Original Article

DIA

Therapeutic Innovation & Regulatory Science
1-8

DOI: 10.1177/2168479018772569
tirs.sagepub.com

Developing and Implementing Performance Outcome Assessments: Evidentiary, Methodologic, and Operational Considerations

Elizabeth Richardson, MSc[1], Jessica Burnell, MPA[1], Heather R. Adams, PhD[2], Richard W. Bohannon, EdD, DPT[3], Elizabeth Nicole Bush, MHS[4], Michelle Campbell, PhD[5], Wen Hung Chen, PhD[5], Stephen Joel Coons, PhD[6], Elektra Papadopoulos, MD, MPH[5], Bryce R. Reeve, PhD[7], Daniel Rooks, PhD[8], and Gregory Daniel, MPH, PhD[1]

Abstract

The use of performance outcome (PerfO) assessments to measure cognitive or physical function in drug trials presents several challenges for both sponsors and regulators, owing in part to a relative lack of scientific guidance on their development, implementation, and interpretation. In December 2016, the Duke-Margolis Center for Health Policy convened a 2-day workshop to explore the evidentiary, methodologic, and operational challenges associated with PerfO measures, and to identify potential paths to addressing these challenges. This paper presents both a summary of the discussion as well as additional input from a working group of experts from FDA, industry, academia, and public-private consortia. It is intended to advance the discussion around the development and use of PerfO measures to assess patient functioning in clinical trials intended to support registration of new treatments, and to highlight the key gaps in knowledge where additional research, collaboration, and discussion are needed.

Keywords

clinical outcome assessment, performance outcome assessments, concept of interest, context of use

Introduction

An important component of the move toward patient-focused drug development is the successful implementation of fit-for-purpose clinical outcome assessments (COAs) (see Table 1) that can be used to obtain valid, reliable, and meaningful endpoints in populations of interest. COAs measure outcomes that describe or reflect how an individual feels, functions or survives.[1] The US Food and Drug Administration (FDA) published guidance for one type of COA—patient-reported outcome (PRO) measures—in 2009, and while many of the principles outlined in that guidance can generally be applied to other types of COAs, how these principles are applied may vary across the types of COAs. Performance outcome (PerfO) assessments—which measure concepts such as cognitive or physical function—present particular challenges, owing in part to a relative lack of scientific guidance.

In December 2016, the Duke–Margolis Center for Health Policy convened a 2-day workshop to explore the evidentiary, methodologic, and operational challenges associated with PerfO measures, and to identify potential paths to addressing these challenges. This article presents both a summary of the discussion held over those 2 days and additional input from a working group of experts from FDA, industry, academia, and public-private consortia. It is intended to advance the discussion around the development and use of PerfO measures[i] to assess patient functioning in clinical trials intended to support registration of new treatments, and to highlight the key gaps in

[1] Duke–Robert J. Margolis Center for Health Policy, Washington, DC, USA
[2] University of Rochester Medical Center, Rochester, NY, USA
[3] Campbell University, Buies Creek, NC, USA
[4] Eli Lilly and Company, Indianapolis, IN, USA
[5] US Food and Drug Administration, Silver Spring, MD, USA
[6] Critical Path Institute, Tucson, AZ, USA
[7] Department of Population Health Sciences, Duke University, Durham, NC, USA
[8] Novartis Institutes for BioMedical Research, Boston, MA, USA

Submitted 12-Oct-2017; accepted 30-Mar-2018

Corresponding Author:
Elizabeth Richardson, MSc, Duke–Robert J. Margolis Center for Health Policy, 1201 Pennsylvania Ave NW, Suite 500, Washington, DC 20009, USA.
Email: elizabeth.richardson@duke.edu

Table 1. Clinical Outcome Assessments: Definitions of Key Terms Adapted From the BEST Glossary.

Patient-reported outcome (PRO) assessment: A measurement based on a report that comes directly from the patient (ie, study subject) about the status of a patient's health condition without amendment or interpretation of the patient's response by a clinician or anyone else. A PRO can be measured by self-report or by interview provided that the interviewer records only the patient's response. Symptoms or other unobservable concepts known only to the patient can only be measured by PRO measures. PRO measures can also assess the patient perspective on functioning or activities that may also be observable by others.
Clinician-reported outcome (ClinRO) assessment: A measurement based on a report that comes from a trained health-care professional after observation of a patient's health condition. Most ClinRO measures involve a clinical judgment or interpretation of the observable signs, behaviors, or other manifestations related to a disease or condition. ClinRO measures cannot directly assess symptoms that are known only to the patient.
Observer-reported outcome (ObsRO) assessment: A measurement based on a report of observable signs, events, or behaviors related to a patient's health condition by someone other than the patient or a health professional. Generally, ObsROs are reported by a parent, caregiver, or someone who observes the patient in daily life and are particularly useful for patients who cannot report for themselves (eg, infants or individuals who are cognitively impaired). An ObsRO measure does not include medical judgment or interpretation.
Performance outcome (PerfO) assessment[a]*:* A measurement based on a task(s) performed by a patient according to instructions that is administered by a health care professional. PerfO assessments require patient cooperation and motivation.

[a] The working group is proposing an alternative definition for performance outcome assessments. See Table 2.

Table 2. New Proposed Definition of Performance Outcome Assessments.

Following discussions at the December 2016 workshop, the authors jointly developed and are proposing a new working definition of performance outcome assessments:
A measurement based on a standardized task performed by a patient that is administered and evaluated by an appropriately trained individual or is independently completed

knowledge where additional research, collaboration, and discussion are needed. Other uses of PerfO measures (eg, use in clinical practice) are outside the scope of this paper (Table 2).

Background

Most clinical studies submitted to FDA to support registration of a drug utilize at least one COA, either as part of the primary or secondary endpoint definition to evaluate the effect a treatment has on how a patient feels or functions, or as an exploratory endpoint measure that informs future research.[2] The selection of a particular COA for use in a study depends on the concept of interest (eg, pain intensity or frequency) and the context of use in which the measure will be applied (ie, the key study aspects that can impact the adequacy of the measurement, such as the disease, patient population, method of administration, and frequency and timing of assessment).

COAs are generally divided into 4 broad categories according to how the measure is conducted and reported: patient-reported outcome (PRO) measures, clinician-reported outcome (ClinRO) measures, observer-reported outcome (ObsRO) measures, and performance outcome (PerfO) measures. The selection of a COA for use in a trial begins with an understanding of the outcome(s) that is/are most relevant and meaningful to patients with the target condition. The type of COA to select will depend on the context of use and the concept of interest that will be measured, and more than one COA may be appropriate.

PerfO measures may assess an array of patient functions including physical, cognitive, or perceptual/sensory function through tasks completed by the patient. These measures may consist of completing one or a series of standardized tasks in order to assess the function or functions of interest. The patient's performance on these tasks is then quantified and reported using defined procedures. Because PerfO measures are typically assessed in a clinical or research setting evaluating the performance of the subjects on a set of standardized tasks, they may be particularly attractive for use in multicenter trials for standardizing assessment, as they can reduce the variability of the daily function and activities performed by the subjects in their natural environment and may also enable direct comparison on specific standardized tasks.[ii] The use of a PerfO measure may also overcome some limitations of PRO measures, such as patients' limited ability to accurately recall their daily functioning or to assess their functional abilities. The latter may result if a patient's self-report of ability (ie, his or her perceived ability) differs from his or her actual ability to perform a particular task. Patients may also differ in the daily activities they routinely perform, resulting in challenges with PRO items querying about activities that not all patients perform in daily life (eg, climbing stairs).

PerfO measures have been used successfully to support the regulatory approval of treatments in several therapeutic areas. Examples of PerfO assessments include measures of memory (eg, word recall test) in the context of a drug intended to improve memory, and gait speed (eg, timed 25-foot walk test) in the context of a drug intended to improve mobility.[iii] However, there are unique challenges associated with this particular type of COA. For example, there remain a number of unanswered questions regarding how best to establish the validity of a PerfO measure, what level of evidence is necessary to support that validity (including whether what is being measured is translatable to an important aspect of patient functioning), and whether a PerfO measure retains its validity for different patient populations (eg, pediatrics vs adults vs older adults). Additionally, the considerations for evaluating measurement properties may vary depending upon the type of PerfO measure (eg, measures of cognitive or physical function).

One of the key challenges in developing and implementing COA instruments more generally, but which may be particularly

difficult for PerfO measures, is determining the degree of within-patient change—that is, individual-level change from a baseline assessment to a postintervention assessment—on a given measure that is considered clinically important and meaningful, rather than simply statistically significant. Anchor-based methods appear to have considerable support for estimating meaningful change, but there is no established consensus on the ideal approach.

Other major challenges arise when using a PerfO measure in multinational clinical trials with culturally diverse populations. Obtaining culturally appropriate and accurate translations of an instrument often involves specific considerations. It is also important to ensure that the PerfO measure task or tasks are administered in a standardized manner to minimize interscorer variability, regardless of country or site. Standardizing PerfO measures typically requires that the equipment or other materials used in the execution of tasks is consistent across study sites. In some cultures, the specific tasks being assessed by a PerfO measure may be less relevant or meaningful in some populations, which can impact the interpretation and value of the resulting data in a global trial. However, the issue of translatability across cultural groups may be less of a limitation with PerfO measures in comparison with PRO measures, as specific activities may differ across cultures while the component functions (eg, mobility, recall) are often more generalizable.

Determining Whether a PerfO Measure Is Fit-for-Purpose: Major Considerations

Concept of Interest and Context of Use

Developing a well-defined and psychometrically appropriate PerfO measure begins with identifying and clearly defining the target concept of interest (ie, the aspect of an individual's clinical, biological, physical, or functional state or experience that the assessment is intended to capture [or reflect]), and determining if a PerfO assessment is the most appropriate type of COA to capture that concept. The concept of interest might include something like usual walking speed or muscle strength (for PerfO measures assessing physical function) or memory recall (for PerfO measures assessing cognitive function). For any PerfO measure, the manner in which the target concept of interest relates to relevant and important aspects of the patient's functional impairment associated with the condition, and/or how the concept of interest informs the understanding of the underlying disease state, should be clear.

The task(s) of the PerfO measure should also be clearly connected to the concept of interest. The interpretation of the result of the PerfO measure should be able to reflect an important aspect of the patient's functioning, which is best achieved when both the tasks and testing conditions reflect the demands of the patient's day-to-day activities as closely as possible. This approach aligns with the World Health Organization (WHO) International Classification of Functioning, Disability, and Health (ICF), a multidimensional framework that considers not only the person's intrinsic "functioning at the level of the body," but also the impact of impairments upon abilities and participation in life activities, and potential environmental factors that may aid or impede one's participation.[3] For cognitive PerfO measures, it is necessary to differentiate between concepts of interest that involve cognition (eg, processing speed, working memory) and concepts of interest that involve cognition-dependent behavior (eg, instrumental activities of daily living). Additional guidance regarding identification of the concept of interest and context of use for PerfO measures can be found in the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) task force report on clinical outcome assessments.[2]

As no measure will be able to fully capture every concept that matters within a particular therapeutic context, PerfO measures should be designed to assess the most important concept within the targeted context of use.

Evidence for Validity

Validity refers to the extent to which the PerfO measure assesses the concept of interest, and involves establishing both the content validity and the construct validity of the measure.[4] Establishing content validity involves gathering evidence that demonstrates that the tasks and domains of a measure are both appropriate and comprehensive with regard to the concept of interest, target population, and intended use.[1] Content validity is informed by qualitative research (eg, literature review, clinician input, patient input, and caregiver input) into what matters most to patients with the disease or condition. For example, a PerfO measure assessing walking speed will not be a content valid measure as an endpoint of upper body functioning since dexterity and strength of upper arms are what matter in this situation.

Although PerfO measures can involve tasks that clearly relate to or resemble some aspect of a patient's daily functioning (eg, the timed 25-foot walk test to assess gait speed), sometimes the link between the tasks and the outcome being studied is not obvious. This is particularly true of PerfO measures assessing cognition, which may be designed to capture complex underlying cognitive processes that are not as obviously linked to real-world functioning (eg, neuropsychological tests assessing certain cognitive abilities that underpin the performance of many varied mental tasks). The type and/or level of evidence used to establish the content validity of a PerfO measure that is indirectly linked to real-world functioning may be different from that used for PerfO measures where the link to real-world functioning is more direct and translatable.

Establishing construct validity involves using quantitative methods to assess the extent to which the PerfO measure's scores conform to a priori hypotheses concerning logical relationships that should exist with other measures or patient characteristics (eg, disease severity). Additional objectives that the quantitative methods can achieve are to assess the reliability of the PerfO measure, its ability to detect change, and the

meaningful change score. Methods for assessing reliability and ability to detect change are well-established and published for other types of COAs, and these methods are also applicable to PerfO measures. Approaches to determining the amount of change on a PerfO measure that is meaningful to patients is discussed later in this paper.

Additional Considerations in Choosing a PerfO Measure

FDA encourages the use of existing measures where applicable rather than developing one de novo. This may be particularly important when evaluating children, where existing standardized measures of children's cognitive, physical, academic, or adaptive skills are designed to assess the wide variability in function by age. In some cases, an existing measure can be used as is, but sometimes a measure will require modification to ensure it is fit for the relevant context of use. Determining when it is appropriate to use an existing measure in its current iteration, modify it, or develop an entirely new measure can be difficult. Additionally, some existing measures, depending on their age, may require updating so that item content and norm-referenced data are representative of current populations. Consider, for example, the hypothetical case of a PerfO measure that asks patients to name or describe the use of outmoded objects such as a rotary phone or an audio cassette.

Additional tools or best practices are needed to aid stakeholders (eg, instrument developers, researchers, clinical trial sponsors, and patients) in evaluating the appropriateness of such measures, and how they might be modified for a particular context of use. It may be useful, for example, to develop a "checklist" that can help stakeholders determine when and to what extent a measure should be modified for a particular context of use. FDA has created a "Wheel and Spokes" diagram to guide stakeholders through the process of designing, testing, and modifying a COA, and it would be helpful to develop a modified version of this diagram that is specific to PerfO measures and can be used to guide stakeholders in both the creation of new PerfO measures and the modification of existing instruments.[5]

Patient, caregiver, and test administrator (or "administrator") burden is another important consideration. PerfO measures that are too onerous, stressful, or painful for patients to complete or that are too time-consuming, difficult, or complex for study staff to administer and/or score can discourage uptake and risk compromising data quality. The feasibility of the assessment including the timing of administration, the training and equipment required, and duration of the test should be considered to determine whether a PerfO measure will be appropriate within a particular context of use.

Thorough documentation of the PerfO assessment's design and measurement underpinnings is also critical to both the development of new measures and the use of existing measures. This includes information on how the measure was developed, what stakeholders were involved in the development, qualitative and quantitative methods used to evaluate the measure, psychometric evidence for the measure (eg, reliability and validity), the availability of training and instruction guides and a scoring manual, and a description of how any normative data, if available, were generated. FDA has developed several tools for consultation to provide further guidance, including the Roadmap to Patient-Focused Outcome Measurement, which outlines how stakeholders can identify the concept of interest, define the specific context of use in which the measure will be used, and identify the appropriate type of COA.[6]

Engaging Stakeholders in the PerfO Measure Development Process

Similar to the PRO measure development process, a wide range of stakeholders need to be involved in the development, adoption, or modification of a PerfO measure. These stakeholders can include, but are not limited to, patients, caregivers, clinical trial sponsors, health care providers, payers, disease experts, regulators, advocacy groups, measurement specialists, and experts in the concept or construct being measured. The appropriate stakeholders with which to engage will depend in part on the disease area being studied as well as the measure itself. For example, diseases that affect cognitive function may require certain stakeholders, such as caregivers, to play more of a role than they would in the development of a PerfO measure for a physical condition. The potential role for stakeholders includes, but is not limited to, the following.

Concept elicitation

Where possible, it is critical to elicit the concepts that are most important to patients (and sometimes caregivers) in a given disease area, and to consider that information when selecting or creating a PerfO measure for a given context of use. For example, patients could provide input on the areas of physical or cognitive function where they experience challenges, such as walking across the street, rising from a chair or toilet, climbing stairs, remembering to take medication, or keeping an appointment. The level of input that various stakeholders should provide in shaping what PerfO measures assess and how they assess it varies depending on the condition and what is to be measured. For example, patients with serious cognitive impairments may not be able to meaningfully contribute to this stage of measure development, but clinical experts or caregivers may be able to provide important insight or feedback on meaningful concepts.

Item/task generation, selection, or modification

From the prioritized list reported by patients and other stakeholders obtained through concept elicitation, tasks that would be used to most directly assess those concepts can be created, selected, or modified to assess functions or activities that are clinically relevant. However, this approach may overlook some tasks that assess useful concepts but are not reported by patients and other stakeholders, as they are not tasks that we normally perform in our daily lives (e.g., as in certain neuropsychological

tests). Furthermore, the generation of items or tasks is not a linear process; as modifications to a measure are made, stakeholders should be included to assess relevance to daily life, feasibility of administration, and appropriateness of the measure with respect to the anticipated context of use.

Ensuring Patient and Administrator Understanding of the PerfO Measure

It is critical that both patients and administrators are able to understand what is required to perform the PerfO measure tasks, and that the measure is administered in a consistent manner, both within and among patients and administrators, as well as across repeated assessments. The following strategies can help stakeholders detect and resolve problems in patient and administrator understanding.

Patient-focused strategies

Pilot testing the PerfO measure with patients in the target population can help identify any aspects of the measure that may need to be altered to increase patient understanding and engaged participation in the required task(s). Cognitive interviewing can also be used to find out from patients directly whether the instructions and task(s) are clear and easy to understand. Determining whether a patient understands the instructions may be more challenging when studying conditions that affect cognitive function, and specific patient populations may require additional help in understanding or completing the task(s). Stakeholders should consider additional strategies to address these issues. For example, the test should be designed in such a way to ensure that the patient understands how to complete the task; considerations include the physical "look and feel" of the test materials and the standardized task instructions presented by the administrator.

Another feature that is common to PerfO measures is the "practice effect," wherein patients improve their performance after repeated exposure to the same tasks. There are several approaches to addressing this effect, such as using alternate forms of a task or administering the PerfO measure multiple times during a run-in period prior to baseline assessment. It is also possible to statistically adjust for the practice effect, for example, by estimating the practice effect on a control group and then using the resulting data to statistically adjust the scores for the treatment group. Regardless of the approach selected, it should be well-justified in the trial design, documented, and implemented consistently across a trial.

Administrator-focused strategies

Developing clear, comprehensive, and standardized instructions for all administrators is key to making sure the PerfO measure will be implemented properly in clinical trials, particularly across multiple study sites. Instructions should include guidance on both administration and data collection, including how to maintain the fidelity of the data being collected. Administrators need to undergo training in how to administer the PerfO measure before working with patients; however, the level of training will depend on the complexity and nature of the PerfO measure and the administrator's experience. In some cases, administrators may need to undergo extensive training before being able to administer the PerfO measure in a trial. The level of education required of administrators may also differ depending on the nature and complexity of the PerfO measure, with some needing administrators to hold an advanced degree while others may require less formal academic training. In some instances, an advanced degree is not required by the administrator but should be held by the individual who implements training and/or ongoing supervision of the PerfO assessment during the trial.

Once a PerfO assessment is implemented in a clinical trial, regular monitoring and periodic re-training of administrators should also be considered to ensure that implementation of the PerfO measure is consistent over time. If videotaping (or audio recording) is used during the trial to record patient performance, the recording should be unobtrusive and used consistently throughout the trial, so as to improve patient and administrator comfort and mitigate bias. When PerfO measures are used to assess a patient's fitness for participation in a clinical trial, administrators should be blinded to the trial entry criteria so that it does not influence administration or scoring.

Applying PerfO Measures in Differing Populations

A number of factors can affect whether a PerfO measure is appropriate for a given population. As noted above, pediatric populations, in particular, require special consideration. Physical and cognitive skills in young children develop rapidly across relatively narrow age ranges (in some cases only several months: eg, the change in gross motor skill from sitting, crawling, standing, to walking, or the transition from babbling to language acquisition). Stakeholders must be careful not to assume that all children in a desired population will have the skills necessary to complete the PerfO measure if a wide age range is being targeted for the trial. Though (as noted above), assessment of pediatric populations may be best conducted with existing standardized measures that capture a wide range of development, a traditional approach of comparing patients' performance to age-referenced normative data may be challenging in special subpopulations. For example, in patients with pediatric neurodegenerative conditions, existing standardized assessments may be subject to floor effects, which occur when the measure has insufficient range at its "low" end (eg, at the level of greater impairment or limitations in abilities) to measure function with adequate detail or sensitivity.[7] Further discussion of special considerations for development of outcome assessments in rare diseases was covered in a workshop convened in 2015 by the FDA, "Assessing Neurocognitive Outcomes in Inborn Errors of Metabolism."[8]

It is often helpful to have observational or natural history data that can provide more detailed insights into how symptoms and functional impacts of a disease affect patients and change over time. This is particularly true in rare diseases or diseases with

slow progression, where the body of evidence on disease or symptom progression and impact on daily life functioning is less robust. Gaining a better longitudinal sense of a disease can allow concepts of interest that are relevant for heterogeneous patient populations—including populations in different stages of the same disease—to be targeted more accurately.

Cross-cultural adaptation should also be considered in the development or modification of a PerfO measure. Although such adaptation of an instrument can be time consuming, this burden can be mitigated through appropriate upfront planning. This begins with the chosen concept of interest, which may have varying degrees of cross-cultural equivalence. Stakeholders should then attempt to anticipate the regions or countries in which the PerfO measure may be used for multinational studies, and design or adapt the measure to include tasks that have high relevance across the majority of those regions or countries.

Cross-cultural adaptation will also require detailed and standardized training for administrators to ensure the instructions are appropriately communicated across sites. This goes beyond simple translation of written instructions. In some situations, training via video instruction may be more useful and will also require cross-cultural adaptation. For example, PerfO measure instructions may require an administrator to demonstrate to patients how to perform the tasks in the way that is appropriate for that culture (eg, when assessing the ability to stand up from sitting position, in Japan, it may be necessary to assess both standing up from sitting in a chair as well as from sitting on the floor, as sitting on a flooring material called tatami is very common). The type of PerfO measure may also affect its cross-cultural adaptability. For example, a PerfO measure that assesses cognition-dependent function (eg, preparing a shopping list) may be more strongly influenced by culture differences than a PerfO measure that assesses physical function (eg, tests of walking speed). Additionally, differences in gender roles and norms can also impact the appropriateness of a PerfO assessment, as certain tasks (eg, cooking, laundry, yardwork, or driving) may be more gendered in some regions than in others.

Where possible, stakeholders should also strive to include people in the development process who are native to the cultures where the PerfO measure will be used. This can allow local cultural differences to be identified and addressed as early as possible. In some cases, it may be preferable to avoid sites where cultural issues may prove challenging; however, this decision will reduce the generalizability of study findings. Stakeholders should consult resources such as the ISPOR task force report on COAs for further guidance on developing PerfO measures for differing populations.[2]

Implementing PerfO Measures in Clinical Trials

Administration and Scoring of a PerfO Measure

Trial investigators should carefully consider how best to administer the measure and calculate the resulting score. Although PerfO measures are typically administered in the presence of the patient, remote or even patient self-administration may be possible in some cases, which can help to reduce some of the burdens associated with bringing patients and/or caregivers to a particular site. However, remote or self-administration approaches present their own challenges in assessment of motor or cognitive task performance. If the environment in which the PerfO measure is administered influences the outcome (eg, the patient's home vs a clinic) or influences the conduct of the PerfO measure (eg, insufficient space to perform 6-minute walk test, but adequate for 4 meter gait speed), the quality or consistency of the data being collected may be compromised. In addition, patients may not follow standardized PerfO measure instructions exactly as they are intended and may not be as motivated when performing the test on their own. Finally, some physical PerfO measures may require direct (non-remote) assessment in order to ensure patient safety. Certain patient populations, particularly children and older adults, may also need guidance to stay engaged with and complete the tasks. Additionally, self-administration of PerfO measures may require the use of digital technology to measure and record task performance and to communicate with the administrator, which should be considered and evaluated carefully during PerfO measure development. Older populations or those with cognitive impairment in particular may not be familiar with using a device.

Stakeholders may also consider a strategy that combines a remote administrator with local assistance for the patient. An administrator may give the patient instructions and collect data remotely, while a local assistant works directly with the patient to ensure that the assessment takes place in a controlled environment, to keep patients on task, or to facilitate technology use for those unfamiliar with it.

PerfO measures may be scored using either central or on-site scorers, and each approach presents advantages and disadvantages. Centralized scorers can mitigate the variation associated with multiple on-site scorers and provide more consistent scoring across sites. If a patient's performance on a PerfO measure requires experience and skill to assess, relying on centralized scorers reduces the need to ensure such expertise at every administration site. For rare diseases in particular, there may be few individuals with expertise in the disease or its assessment, which can make centralized scoring by carefully selected scorers particularly useful.[9] Conversely, some aspects of functioning may require in-person assessment by a trained administrator (eg, handling of physical test materials such as blocks or puzzle pieces). It is also possible to have both an on-site scorer and centralized scorers who can provide a second score or ensure quality control across on-site scorers. The question of when central administration or scoring should be employed within a trial requires additional research and may very well depend on the disease or condition, the target patient population, the specific PerfO measure, and the study design.

The Role of Technology

Regardless of how the PerfO measure is administered or scored, the technology used to support its implementation should be carefully considered upfront as part of the measure development process, as this will help to enhance both its reliability and validity, as well as the feasibility of its implementation. Where feasible, stakeholders may want to ensure that backup methods of administration are available.

While newer technologies, such as tablets or specialized mobile apps, may be used to support PerfO measure administration, questions remain regarding both the ability of legacy measures to be adapted to these new technologies, as well as the additional research that will be needed to establish equivalence between the methods of administration for legacy measures. Whatever platform is used, the method of administration must effectively assess the concept of interest within the specified context of use. If more than one method of administration is being used in a trial, measurement comparability or equivalence should be demonstrated to enable the pooling of trial data from the different data collection methods. When possible, mixed modes of administration should be avoided within a trial. Stakeholders should also consider the broader implementation issues associated with newer technologies, as different research sites or patient populations may have varying access to the Internet and/or familiarity and comfort with these technologies, which would introduce additional costs and special training requirements.

There is also significant interest among stakeholders regarding the use of mobile technologies such as wearable devices (eg, activity trackers) in clinical trials, either to record or measure patient performance of defined PerfO assessment tasks, or to collect data passively as patients go about their day-to-day activities. Consideration of wearable devices to collect performance-based clinical trial endpoint data is becoming more feasible as technology is improved. Technology-based data collection is anticipated to become a key part of conducting clinical trials in the future, and will become easier to integrate into trials as we gain more experience with their use in clinical trials, including learning how best to select appropriate technology-based endpoints as well as interpret the meaningfulness of the data. PerfO measures and wearables each have distinct characteristics that set them apart and complicate the alignment between the data that can be collected by the wearable versus the data needed to derive the endpoint. While it may be possible to use such technology to collect PerfO data as patients are going about their daily activities, this is an area that requires additional research and discussion among stakeholders, including with health authorities.

Establishing Meaningful Within-Patient Change

The derivation and interpretation of thresholds for meaningful change on COA measures, particularly PerfO measures, can be challenging. One factor contributing to this challenge is the variability among patients regarding what constitutes a meaningful change. A small improvement in function—or even maintenance of baseline function—may be very meaningful for some patients but not for others.

Normative data can be useful in guiding interpretation of the PerfO measure by providing benchmarks of performance of a given reference sample. This is particularly informative for studies in very young and very old populations, where measures that have age-equivalent scores can be useful for understanding a patient's relative/comparative development or status. However, establishing normative data becomes quite challenging for diseases with many phenotypes, and may be less appropriate in certain disease contexts or populations. It is also important to consider whether the normative data should be generated from a healthy (nonpatient) population, from among patients with the same condition with various levels of severity, or both.

In addition to these and other issues that are present across all types of COAs, PerfO measures present specific challenges. For example, some PerfO measures assess concepts or abilities for which the relationship to daily functioning may not be obvious (eg, some neuropsychological tests), and it can be difficult in such cases for patients to determine the level of change that would be important. In other cases, patients may lack the capacity to provide meaningful input owing to their age or level of function, and it may be necessary to seek input from caregivers, other observers, or clinical experts. The level of patient, caregiver, or expert input required to derive and interpret thresholds for meaningful change may depend on the individual PerfO measure, its concept of interest and context of use, and the population in which it will be used.

As with PRO measures, anchor-based methods may be used to derive thresholds for interpreting meaningful change supportive with cumulative distribution functions (CDFs) and distribution-based methods. Other types of COAs, as appropriate, might be used as anchors for PerfO measures. For example, a PRO measure that assesses activities of daily living dependent on mobility may be used as an anchor for a walking test. Other emerging methods may also be considered in certain contexts. However, this is an ongoing question regarding in what contexts each of these established or emerging methods is most appropriate. In general, it is preferable to use more than one method to derive and interpret the threshold for meaningful within-patient change and to use the totality of those results to interpret findings.

Next Steps

Stakeholders have a number of opportunities to advance the understanding and increase the use of PerfO measures in the clinical setting, in multinational trials, and in the regulatory review process. However, many aspects of the PerfO measure development, adoption, and modification processes require additional work and should be prioritized for discussion and consensus-building. First, more information around data

collection methods is necessary to ensure the quality and consistency of the resulting PerfO data. A greater understanding of how to integrate PerfO measures into clinical trials and better standardize their administration, as well as when to use a PerfO measure alone or in conjunction with another type of COA, is also needed. Additionally, unanswered questions remain regarding whether the definition of PerfO measures should be amended or expanded, including the contexts in which wearable devices should be considered PerfO measures, or whether wearables could work alongside a PerfO measure to provide additional data.

Additional discussion and consensus-building is also needed on the development and application of "personalized" COAs, including PerfO measures. Under personalized COAs, the measure may vary across patients in an effort to capture the most important and relevant signs, symptoms, or functional impairment in each individual. In the context of PerfO assessments, such an approach could help developers, providers, and patients address the challenges related to heterogeneity in functioning among a given patient population. However, much work remains to advance the discussion in this area.

Stakeholders have also expressed a desire for additional FDA guidance on the application of PerfO measures in clinical trials, particularly for the purposes of supporting approval and labeling. Further refinement of a PerfO measure-specific "Wheel and Spokes" diagram would also be a useful next step for the Agency to pursue in consultation with stakeholders. Additionally, overarching guidance on more general issues of measurement—such as how to determine the most appropriate type of COA for a given concept of interest—would be beneficial not just for the development of PerfO measures but for all types of COA tools.

Author Note

This paper reflects the perspectives of the individual authors and should not be construed to represent official views or policies of the US Food and Drug Administration.

Declaration of Conflicting Interests

No potential conflicts were declared.

Funding

Funding for this workshop and resulting white paper was provided under a cooperative agreement between the Duke–Robert J. Margolis Center for Health Policy and the US Food and Drug Administration.

ORCID iD

Elizabeth Richardson, MSc http://orcid.org/0000-0001-8516-6672

Notes

i. Throughout this paper, the terms "measures," "assessments," and "instruments" are used interchangeably unless otherwise specified.
ii. Ibid.
iii. See the FDA's Clinical Outcome Assessment Compendium for a select list of measures that have been used in clinical trials to support labeling claims.

References

1. FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource. Silver Spring, MD: Food and Drug Administration; Bethesda, MD: National Institutes of Health; 2016. https://www.ncbi.nlm.nih.gov/books/NBK338448/. Accessed September 20, 2017.
2. Walton MK, Powers JH, Hobart J, et al. Clinical outcome assessments: conceptual foundation—report of the ISPOR Clinical Outcomes Assessment—Emerging Good Practices for Outcomes Research Task Force. *Value Health*. 2015;18:741-752.
3. World Health Organization. *International Classification of Functioning, Disability and Health (ICF)*. Geneva: World Health Organization; 2001.
4. United States Food and Drug Administration Guidance for Industry. Patient-reported outcome measures: use in medical product development to support labeling claims. https://www.fda.gov/downloads/drugs/guidances/ucm193282.pdf. Published 2009. Accessed September 20, 2017.
5. United States Food and Drug Administration. Wheel and spokes diagram: clinical outcome assessments. https://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/UCM370175.pdf. Accessed September 20, 2017.
6. United States Food and Drug Administration. Roadmap to patient-focused outcome measurement in clinical trials. https://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/UCM370174.pdf. Accessed September 20, 2017.
7. Delaney KA, Rudser KR, Yund BD, Whitley CB, Haslett PA, Shapiro EG. Methods of neurodevelopmental assessment in children with neurodegenerative disease: Sanfilippo syndrome. *JIMD Rep*. 2014;13:129-137.
8. Shapiro E, Bernstein J, Adams HR, et al. Neurocognitive clinical outcome assessments for inborn errors of metabolism and other rare conditions. *Mol Genet Metab*. 2016;118:65-69.
9. Kobak KA, Engelhardt N, Williams JBW, Lipsitz JD. Rater training in multicenter clinical trials: issues and recommendations. *J Clin Psychopharm*. 2004;24:113-117.