1521-0081/72/4/899–909$35.00
PHARMACOLOGICAL REVIEWS

https://doi.org/10.1124/pharmrev.120.000028
Pharmacol Rev 72:899–909, October 2020

ASSOCIATE EDITOR: MARTIN MICHEL

Development of Novel, Value-Based, Digital Endpoints for Clinical Trials: A Structured Approach Toward Fit-for-Purpose Validation

M. D. Kruizinga, F. E. Stuurman, V. Exadaktylos, R. J. Doll, D. T. Stephenson, G. J. Groeneveld, G. J. A. Driessen, and A. F. Cohen

Centre for Human Drug Research, Leiden, The Netherlands (M.D.K., F.E.S., V.E., R.J.D., G.J.G., A.F.C.); Juliana Children's Hospital, HAGA Teaching Hospital, The Hague, The Netherlands (M.D.K., G.J.A.D.); Leiden University Medical Center, Leiden, The Netherlands (M.D.K., F.E.S., G.J.G., A.F.C.); and Critical Path for Parkinson's Consortium, Critical Path Institute, Tucson, Arizona (D.T.S.)

***Abstract*——Novel digital endpoints gathered via wearables, small devices, or algorithms hold great promise for clinical trials. However, implementation has been slow because of a lack of guidelines regarding the validation process of these new measurements. In this paper, we propose a pragmatic approach toward selection and fit-for-purpose validation of digital endpoints. Measurements should be value-based, meaning the measurements should directly measure or be associated with meaningful outcomes for patients. Devices should be assessed regarding technological validity. Most importantly, a rigorous clinical validation process should appraise the tolerability, difference between patients and controls, repeatability, detection of clinical events, and correlation with traditional endpoints. When technically and clinically fit-for-purpose, case building in interventional clinical trials starts to generate evidence regarding the response to new or existing health-care interventions. This process may lead to the digital endpoint replacing traditional endpoints, such as clinical rating scales or questionnaires in clinical trials. We recommend initiating more data-sharing collaborations to prevent unnecessary duplication of research and integration of value-based measurements in clinical care to enhance acceptance by health-care professionals. Finally, we invite researchers and regulators to adopt this approach to ensure a timely implementation of digital measurements**

Address correspondence to: A. F. Cohen, Centre for Human Drug Research, Zernikedreef 8, 2333CL Leiden, The Netherlands. E-mail: ac@chdr.nl
https://doi.org/10.1124/pharmrev.120.000028.

and value-based thinking in clinical trial design and health care.

***Significance Statement*——Novel digital endpoints are often cited as promising for the clinical trial of the future. However, clear validation guidelines are lacking in the literature. This paper contains pragmatic criteria for the selection, technical validation, and clinical validation of novel digital endpoints and provides recommendations for future work and collaboration.**

I. Introduction

Traditional clinical endpoints, such as mortality or clinically validated rating scales, have limitations despite their accepted status as the gold standard in clinical trial design. The worldwide improvement in standards of care means that "hard" endpoints, such as mortality, are increasingly rare and necessitate oversized and overly expensive clinical trials (Kruizinga et al., 2019b). Traditional measurements, such as the 6-minute walk test or a single pulmonary function test, capture no more than a snapshot of the burden of disease and are obtained in clinics rather than the real world (Steinhubl et al., 2017), although they are only loosely related to health-related quality of life (Carranza Rosenzweig et al., 2004).

Health care is moving away from these traditional metrics with the implementation of value-based health care (Porter, 2010). However, the implementation of a value-based approach in clinical trials has lagged (Kruizinga et al., 2019b). The use of digital and wearable technology can help in this endeavor because it could ensure the burden of disease is measured in a more realistic manner objectively, more frequently, at home, and on an individual level (Boehme et al., 2019).

Despite this potential, the exact value and measurement properties of many digital measurements in the context of monitoring disease are unclear (Babrak et al., 2019). Full integration as primary or secondary endpoint in clinical trials would imply the inferential value of these measurements is sufficient to change clinical care or lead to the registration of new medicines, which, bar some exceptions (Haberkamp et al., 2019), is not the case at this moment. Therefore, novel value-based and digital endpoints must be validated before they can be accepted by clinicians or regulators (Coravos et al., 2019). Although a framework for the qualification process of novel endpoints has been proposed, digital endpoints may require a more focused and pragmatic approach (Leptak et al., 2017; Coravos et al., 2020). Validation steps for digital endpoints must include technical validation, which focuses on the properties of the measurements of devices or software, and clinical validation, which is focused on the value of the measurements when used as endpoint for patients. Although the FDA and EMA both recognize the need for validation and have released guidance (European Medicines Agency, 2014, U.S. Food and Drug Administration, 2017, U.S. Food and Drug Administration, 2018), there is a lack of clarity regarding the exact criteria a digital endpoint should fulfill. Therefore, the question remains regarding when a novel measurement is fit-for-purpose as an endpoint in clinical trials.

In this paper, we propose a pragmatic stepwise approach toward technical and clinical validation of digital endpoints (Fig. 1) comparable to the fit-for-purpose validation employed for traditional biomarkers (Cummings et al., 2010; Cohen et al., 2015). We relate each step to example cases in the fields of asthma, cystic fibrosis, pediatrics, and orthopedics.

II. Potential Advantages of Value-Based Digital Endpoints

The inherent characteristics of digital endpoints can add value to clinical trials in numerous ways (Kruizinga et al., 2019b). They allow measurements to be conducted completely at home, increasing participation rates and enabling trials to be conducted in vulnerable populations with chronic diseases, such as the elderly, psychiatric patients, and children. These patient groups have traditionally been neglected in clinical research because of a lack of mobility, additional ethical barriers, and low recruitment rates. An added advantage is the ability to measure effects of an intervention in the natural environment of patients, resulting in increased ecological or "real world" validity. The objective nature of measurements can lead to a higher sensitivity and objectivity compared with clinical rating scales. Wearable technology also offers high frequency and situation-relevant measurements, moving away from the artificially contrived intervals used in clinical trials. The move toward the home is in line with a trend in health care, which is also increasingly delivered outside hospitals. The lack of direct supervision on all assessments generates a potential problem in clinical trials, in which complete control of participants via standardized measurements and environments is the standard. However, the inclusion of these measurements to clinic-based trials and data collection adds a new dimension of real-world data to traditional trial designs. If novel endpoints prove to be of near-identical or even superior value, they may eventually lead to reduced visits to the clinic and improved efficiency.

ABBREVIATIONS: ARID1B, AT-rich interactive domain-containing protein 1B; EMA, European Medicines Agency; FDA, Food and Drug Administration; GPS, Global Positioning System.

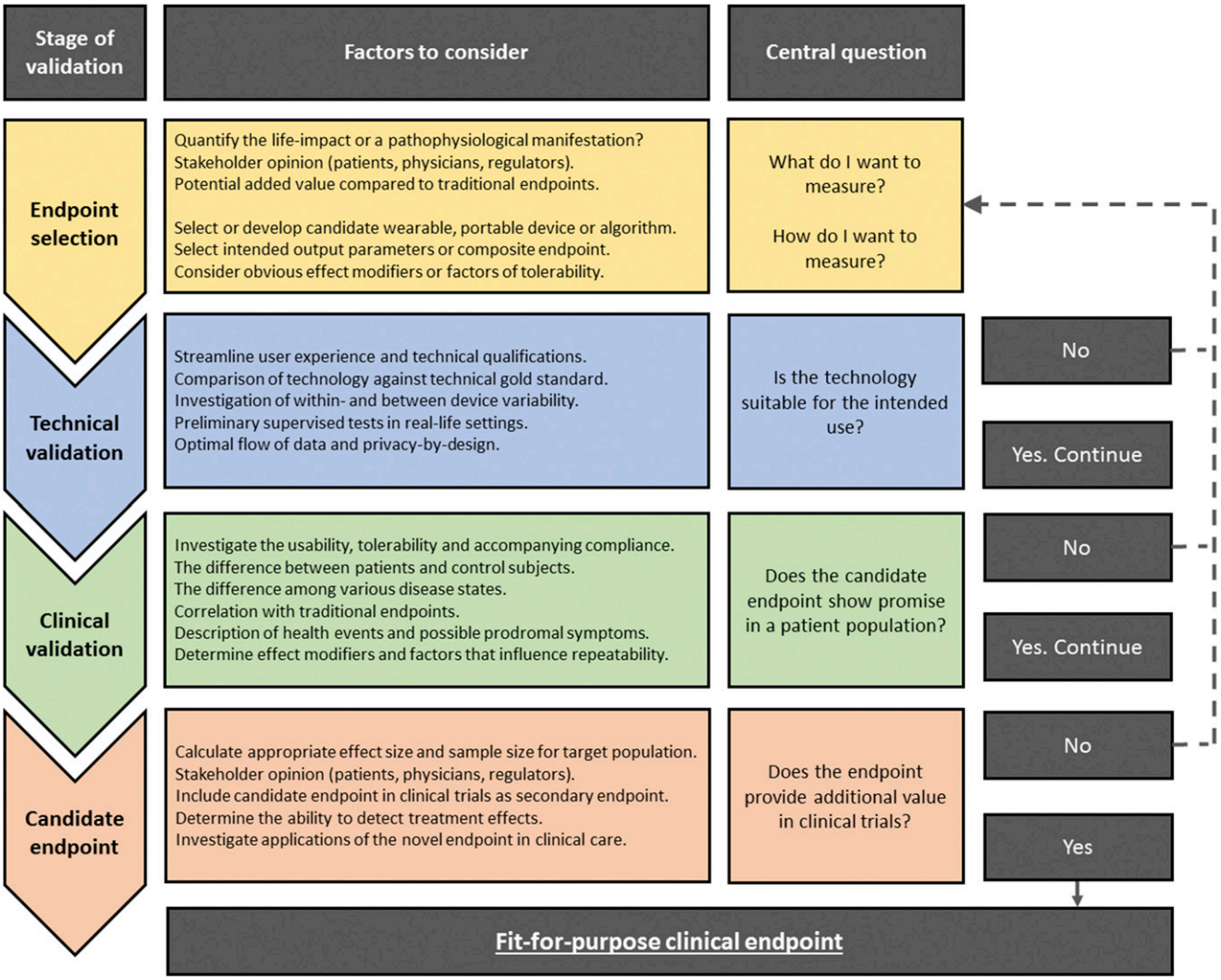

Fig. 1. Structured approach to developing a fit-for-purpose digital endpoint. Factors to consider during development and validation of novel digital endpoints.

III. Selection of Candidate Digital Endpoints

Choosing the right digital biomarker for evaluation in a specific clinical condition is challenging. There is an increasing amount of potential digital biomarkers and a large array of start-up companies vying for investors and patient users. In this paper, we assume that devices used for the monitoring, diagnosis, or prognosis of disease adhere to the medical device regulations, which is the responsibility of the manufacturer (Ben-Menahem et al., 2020). The regulations contain subtle differences between countries and stratify devices in several classes based on the context and risk associated with their use (Chen et al., 2018; Gordon et al., 2020). Although not a device per se, health-related software is regarded as a medical device and must comply with applicable regulations.

Table 1 lists several candidate digital biomarkers, many of which have already been investigated in early feasibility studies (Bakker et al., 2019) or industry-sponsored clinical trials (Chasse et al., 2019). A good candidate endpoint should be accurate, reliable, and value-based and therefore directly measure meaningful outcomes for the individual subject (e.g., sleep, activity, pain, ability to move, gait) or be associated with important clinical outcomes (e.g., occurrence of mortality, morbidity, complications). Furthermore, the assessment or device must be usable, tolerable, and suitable for the intended users. Conceptually, there must be clear benefit of using the digital measurement over existing methods. The candidate endpoint should also have a plausible relationship with the studied disease or general quality of life. These factors together imply that a close collaboration with patients and patient advocacy groups is vital during the selection and validation process.

If the goal at this point is to obtain regulatory endorsement of the novel digital endpoint, regulators advise early engagement to identify and define the

TABLE 1
Potential devices and candidate digital endpoints for use in clinical trials outside of clinical units

Device/Sensor	Candidate Endpoint	Patient Domain
Accelerometer	Physical activity	Mobility
	Steps	Symptom severity
	Gait patterns	Symptom severity
	Tremor analysis	Symptom severity
Blood-pressure meter	Blood pressure	Cardiovascular health
Camera	Dermatological assessments	Dermatological health
	Treatment adherence	Compliance
Dynamometer	Muscle strength	Musculoskeletal health
ECG	Event detection	Cardiovascular health
Glucose monitor	Glucose	Diabetic control
Oximeter	Oxygen saturation	Pulmonary healthy
GPS	GPS mobility	Mobility
	Location type	Social behavior
Light sensor	Light intensity	Environment
Microphone	Event detection	Clinical events
	Voice analysis	Mood
PPG	Heart rate	Cardiovascular health
Smartphone	App use	Social behavior
	Phone use (calls, sms)	Symptom severity
	Patient reported outcomes	Symptom severity
Spirometer	Pulmonary function	Pulmonary health
Thermometer	Temperature	Infection control
		Thermoregulation
Touch screen	Response time	Dexterity
	Speed of typing	Coordination
	Custom tests	Various

PPG, photoplethysmography.

"Concept of Interest" that underpins the digital endpoint from a regulatory point of view. This engagement also serves to identify appropriate regulatory interaction channels going forward and to discuss how a clinically meaningful change can be defined and investigated (Cerreta et al., 2020).

Even when a device to measure digital biomarkers has been selected, choosing the right way to display and analyze a measurement is difficult. Physical activity currently has at least four separate units: step count, duration of moderate-to-vigorous physical activity, accelerometer counts per minute, and average gait speed. In the case of multiple promising and related measurements, machine learning or other artificial intelligence techniques can be used to choose and combine several metrics in an algorithm to produce a composite score comprised of several novel and traditional endpoints (Zhan et al., 2018).

IV. Technical Validation

Before using the novel measurement in patients, a robust assessment of the usability, reliability, and reproducibility of the technology and flow of data should be performed.

A. Minimal Technological Standards

The reliability and consistency of devices should be assessed in the form of interdevice and intradevice variability. The flow of data should be automated, requiring as little manual input as possible to reduce data (entry) errors. The flow should be consistent and allow for the subjects' privacy by design—for example, via encrypted transmission of data (Angeletti et al., 2018). Furthermore, the data flow must be part of the validation and comply with the necessary FDA regulations regarding audit trails and the storage and processing of source data (U.S. Food and Drug Administration, 2017). In the case of technological gold standards that can be used as a reference, a head-to-head comparison should be conducted in a standardized setting to determine the bias and limits of agreement, specificity, and sensitivity, depending on the type of measurement.

Furthermore, an analysis should be conducted with nonpatient test subjects to ensure that a novel measurement truly captures the behavior, symptom, or activity it attempts to quantify in various real-life situations. For example, a smartphone accelerometer is capable of counting steps taken per day, but some may leave their phones on a desk or in a locker, bag, or jacket

when going for a walk, which leads to low accelerometer counts with little information of the underlying reason. Although it is not possible to simulate all situations that might occur in daily life, a supervised test of limited duration may result in the detection of easily addressable confounders. In this case, the exact clinical relevance will not yet be determined, and the test subject merely functions as a free-living data generator in a closely observed setting. In this phase of validation, it is also advised to consider the amount of training and instruction that will be necessary to ensure measurements are conducted correctly by patients.

Not all these steps are feasible in all cases. For example, when there is no obvious technical gold standard or when the measurement is a completely new concept, a head-to-head comparison is impossible. In this case, technical validation is necessarily more limited, and clinical validation gains a larger role.

B. Handling of Discordance

Technological validation of novel measurements is vital for the validation process, and if there is a near-perfect agreement with an unchallenged technical gold standard, the validation process may end there. However, often there will be some "technical noise" or measurement bias associated with the miniaturization process, potentially undermining the accuracy of the home-based measurements. Although medical-grade devices will likely have less technical noise compared with consumer devices, they are often extremely expensive and unwieldy, which limits widespread implementation and causes a reduced compliance compared with more user-friendly (consumer) devices (Schrack et al., 2016). Furthermore, technical noise or bias is not necessarily disqualifying, and random deviations from the gold standard would not significantly alter treatment-effect estimates in a clinical trial (Buyse et al., 2017). The major advantage of home-based monitoring is that the resolution of data can be very high and will be likely to outperform traditional endpoints despite suboptimal technological validity. To illustrate this further, Fig. 2 shows serial home-based blood-pressure measurements of a patient with a history of hypertension whose antihypertensive was switched by the pharmacy. The graph shows a gradual but clinically significant increase of the systolic blood pressure over time. However, when the data would have been limited to a small amount of clinic-based measurements (e.g., at the time points indicated by the black arrow), one could easily have concluded that blood pressure was much lower on Irbesartan compared with when that patient was on Valsartan. This example shows that, especially for measurements with a high intraindividual variability, increasing the measurement frequency leads to more valid conclusions on an individual level. Even the presence of a random measurement error will still lead to a better clinical overview compared with six weekly, or even more infrequent, measurements, with the caveat that the error must be significantly lower than the intraindividual variability.

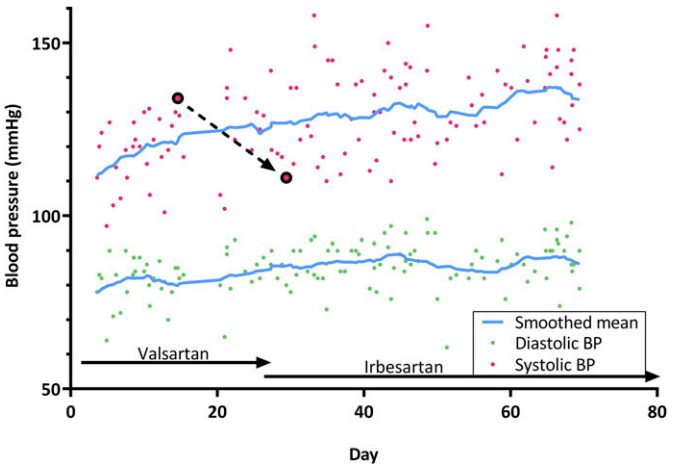

Fig. 2. Increased frequency can overcome variability. Individual patient who measured blood pressure in a home setting for a period of 70 days. The patient was switched to a different antihypertensive drug by his pharmacy and, on average, measured slightly higher systolic blood pressures over time. When monitored during regular outpatient clinic visits—for example, at the timepoints indicated by the black arrow—this effect could have been completely missed, and a completely opposite conclusion could have been drawn. BP, blood pressure.

C. Pitfalls

Investigators should also not be too focused on theoretical measurement errors inherent to home monitoring. For example, a recent FDA advice requested that activities that may resemble "steps," such as repetitive movements of the arm, must be discriminated from actual steps. Furthermore, a plan had to be in place to make certain that devices are not used by anyone else (Papadopoulos and Norman, 2018). Although these requests can certainly be relevant on a technical level, there may be many reasons why this may be irrelevant in the context of a clinical trial—for example, if the conditions when this inaccuracy might occur are very rare or if the impact of the perceived inaccuracy when measuring an association with the severity of a clinical condition is negligible. In such cases, additional requirements such as these might inhibit implementation unnecessarily. Regulators should be wary of a slippery slope: The unsupervised nature of home-based measurements introduces many uncertainties, and new uncertainties could be identified during every evaluation. During the qualification process of the 6-minute walk test, we imagine there was no request for a plan to ensure a subject does not walk slower on purpose. In practice, most of these randomly generated data-quality issues will be mitigated by the application of randomization and blinding (Buyse et al., 2017), which will remain the standard in pivotal clinical trials. When regulators focus on a clinical validation process that is more rigorous than that for traditional rating

scales, any inaccuracies and uncertainties that are found can be appraised in the perspective of the clinical condition.

V. Clinical Validation

A rigorous clinical evaluation of the candidate endpoint must be performed to determine the potential clinical value. We propose five criteria that should be assessed during this process (Fig. 1). Most of the characteristics can be assessed in observational studies, and some criteria may not be applicable to some measurements. This applies to measurements that are well known in clinic-based trials and are merely miniaturized or streamlined for use in a home setting. For example, an absolute difference between patients and healthy controls is less important for a wearable device for glucose monitoring (Ólafsdóttir et al., 2017) than it is for a novel algorithm quantifying negative symptoms in schizophrenia (Depp et al., 2019). Once the glucose-monitoring device has been shown to adequately measure glucose in various situations in a home setting, it would already be close to fit-for-purpose for clinical trials. It would be unnecessarily burdensome to also associate use of the device with traditional indicators of disease, such as a decrease in the incidence of cardiovascular complications. Furthermore, multiple criteria can be investigated in a single observational study. It is therefore important to critically appraise the novel endpoint prior to initiation of the clinical validation process to define the most important questions and investigate these early and extensively. This question-based approach has been described for prototypical drug development and can be adapted to digital endpoint development as well (Cohen et al., 2015).

A. Tolerability and Usability for Patients

First, assessments should be tolerable for the target population. In general, that means the assessment should be minimally invasive and require as little manual input as possible. Since the use of digital endpoints means that clinical trials will be increasingly decentralized, the user experience of participants is vital to optimize adherence and retention in trials. Technology may allow for a longer follow-up period with a lower burden for subjects but only when the study assessments are tolerable to conduct for extended periods of time. Tolerability is also important when investigating vulnerable populations, such as children, the elderly, and patients that are otherwise impaired. Although problems may appear trivial at first, small technical or usability issues, such as decreased smartphone battery life, may lead to low compliance, workarounds by users, or even dropouts. Researchers must adapt to population-specific needs in an early stage or conclude that the included measurement is unsuitable. An example is the use of a smartwatch in children. In a recent observational study including 391 pediatric subjects, we observed children aged 6–16 who were enthusiastic and demonstrated a drop-out rate of 1.4%. On the other hand, children aged 2–5 were less amused: 17% did not complete the study period (unpublished data). Although the user experience should be optimized to increase compliance, low compliance rates in decentralized studies may still lead to a high number of observations that can provide valuable results. A study by Lipsmeier et al. (2018), which investigated novel smartphone-based tests in 44 patients with Parkinson disease for a duration of 6 months, demonstrated that patients exhibited an average compliance of only 61%. However, this resulted in a data set consisting of 5135 test outcomes and appeared to allow for the detection of subtle symptomatology unlikely to result in a change in traditional symptom-questionnaire scores (Lipsmeier et al., 2018). Still, compliance with the use of digital devices has been a challenge, and there is growing recognition of the need to improve alignment with patients at all stages of development (Bot et al., 2016; Pratap et al., 2020).

B. Difference Between Patients and Controls

An important validation criterion is the difference between patient groups and control groups, which should be assessed vigorously. The magnitude of the difference and the accompanying variability can be used to determine what improvement could be considered clinically relevant for new measurements. The data can also aid in the prospective calculation of the sample size needed to detect an appropriate treatment effect. Further development of novel measurements seems futile when there is no detectable or clinically significant difference since an effective treatment would have to result in the patient group outperforming the control group for the particular measurement. An example from the field of pediatric asthma: Fig. 3A shows the difference in physical activity between children with (un)controlled asthma and healthy children. The candidate endpoint appears to be able to differentiate both between healthy children and asthmatics but also between the two disease states. Increasing the resolution of data can provide interesting insights regarding the time of day responsible for group differences (Fig. 3, B and C)—in this example, between healthy subjects and subjects with ARID1B-related intellectual disability (Kruizinga et al., 2020). In the case of completely new measurements or algorithms, we believe reference values should be obtained for the target populations with special interest toward the influence of age, gender, lifestyle choices, and socioeconomic status because these factors are undoubtedly influential in home-based measurement outcomes. When appraising the difference between patients and controls and estimating relevant treatment effects, a distinction can be made between (partially) reversible and invariably

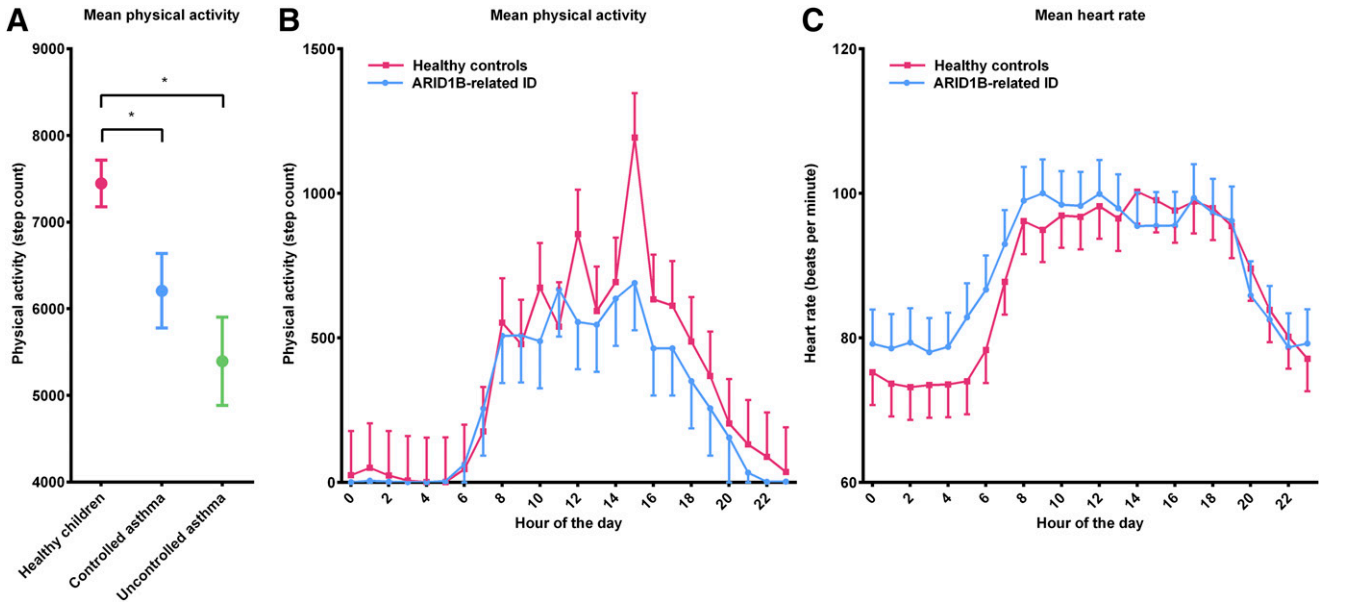

Fig. 3. Difference between patients and healthy controls. (A) Average physical activity (95% CI) per day of healthy children, children with controlled asthma, and children with uncontrolled asthma. (B) Average physical activity per hour of the day of healthy children, children with controlled asthma, and children with uncontrolled asthma. The estimated averages are corrected for age and sex in a mixed-model ANOVA. *$P < 0.05$. (B) Mean (95% CI) physical activity per hour of the day of children with ARID1B-related intellectual disability (ID) and healthy age-matched controls. (C) Mean (95% CI) heart rate per hour of the day of children with ARID1B-related intellectual disability and healthy age-matched controls. CI, confidence interval.

progressive diseases. In the case of progressive disease, the expected gain from treatment may be a decreased speed of deterioration and not an absolute improvement toward reference values.

C. Repeatability and Variability

Measurements should be stable over time in the absence of an intervention or change in disease activity, and a change of the outcome variable should be either due to improvement or exacerbation of the disease. Although this may seem unrealistic for home-based and continuous measurements, it implies a need to consider the impact of real-world variability induced by factors, such as season, weather, and location (Chan and Ryan, 2009), and the impact of baseline factors, such as age and social circumstances. Additionally, concomitant drug use can be a confounding factor in free-living conditions. For example, an improvement of osteoarthritis symptoms may not lead to detectable behavioral changes in the individual patient detected by digital endpoints. When the improvement leads to a reduction in the use of opiates, the improvement is certainly clinically meaningful. A better understanding of the effects of real-world variability also allows for a better quantification of treatment effects in specific individuals, which is one of the hallmarks of value-based health care. Furthermore, in diseases that are progressive by nature, natural history studies regarding the novel measurement can help to quantify the rate of progression and estimate relevant treatment effects in this regard (Jewell, 2016).

D. Correlation with Existing Disease Metrics

The next step is to correlate the novel endpoint with traditional endpoints—ideally, the gold standard. However, perfect correlations will never be achieved considering the nature of both digital and traditional endpoints. Investigators therefore must critically appraise the data to determine whether a suboptimal correlation is due to limitations of the novel endpoint, limitations of the "gold" standard, or because both quantify different aspects of the disease. An uncomplicated example would be to correlate number of steps taken per day versus the 6-minute walk test in patients with chronic obstructive pulmonary disease (Steele et al., 2000). Here, both endpoints are conceptually the same but measured and expressed differently. Interpretation becomes more challenging when correlating GPS mobility with schizophrenia symptom burden (Depp et al., 2019) or asthma control diary scores with steps taken per day in pediatric asthma (Fig. 4) (Kruizinga et al., 2019a). In that case, there appears to be no correlation between daily symptoms and daily physical activity for subjects with controlled asthma (Fig. 4B). This may be because for these patients, asthma symptomology is too limited to interfere with daily life. However, when looking at subjects with uncontrolled asthma (Fig. 4C), the burden of symptoms is higher and appears to significantly impact physical activity as a result. This may lead to the conclusion that physical activity has added value for monitoring of symptoms of children with uncontrolled asthma only. Interpretation should be done carefully while also accounting for the analyses performed during the other

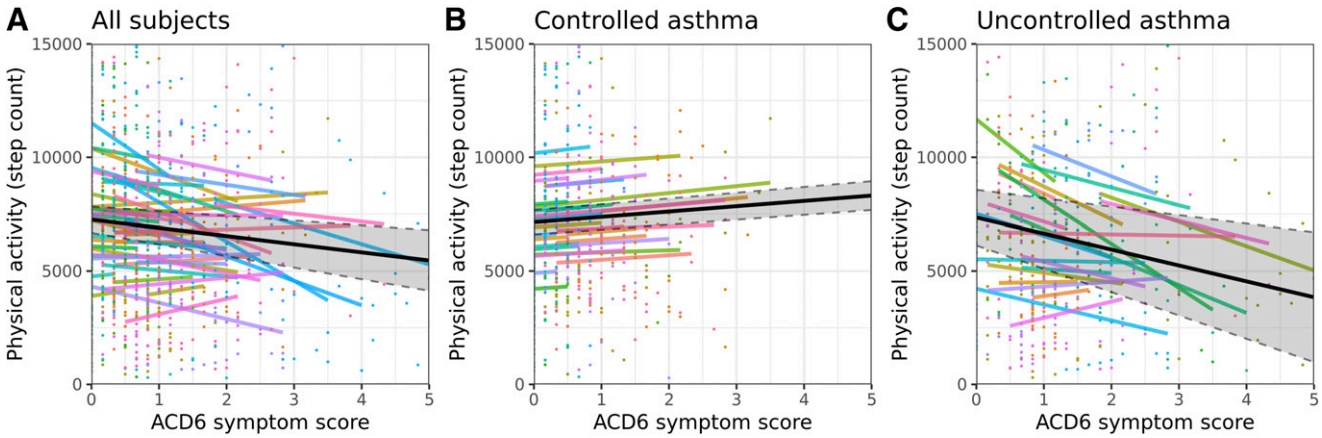

Fig. 4. Modeled relationship of asthma symptoms with physical activity. Relationship of Asthma Control Diary 6 Questions (ACD6) score with physical activity. A mixed-model ANOVA was developed to estimate the relationship with subject as random factor, with a random slope and random intercept. Each dot represents a single day; each color represents a subject. The black line represents the average slope and intercept (95% CI). (A) Analysis including all subjects with asthma. (B) Subgroup analysis of subjects with controlled asthma; (C) subgroup analysis of subjects with uncontrolled asthma. CI, confidence interval.

phases of validation. At this point, the relationship of the novel endpoint with general and disease-related quality of life can also be investigated to assess whether the novel endpoint captures a disease state that is meaningful for patients.

E. Responsive to Change in Disease State

The final step before declaring a candidate endpoint fit-for-purpose is to investigate whether the endpoint will respond to changes in burden of disease. One method is to investigate the effect of specific health events with expected negative effects, such as a pulmonary exacerbation or sickle-cell crisis or events with expected positive effects, such as a surgical intervention. Several features can be extracted from events, such as prodromal symptoms, the slope or rate of recovery, or the time needed to return to baseline values (Fig. 5A). As an example, Fig. 5B shows physical activity patterns of patients who were elderly with osteoarthritis before and after knee replacement surgery. In this graph, a sharp decline in physical activity is visible postoperatively, whereas the recovery van be visualized over the course of several months. In the future, it may be possible to identify good responders to treatment as the subjects who recover above baseline physical activity, although there are other variables that could reflect improvement, such as the earlier mentioned concomitant medication. Furthermore, Fig. 5C displays physical activity and pulmonary function of a single subject with cystic fibrosis undergoing a moderate pulmonary exacerbation (Kruizinga et al., 2019). The clinical event and recovery period are clearly identifiable, and several proposed features in Fig. 5A can be extracted. Another method is to investigate the effects of known effective treatments on the novel endpoint. However, this final step in the validation process is more easily performed in conditions with an approved disease-modifying therapy or disease with known risk of rapid exacerbations, as opposed to many rare or slowly progressive diseases with no proven treatment. In such cases, validation of novel endpoints could be performed in other similar conditions before turning toward the disease of primary interest.

VI. Case Building

The ultimate test for novel endpoints is the interventional clinical trial, which allows the detection of beneficial or deleterious effects of novel or existing treatments. Although it may be tempting to immediately include a novel measurement in trials, it is important to systematically complete the validation process to reliably assess the usability, potential value, and, most importantly, the expected effect size necessary for clinical benefit. When a novel measurement or technology is assessed positively on all criteria, it is fit-for-purpose as clinical endpoint in future interventional clinical trials. Then, case building starts by including the biomarker as exploratory or secondary endpoint in relevant trials and, eventually, when the novel endpoint has proven superior value compared with traditional endpoints regarding study compliance, detection of response to treatment, or the capability of distinguishing various disease states, as primary endpoint in the clinical trial of the future.

VII. Discussion

A. Regulatory Engagement

The current state of regulatory guidance should not deter investigators from including digital measurements in clinical trials. Although guidelines are lacking, the FDA and EMA have provided strong indications that they support the use of wearable, biosensor, and other real-world data in regulatory decision making (https://www.fda.gov/media/120060/download; Cerreta et al., 2020). Every single state-of-the-art biomarker

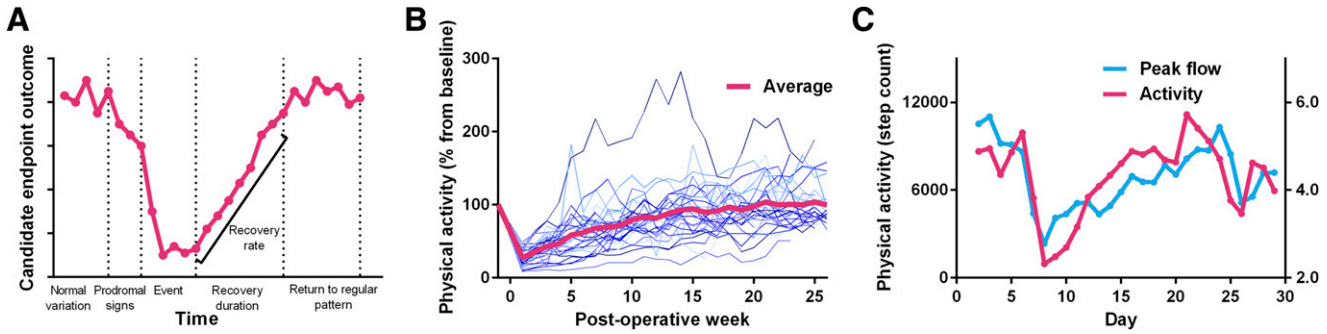

Fig. 5. Response of physical activity to health events. (A) Fictional data of a digital endpoint measured during a clinical event. The features that could be extracted from such data are listed in on the *x*-axis. Additionally, the slope of the line of recovery represents the recovery rate. (B) Nonfictional data of the physical activity change from baseline after knee surgery. Individual lines represent individual patients, and the pink line represents the average. (C) Data of an individual pulmonary event of a patient who was pediatric with cystic fibrosis. The running 2-day average of physical activity (pink) and peak flow (blue) are displayed.

was exploratory at one time. For example, human immunodeficiency virus viral load is the unchallenged gold standard to quantify disease activity in 2020, but this was completely exploratory 25 years prior (Piatak et al., 1993). Although it is not compulsory to only use qualified clinical outcome assessments in clinical trials, early engagement with regulators may streamline the process to qualify novel digital endpoints. Until more experience with digital endpoints has been obtained by both regulators and researchers, the qualification process will be unpredictable. Pioneering work in engagement with EMA has been performed in this regard with the qualification of the stride velocity 95th centile as secondary endpoint in Duchenne muscular dystrophy (Committee for Medicinal Products for Human Use (CHMP), 2019; Haberkamp et al., 2019). The EMA qualification opinion describes an iterative process with multiple discussion sessions and a public consultation. Each session provided additional points needing clarification, and many of the requested clarifications feature in this manuscript. Early adoption of the proposed framework may streamline future iterative discussions with regulators.

The lack of complete control over subjects is a major disadvantage of digital endpoints. Real-world data collection in free-living conditions will invariably add a factor of uncertainty and, although difficult to quantify, this factor may remain a recurring theme during regulatory appraisal. However, there is currently no data regarding the consequences of this loss of control in terms of bias in estimated treatment effects in clinical trials. It is up to investigators to provide enough data to prove the added value of digital endpoints for clinical trials, and the case-building process should start now. Table 2 outlines the various steps presented in this paper and can be used to prepare and plan the validation process and to

TABLE 2
Checklist to use during the planning and execution of the validation process of novel digital endpoints for clinical trials

Step	Question	Applicable (Y/N)	Assessment
Endpoint selection	What aspect of the disease must be measured?	☐	
	How is this aspect directly relevant to patients?	☐	
	What is the optimal device, wearable, or algorithm to measure this aspect?	☐	
	What are the output parameters of this device?	☐	
	How can the measurement be expressed as a single outcome measure?	☐	
	Is there a difference expected between patients and controls?	☐	
Technical validation	Are the technical qualifications on par with researcher requirements?	☐	
	Can the user experience be streamlined?	☐	
	What kind of user training would optimize uptake and retention?	☐	
	How does the measurement compare with the technical gold standard?	☐	
	What is the intradevice and interdevice variability?	☐	
	What are the sensitivity and specificity?	☐	
	Does the measurement capture the aimed behavior or symptom?	☐	
	Is the privacy of subjects guaranteed?	☐	
	Is the data flow stable and compliant with regulatory requirements?	☐	
Clinical validation	Is the device tolerable and usable for the target population?	☐	
	What is the difference between patients and control subjects?	☐	
	What is the difference between the several states of disease?	☐	
	What influences the day-to-day variability within subjects?	☐	
	Can the endpoint detect and describe clinical events or interventions?	☐	
	Is the endpoint correlated to traditional endpoints?	☐	
Candidate endpoint – Application and case building	What phase of clinical research is the device most suitable for?	☐	
	Integration in ongoing or upcoming trials as exploratory endpoint?	☐	
	Can the collected data be extrapolated to other populations?	☐	
	Can anonymized data of control subjects be shared with other parties?	☐	

select the various assessments that are applicable to the candidate endpoint.

B. Recommendations

There are precompetitive collaborative efforts by various stakeholders to support the developmental process of digital endpoints (Coran et al., 2019). Examples are the Critical Path Institute's consortia and the Clinical Trials Transformation Initiative; however, more could be done to stimulate and incentivize implementation and standardize reporting (Byrom and Rowe, 2016; Arnerić et al., 2017; Izmailova et al., 2018; Badawy et al., 2019). International adoption of a single device or algorithm is unrealistic, but there is an established difference between manufacturers in, for example, smartwatches. Furthermore, all algorithms are invariably dependent on the original data set from which they are derived. Nevertheless, data from multiple studies with different devices may eventually be combined for analysis by regulatory agencies or in the context of meta-analysis. For those cases, head-to-head comparisons can be conducted to demonstrate equivalency or to develop conversion factors enabling the comparison of the results of different studies (O'Connell et al., 2016). Regulatory agencies have recommended the creation of normative databases for device platforms, but this has been adopted only on rare occasions (Haberkamp et al., 2019).

Continuing collaboration by academia and industry aimed toward data sharing is crucial to avoid unnecessary duplication (Nature Biotechnology Editorial Team, 2019)—for example, by sharing reference values of novel measurements generated by healthy participants. Adoption of universal data sharing at the level of consent of subjects could accelerate the move forward (Hake et al., 2017). Socially aware wearable companies may also aid in this purpose via data sharing and could be more open toward researchers regarding their proprietary algorithms and raw data for the common interest of improving health care. In the case of devices subject to firmware updates, care must be taken that changes do not impact the validity and repeatability of the measurements.

Furthermore, results should be shared and discussed with the same patients and patient advocacy groups that were consulted during initial candidate endpoint selection to ascertain that the novel measurements truly capture aspects of the disease important to patients.

C. Potential for Clinical Care

To further stimulate acceptance by both patients and health-care providers, suitable digital biomarkers with potential in clinical care should be introduced in the clinic as soon as clinically validated. The high sampling frequency of measurements that are important to patients may allow physicians to obtain a holistic overview of patients' well-being. Digital biomarkers have a wide range of possible applications in clinical care. For example, they can be used to support diagnosis of challenging patients, stratify patients in risk categories, serve as pharmacodynamic response marker, provide monitoring in addition to or in place of traditional visits to the outpatient clinic, and even aid in the prediction of health outcomes after a hospital admission (Burnham et al., 2018; Coravos et al., 2019). Realizing this potentially disruptive new component in value-based health care necessitates a proactive approach toward an improved data infrastructure in hospitals, which must be capable of processing algorithms for diagnostic and follow-up purposes (Panch et al., 2019). Not all digital endpoints with value in clinical trials will add value in clinical care. The context of participating in a clinical trial with incentives, such as access to new treatments or financial reward, might be quite different from the general clinical health setting when determining the value of digital endpoints for patients. The usability and tolerability of measurements with potential in clinical care need to be assessed in this light. Ultimately, addition of novel digital endpoints in standard care requires measurements that are minimally invasive that can reliably detect the individual response to health-care interventions and, finally, are cost-effective as well. These are demanding requirements that will require a long process of clinical validation beyond the steps described in this review.

VIII. Conclusion

The proposed stepwise approach toward technical and clinical validation of novel endpoints in clinical trials is pragmatic and can be applied to most types of digital data. We invite researchers and regulators to endorse and adopt this framework to ensure a timely implementation of digital measurements and value-based thinking in clinical trial design and health care.

Acknowledgments

The authors wish to thank all study participants who contributed to the data used to produce the figures in this manuscript, Marcus van Diemen for designing and conducting the knee replacement study, and other investigators involved in the studies used for the examples.

Authorship Contributions

Participated in research design: Kruizinga, Stuurman, Groeneveld, Driessen, Cohen.

Performed data analysis: Kruizinga.

Wrote or contributed to the writing of the manuscript: Kruizinga, Stuurman, Exadaktylos, Doll, Stephenson, Groeneveld, Driessen, Cohen.

References

Angeletti F, Chatzigiannakis I, and Vitaletti A (2018) Towards an architecture to guarantee both data privacy and utility in the first phases of digital clinical trials. *Sensors (Basel)* **18**:4175.

Arnerić SP, Cedarbaum JM, Khozin S, Papapetropoulos S, Hill DL, Ropacki M, Rhodes J, Dacks PA, Hudson LD, Gordon MF, et al. (2017) Biometric monitoring

devices for assessing end points in clinical trials: developing an ecosystem. *Nat Rev Drug Discov* **16**:736.
Babrak LM, Menetski J, Rebhan M, Nisato G, Zinggeler M, Brasier N, Baerenfaller K, Brenzikofer T, Baltzer L, Vogler C, et al. (2019) Traditional and digital biomarkers: two worlds apart? *Digit Biomark* **3**:92–102.
Badawy R, Hameed F, Bataille L, Little MA, Claes K, Saria S, Cedarbaum JM, Stephenson D, Neville J, Maetzler W, et al. (2019) Metadata concepts for advancing the use of digital health technologies in clinical research. *Digit Biomark* **3**:116–132.
Bakker JP, Goldsack JC, Clarke M, Coravos A, Geoghegan C, Godfrey A, Heasley MG, Karlin DR, Manta C, Peterson B, et al. (2019) A systematic review of feasibility studies promoting the use of mobile technologies in clinical research. *NPJ Digit Med* **2**:47.
Ben-Menahem SM, Nistor-Gallo R, Macia G, von Krogh G, and Goldhahn J (2020) How the new European regulation on medical devices will affect innovation. *Nat Biomed Eng* **4**:585–590.
Boehme P, Hansen A, Roubenoff R, Scheeren J, Herrmann M, Mondritzki T, Ehlers J, and Truebel H (2019) How soon will digital endpoints become a cornerstone for future drug development? *Drug Discov Today* **24**:16–19.
Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, Doerr M, Pratap A, Wilbanks J, Dorsey ER, et al. (2016) The mPower study, Parkinson disease mobile data collected using ResearchKit. *Sci Data* **3**:160011.
Burnham JP, Lu C, Yaeger LH, Bailey TC, and Kollef MH (2018) Using wearable technology to predict health outcomes: a literature review. *J Am Med Inform Assoc* **25**:1221–1227.
Buyse M, Squifflet P, Coart E, Quinaux E, Punt CJA, and Saad ED (2017) The impact of data errors on the outcome of randomized clinical trials. *Clin Trials* **14**:499–506.
Byrom B and Rowe DA (2016) Measuring free-living physical activity in COPD patients: deriving methodology standards for clinical trials through a review of research studies. *Contemp Clin Trials* **47**:172–184.
Carranza Rosenzweig JR, Edwards L, Lincourt W, Dorinsky P, and ZuWallack RL (2004) The relationship between health-related quality of life, lung function and daily symptoms in patients with persistent asthma. *Respir Med* **98**:1157–1165.
Cerreta F, Ritzhaupt A, Metcalfe T, Askin S, Duarte J, Berntgen M, and Vamvakas S (2020) Digital technologies for medicines: shaping a framework for success. *Nat Rev Drug Discov* DOI: 10.1038/d41573-020-00080-6 [published ahead of print].
Chan CB and Ryan DA (2009) Assessing the effects of weather conditions on physical activity participation using objective measures. *Int J Environ Res Public Health* **6**:2639–2654.
Chasse R, Coravos A, and Goldsack JC (2019) The Digital Medicine Society (DiMe): advancing the use of digital medicine to optimize health. *Innov Clin Neur* 2019;**16**: S5-S19.2158-8333
Chen YJ, Chiou CM, Huang YW, Tu PW, Lee YC, and Chien CH (2018) A comparative study of medical device regulations: US, Europe, Canada, and Taiwan. *Ther Innov Regul Sci* **52**:62–69.
Cohen AF, Burggraaf J, van Gerven JM, Moerland M, and Groeneveld GJ (2015) The use of biomarkers in human pharmacology (Phase I) studies. *Annu Rev Pharmacol Toxicol* **55**:55–74.
Committee for Medicinal Products for Human Use (CHMP) (2019) Qualification opinion on stride velocity 95th centile as a secondary endpoint in Duchenne Muscular Dystrophy measured by a valid and suitable wearable device. European Medicines Agency; EMA/CHMP/SAWP/178058/2019.
Coran P, Goldsack JC, Grandinetti CA, Bakker JP, Bolognese M, Dorsey ER, Vasisht K, Amdur A, Dell C, Helfgott J, et al. (2019) Advancing the use of mobile technologies in clinical trials: recommendations from the clinical trials transformation initiative. *Digit Biomark* **3**:145–154.
Coravos A, Doerr M, Goldsack J, Manta C, Shervey M, Woods B, and Wood WA (2020) Modernizing and designing evaluation frameworks for connected sensor technologies in medicine. *NPJ Digit Med* **3**:37.
Coravos A, Khozin S, and Mandl KD (2019) Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *NPJ Digit Med* **2**:14.
Cummings J, Ward TH, and Dive C (2010) Fit-for-purpose biomarker method validation in anticancer drug development. *Drug Discov Today* **15**:816–825.
Depp CA, Bashem J, Moore RC, Holden JL, Mikhael T, Swendsen J, Harvey PD, and Granholm EL (2019) GPS mobility as a digital biomarker of negative symptoms in schizophrenia: a case control study. *NPJ Digit Med* **2**:108.
European Medicines Agency (2014) Qualification of novel methodologies for drug development: guidance to applicants; EMA/CHMP/SAWP/72894/2008.
Gordon WJ, Landman A, Zhang H, and Bates DW (2020) Beyond validation: getting health apps into clinical practice. *NPJ Digit Med* **3**:14.
Haberkamp M, Moseley J, Athanasiou D, de Andres-Trelles F, Elferink A, Rosa MM, and Magrelli A (2019) European regulators' views on a wearable-derived performance measurement of ambulation for Duchenne muscular dystrophy regulatory trials. *Neuromuscul Disord* **29**:514–516.
Hake AM, Dacks PA, and Arnerić SP; CAMD ICF Working Group (2017) Concise informed consent to increase data and biospecimen access may accelerate innovative Alzheimer's disease treatments. *Alzheimers Dement (N Y)* **3**:536–541.
Izmailova ES, Wagner JA, and Perakslis ED (2018) Wearable devices in clinical trials: hype and hypothesis. *Clin Pharmacol Ther* **104**:42–52.
Jewell NP (2016) Natural history of diseases: statistical designs and issues. *Clin Pharmacol Ther* **100**:353–361.
Kruizinga M, van der Heide N, Nuijsink M, Stuurman R, Cohen A, and Driessen G (2019a) Activity and pulmonary function collected via a non invasive platform differentiate healthy and asthmatic children - Selected abstracts from pharmacology 2019. *Br J Clin Pharmacol* **86**:1182–1183.
Kruizinga MD, Stuurman FE, Groeneveld GJ, and Cohen AF (2019b) The future of clinical trial design: the transition from hard endpoints to value-based endpoints. *Handb Exp Pharmacol* **260**:371–397.
Kruizinga MD, Zuiker RGJA, Sali E, de Kam ML, Doll RJ, Groeneveld GJ, Santen GWE, and Cohen AF (2020) Finding suitable clinical endpoints for a potential treatment of a rare genetic disease: the case of ARID1B. *Neurotherapeutics* DOI: 10.1007/s13311-020-00868-9 [published ahead of print].
Leptak C, Menetski JP, Wagner JA, Aubrecht J, Brady L, Brumfield M, Chin WW, Hoffmann S, Kelloff G, Lavezzari G, et al. (2017) What evidence do we need for biomarker qualification? *Sci Transl Med* **9**:eaal4599.
Lipsmeier F, Taylor KI, Kilchenmann T, Wolf D, Scotland A, Schjodt-Eriksen J, Cheng WY, Fernandez-Garcia I, Siebourg-Polster J, Jin L, et al. (2018) Evaluation of smartphone-based testing to generate exploratory outcome measures in a phase 1 Parkinson's disease clinical trial. *Mov Disord* **33**:1287–1297.
Nature Biotechnology Editorial Team (2019) Getting real with wearable data. *Nat Biotechnol* **37**:331.
O'Connell S, ÓLaighin G, Kelly L, Murphy E, Beirne S, Burke N, Kilgannon O, and Quinlan LR (2016) These shoes are made for walking: sensitivity performance evaluation of commercial activity monitors under the expected conditions and circumstances required to achieve the international daily step goal of 10,000 steps. *PLoS One* **11**:e0154956.
Ólafsdóttir AF, Attvall S, Sandgren U, Dahlqvist S, Pivodic A, Skrtic S, Theodorsson E, and Lind M (2017) A clinical trial of the accuracy and treatment experience of the flash glucose monitor FreeStyle libre in adults with type 1 diabetes. *Diabetes Technol Ther* **19**:164–172.
Panch T, Mattie H, and Celi LA (2019) The "inconvenient truth" about AI in healthcare. *NPJ Digit Med* **2**:77.
Papadopoulos E and Norman L (2018) Request for qualification plan. FDA. DDT COA #000114.
Piatak M, Saag MS, Yang LC, Clark SJ, Kappes JC, Luk KC, Hahn BH, Shaw GM, and Lifson JD (1993) High levels of HIV-1 in plasma during all stages of infection determined by competitive PCR. *Science* **259**:1749–1754.
Porter ME (2010) What is value in health care? *N Engl J Med* **363**:2477–2481.
Pratap A, Neto EC, Snyder P, Stepnowsky C, Elhadad N, Grant D, Mohebbi MH, Mooney S, Suver C, Wilbanks J, et al. (2020) Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. *NPJ Digit Med* **3**:21.
Schrack JA, Cooper R, Koster A, Shiroma EJ, Murabito JM, Rejeski WJ, Ferrucci L, and Harris TB (2016) Assessing daily physical activity in older adults: unraveling the complexity of monitors, measures, and methods. *J Gerontol A Biol Sci Med Sci* **71**:1039–1048.
Steele BG, Holt L, Belza B, Ferris S, Lakshminaryan S, and Buchner DM (2000) Quantitating physical activity in COPD using a triaxial accelerometer. *Chest* **117**: 1359–1367.
Steinhubl SR, McGovern P, Dylan J, and Topol EJ (2017) The digitised clinical trial. *Lancet* **390**:2135.
U.S. Food and Drug Administration (2017) Use of Electronic Records and Electronic Signatures in Clinical Investigations Under 21 CFR Part 11 – Questions and Answers: Guidance for Industry. FDA-2017-D-1105.
U.S. Food and Drug Administration (2018a) Biomarker Qualification: Evidentiary Framework Guidance for Industry and FDA Staff [DRAFT GUIDANCE]. FDA. FDA-2018-D-4267.
Zhan A, Mohan S, Tarolli C, Schneider RB, Adams JL, Sharma S, Elson MJ, Spear KL, Glidden AM, Little MA, et al. (2018) Using smartphones and machine learning to quantify Parkinson disease severity: the mobile Parkinson disease score. *JAMA Neurol* **75**:876–880.