure 1.

Statement of Ethics

This work involves no human subjects, animals, or trial data of any kind, and as such did not require ethics committee approval.

Conflict of Interest Statement

C.R.H. is an employee of Vector Psychometric Group, LLC. B.P.-L. is an employee of and holds stock options in Evidation Health. She consults for Bayer and is on the Scientific Leadership Board of the Digital Medicine Society. I.C. is an employee of and holds stock options in Evidation Health. He has received payment for lecturing on Digital Health at the ETH Zurich and FHNW Muttenz. He is an Editorial Board Member at Karger Digital Biomarkers and a founding member of the Digital Medicine Society. R.J.W. is a managing member and employee of Vector Psychometric Group, LLC.

Funding Sources

This work received no direct funding.

Author Contributions

All authors contributed to the conceptualization, design, drafting, and final approval of the manuscript.

References

1 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.
2 Clay I. Impact of Digital Technologies on Novel Endpoint Capture in Clinical Trials. Clin Pharmacol Ther. 2017 Dec;102(6):912–3.
3 Papadopoulos E. Advancing Real-World Evidence to Incorporate Patient-Generated Health Data. 2020 May [cited July 20, 2020]. Available from: https://www.ispor.org/docs/default-source/intl2020/ispor-2020workshop10.pdf?sfvrsn=93ed73d8_0.
4 Marra C, Chen JL, Coravos A, Stern AD. Quantifying the use of connected digital products in clinical research. NPJ Digit Med. 2020 Apr;3(1):50.
5 Office of the Commissioner. Real-World Evidence [Internet]. US Food and Drug Administration. 2020 Jul [cited July 19, 2020]. Available from: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence.
6 Library of Digital Endpoints – Digital Medicine Society (DiMe) [Internet]. [cited July 19, 2020]. Available from: https://www.dimesociety.org/index.php/knowledge-center/library-of-digital-endpoints.
7 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.
8 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.
9 Weinfurt KP. Clarifying the Meaning of Clinically Meaningful Benefit in Clinical Research: Noticeable Change vs Valuable Change. JAMA. 2019 Dec;322(24):2381.
10 Center for Drug Evaluation. Research. CDER's Patient-Focused Drug Development [Internet]. US Food and Drug Administration. 2020 [cited July 12, 2020]. Available from: https://www.fda.gov/drugs/development-approval-process-drugs/cder-patient-focused-drug-development.
11 Center for Drug Evaluation. Research. Patient-Focused Drug Development Guidance Series [Internet]. US Food and Drug Administration. 2020 [cited July 19, 2020]. Available from: https://www.fda.gov/drugs/development-approval-process-drugs/fda-patient-focused-drug-development-guidance-series-enhancing-incorporation-patients-voice-medical.
12 Servais L, Gidaro T, Seferian A, Gasnier E, Daron A, Ulinici A, et al. Maximal stride velocity detects positive and negative changes over 6- month-time period in ambulant patients with Duchenne muscular dystrophy. Neuromuscul Disord. 2019 Oct;29:S105.
13 Haberkamp M, Moseley J, Athanasiou D, de Andres-Trelles F, Elferink A, Rosa MM, et al. European regulators' views on a wearable-derived performance measurement of ambulation for Duchenne muscular dystrophy regulatory trials. Neuromuscul Disord. 2019 Jul;29(7):514–6.
14 Center for Drug Evaluation. Research. Clinical Outcome Assessment Compendium [Internet]. 2019 [cited September 9, 2020]. Available from: https://www.fda.gov/drugs/development-resources/clinical-outcome-assessment-compendium.
15 Library of Digital Endpoints – Digital Medicine Society (DiMe) [Internet]. [cited September 9, 2020]. Available from: https://www.dimesociety.org/index.php/knowledge-center/library-of-digital-endpoints.
16 Creagh AP, Simillion C, Scotland A, Lipsmeier F, Bernasconi C, Belachew S, et al. Smartphone-based remote assessment of upper extremity function for multiple sclerosis using the Draw a Shape Test. Physiol Meas. 2020 Jun;41(5):054002.
17 Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci Data. 2016 Mar;3(1):160011.
18 Cormack F, McCue M, Taptiklis N, Skirrow C, Glazer E, Panagopoulos E, et al. Wearable Technology for High-Frequency Cognitive and Mood Assessment in Major Depressive Disorder: Longitudinal Observational Study. JMIR Ment Health. 2019 Nov;6(11):e12814.
19 Hamy V, Garcia-Gancedo L, Pollard A, Myatt A, Liu J, Howland A, et al. Developing Smartphone-Based Objective Assessments of Physical Function in Rheumatoid Arthritis Patients: the PARADE Study. Digit Biomark. 2020 Apr;4(1):26–43.
20 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 Mar;47:172–84.
21 Petersen J, Austin D, Sack R, Hayes TL. Actigraphy-based scratch detection using logistic regression. IEEE J Biomed Health Inform. 2013 Mar;17(2):277–83.

22 Byrom B, Watson C, Doll H, Coons SJ, Eremenco S, Ballinger R, et al.; ePRO Consortium. Selection of and Evidentiary Considerations for Wearable Devices and Their Measurements for Use in Regulatory Decision Making: recommendations from the ePRO Consortium. Value Health. 2018 Jun;21(6):631–9.
23 Recommendations CT. Advancing the Use of Mobile Technologies for 2 Data Capture & Improved Clinical Trials [Internet]. [cited July 12, 2020]. Available from: https://www.ctti-clinicaltrials.org/sites/www.ctti-clinicaltrials.org/files/mobile-devices-recommendations.pdf.
24 Walton MK, Cappelleri JC, Byrom B, Goldsack JC, Eremenco S, Harris D, et al. Considerations for development of an evidence dossier to support the use of mobile sensor technology for clinical outcome assessments in clinical trials. Contemp Clin Trials. 2020 Apr;91:105962.
25 European Medicines Agency. Innovation in medicines – European Medicines Agency [Internet]. 2018 Sep [cited May 27, 2020]. Available from: https://www.ema.europa.eu/en/human-regulatory/research-development/innovation-medicines.
26 Cerreta F, Ritzhaupt A, Metcalfe T, Askin S, Duarte J, Berntgen M, et al. Digital technologies for medicines: shaping a framework for success. Nat Rev Drug Discov. 2020 Sep;19(9):573–4.
27 European Medicines Agency Human Medicines Division. Questions and Answers: Qualification of digital technology-based methodologies to support approval of medicinal products [Internet]. 2020 Jun [cited May 28, 2020]. Available from: https://www.ema.europa.eu/en/documents/other/questions-answers-qualification-digital-technology-based-methodologies-support-approval-medicinal_en.pdf.
28 Center for Drug Evaluation and Research. Drug Development Tool (DDT) qualification process [Internet]. US Food and Drug Administration. 2019 Oct [cited May 27, 2020]. Available from: https://www.fda.gov/drugs/drug-development-tool-ddt-qualification-programs/drug-development-tool-ddt-qualification-process.
29 Fitt H. Qualification of novel methodologies for medicine development - European Medicines Agency [Internet]. European Medicines Agency. 2020 May [cited May 28, 2020]. Available from: https://www.ema.europa.eu/en/human-regulatory/research-development/scientific-advice-protocol-assistance/qualification-novel-methodologies-medicine-development-0.
30 Izmailova ES, Wagner JA, Ammour N, Amondikar N, Bell-Vlasov A, Berman S, et al. Remote Digital Monitoring for Medical Product Development. Clin Transl Sci. 2020 Aug;3:cts.12851.
31 Goldsack JC, Izmailova ES, Menetski JP, Hoffmann SC, Groenen PM, Wagner JA. Remote digital monitoring in clinical trials in the time of COVID-19. Nat Rev Drug Discov. 2020 Jun;19(6):378–9.
32 Clinical Trials Transformation Initiative. Novel Endpoints [Internet]. Clinical Trials Transformation Initiative. 2016 May [cited July 22, 2020]. Available from: https://www.ctti-clinicaltrials.org/projects/novel-endpoints.
33 Walton MK, Powers JH 3rd, Hobart J, Patrick D, Marquis P, Vamvakas S, 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 Sep;18(6):741–52.
34 NPR Choice page [Internet]. [cited July 22, 2020]. Available from: https://www.npr.org/transcripts/893227074.
35 IMI MOBILISE-D [Internet]. Mobilise-D. [cited May 26, 2020]. Available from: https://www.mobilise-d.eu/.
36 Mueller A, Hoefling HA, Muaremi A, Praestgaard J, Walsh LC, Bunte O, et al. Continuous Digital Monitoring of Walking Speed in Frail Elderly Patients: Noninterventional Validation Study and Longitudinal Clinical Trial. JMIR Mhealth Uhealth. 2019 Nov;7(11):e15191.
37 Depp CA, Bashem J, Moore RC, Holden JL, Mikhael T, Swendsen J, et al. GPS mobility as a digital biomarker of negative symptoms in schizophrenia: a case control study. npj. Digit Med. 2019 Nov;2(1):1–7.
38 Jacobson NC, Summers B, Wilhelm S. Digital Biomarkers of Social Anxiety Severity: Digital Phenotyping Using Passive Smartphone Sensors. J Med Internet Res. 2020 May;22(5):e16875.
39 Jacobson NC, Weingarden H, Wilhelm S. Digital biomarkers of mood disorders and symptom change. NPJ Digit Med. 2019 Feb;2:3.
40 Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. J Med Internet Res. 2015 Jul;17(7):e175.
41 Sano A, Taylor S, McHill AW, Phillips AJ, Barger LK, Klerman E, et al. Identifying Objective Physiological Markers and Modifiable Behaviors for Self-Reported Stress and Mental Health Status Using Wearable Sensors and Mobile Phones: observational Study. J Med Internet Res. 2018 Jun;20(6):e210.
42 Maich KH, Lachowski AM, Carney CE. Psychometric Properties of the Consensus Sleep Diary in Those With Insomnia Disorder. Behav Sleep Med. 2018 Mar-Apr;16(2):117–34.
43 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.
44 Douma JA, Verheul HM, Buffart LM. Are patient-reported outcomes of physical function a valid substitute for objective measurements? Curr Oncol. 2018 Oct;25(5):e475–9.
45 Luna IE, Kehlet H, Peterson B, Wede HR, Hoevsgaard SJ, Aasvang EK. Early patient-reported outcomes versus objective function after total hip and knee arthroplasty: a prospective cohort study. Bone Joint J. 2017 Sep;99-B(9):1167–75.
46 Bini SA, Shah RF, Bendich I, Patterson JT, Hwang KM, Zaid MB. Machine Learning Algorithms Can Use Wearable Sensor Data to Accurately Predict Six-Week Patient-Reported Outcome Scores Following Joint Replacement in a Prospective Trial. J Arthroplasty. 2019 Oct;34(10):2242–7.

47 Bahej I, Clay I, Jaggi M, De Luca V. Prediction of Patient-Reported Physical Activity Scores from Wearable Accelerometer Data: A Feasibility Study: Proceedings of the 4th International Conference on NeuroRehabilitation (ICNR2018), October 16–20, 2018, Pisa, Italy. In: Masia L, Micera S, Akay M, Pons JL, editors. Converging Clinical and Engineering Research on Neurorehabilitation III. Cham: Springer International Publishing; 2019; pp 668–72.
48 Chen R, Jankovic F, Marinsek N, Foschini L, Kourtis L, Signorini A, et al. Developing Measures of Cognitive Impairment in the Real World from Consumer-Grade Multimodal Sensor Streams. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery; 2019; pp 2145–55.
49 Asgari M, Kaye J, Dodge H. Predicting mild cognitive impairment from spontaneous spoken utterances. Alzheimers Dement (N Y). 2017 Feb;3(2):219–28.
50 Price A, Cohen DE. Assessment of pruritus in patients with psoriasis and atopic dermatitis: subjective and objective tools. Dermatitis. 2014 Nov-Dec;25(6):334–44.
51 Bringhurst C, Waterston K, Schofield O, Benjamin K, Rees JL. Measurement of itch using actigraphy in pediatric and adult populations. J Am Acad Dermatol. 2004 Dec;51(6):893–8.
52 Bender BG, Ballard R, Canono B, Murphy JR, Leung DY. Disease severity, scratching, and sleep quality in patients with atopic dermatitis. J Am Acad Dermatol. 2008 Mar;58(3):415–20.
53 Bender BG, Leung SB, Leung DY. Actigraphy assessment of sleep disturbance in patients with atopic dermatitis: an objective life quality measure. J Allergy Clin Immunol. 2003 Mar;111(3):598–602.
54 Definition of ITCH [Internet]. [cited July 20, 2020]. Available from: https://www.merriam-webster.com/dictionary/itch.
55 Rinaldi G. The Itch-Scratch Cycle: A Review of the Mechanisms. Dermatol Pract Concept. 2019 Apr;9(2):90–7.
56 Nevitt M, Felson D, Lester G. The osteoarthritis initiative. Protocol for the Cohort Study. 2006;1. Available from: https://nda.nih.gov/oai.
57 Schaefer LF, McAlindon TE, Eaton CB, Roberts MB, Haugen IK, Smith SE, et al. The associations between radiographic hand osteoarthritis definitions and hand pain: data from the osteoarthritis initiative. Rheumatol Int. 2018 Mar;38(3):403–13.
58 Song J, Chang AH, Chang RW, Lee J, Pinto D, Hawker G, et al. Relationship of knee pain to time in moderate and light physical activities: Data from Osteoarthritis Initiative. Semin Arthritis Rheum. 2018 Apr;47(5):683–8.
59 Wang Y, Teichtahl AJ, Abram F, Hussain SM, Pelletier JP, Cicuttini FM, et al. Knee pain as a predictor of structural progression over 4 years: data from the Osteoarthritis Initiative, a prospective cohort study. Arthritis Res Ther. 2018 Nov;20(1):250.
60 Stull DE, Leidy NK, Parasuraman B, Chassany O. Optimal recall periods for patient-reported outcomes: challenges and potential solutions. Curr Med Res Opin. 2009 Apr;25(4):929–42.
61 Schork NJ. Personalized medicine: time for one-person trials. Nature. 2015 Apr;520(7549):609–11.
62 Duan N, Kravitz RL, Schmid CH. Single-patient (n-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. J Clin Epidemiol. 2013 Aug;66(8 Suppl):S21–8.
63 Walls TA, Schafer JL. Models for Intensive Longitudinal Data. USA: Oxford University Press; 2006. Available from: https://play.google.com/store/books/details?id=Bo4RDAAAQBAJ, https://doi.org/10.1093/acprof:oso/9780195173444.001.0001.
64 Bolger N, Laurenceau JP. Intensive Longitudinal Methods: An Introduction to Diary and Experience Sampling Research. Guilford Press; 2013. Available from: https://play.google.com/store/books/details?id=5bD4LuAFq0oC.
65 Demidenko E. Mixed Models: Theory and Applications with R. John Wiley & Sons; 2013. Available from: https://play.google.com/store/books/details?id=GWBLAAAAQBAJ.
66 Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G, editors. Longitudinal Data Analysis. Chapman & Hall/CRC; 2008.
67 Patrick-Lake B. Slide credit to @DoctorSwig during the @PFFORG Summit. It's time for objective measures that better assess how patients feel, function and survive so we can inform treatment selection and shared decision making. Twitter. 2019 [cited July 20, 2020]. Available from: https://twitter.com/BrayPatrickLake/status/1193238806633287684.
68 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.