HHS Public Access

Author manuscript

Value Health. Author manuscript; available in PMC 2017 April 04.

Published in final edited form as:

Value Health. 2017 January ; 20(1): 2–14. doi:10.1016/j.jval.2016.11.005.

Clinician-Reported Outcome Assessments of Treatment Benefit: Report of the ISPOR Clinical Outcome Assessment Emerging Good Practices Task Force

John H. Powers III, MD, FACP, FIDSA[1,*], **Donald L. Patrick, PhD, MSPH**[2], **Marc K. Walton, MD, PhD**[3], **Patrick Marquis, MD, MBA**[4], **Stefan Cano, PhD, CPsychol, AFPBsS**[5], **Jeremy Hobart, PhD, FRCP**[6], **Maria Isaac, MASc, MD, PhD**[7], **Spiros Vamvakas, MD**[8], **Ashley Slagle, MS, PhD**[9], **Elizabeth Molsen, RN**[10], and **Laurie B. Burke, RPh, MPH**[11,12]

[1]George Washington University School of Medicine, Washington, DC, USA

[2]Department of Health Services and Seattle Quality of Life Group, University of Washington, Seattle, WA, USA

[3]Janssen Research & Development, Ashton, MD, USA

[4]Modus Outcomes, Newtown, MA, USA

[5]Modus Outcomes, Letchworth Garden City, Hertfordshire, UK

[6]Peninsula College of Medicine and Dentistry, Plymouth, UK

[7]Institute of Medicine of the National Academy of Science, European Medicines Agency, London, UK

[8]European Medicines Agency, London, UK

[9]Aspen Consulting Services, LLC, Philadelphia, PA, USA

[10]ISPOR, Lawrenceville, NJ, USA

[11]Lora Group, LLC, Royal Oak, MD, USA

[12]University of Maryland School of Pharmacy, Baltimore, MD, USA

Abstract

A clinician-reported outcome (ClinRO) assessment is a type of clinical outcome assessment (COA). ClinRO assessments, like all COAs (patient-reported, observer-reported, or performance outcome assessments), are used to 1) measure patients' health status and 2) define end points that can be interpreted as treatment benefits of medical interventions on how patients feel, function, or survive in clinical trials. Like other COAs, ClinRO assessments can be influenced by human choices, judgment, or motivation. A ClinRO assessment is conducted and reported by a trained health care professional and requires specialized professional training to evaluate the patient's health status. This is the second of two reports by the ISPOR Clinical Outcomes Assessment—Emerging Good Practices for Outcomes Research Task Force. The first report provided an

[*]Address correspondence to: John H. Powers, George Washington University School of Medicine, Penfield House, 15915 Emory Lane, Rockville, MD20853. jpowers3@aol.com.

overview of COAs including definitions important for an understanding of COA measurement practices. This report focuses specifically on issues related to ClinRO assessments. In this report, we define three types of ClinRO assessments (readings, ratings, and clinician global assessments) and describe emerging good measurement practices in their development and evaluation. The good measurement practices include 1) defining the context of use; 2) identifying the concept of interest measured; 3) defining the intended treatment benefit on how patients feel, function, or survive reflected by the ClinRO assessment and evaluating the relationship between that intended treatment benefit and the concept of interest; 4) documenting content validity; 5) evaluating other measurement properties once content validity is established (including intra- and inter-rater reliability); 6) defining study objectives and end point(s) objectives, and defining study end points and placing study end points within the hierarchy of end points; 7) establishing interpretability in trial results; and 8) evaluating operational considerations for the implementation of ClinRO assessments used as end points in clinical trials. Applying good measurement practices to ClinRO assessment development and evaluation will lead to more efficient and accurate measurement of treatment effects. This is important beyond regulatory approval in that it provides evidence for the uptake of new interventions into clinical practice and provides justification to payers for reimbursement on the basis of the clearly demonstrated added value of the new intervention.

Keywords

clinical trials; clinician-reported outcomes; end points; outcome assessments

Introduction

Clinical trials evaluate treatment benefit using outcome assessments that measure, directly or indirectly, how patients feel, function, or survive in a disease or condition. Outcome assessments include survival, clinical outcome assessments (COAs), and biomarkers. A COA is any evaluation that can be influenced by human choices, judgment, or motivation. There are four types of COAs: patient-reported outcome (PRO), clinician-reported outcome (ClinRO), observer-reported outcome (ObsRO), and performance outcome (PerfO) [1]. A COA may support either direct or indirect evidence of treatment benefit. Unlike a biomarker that relies completely on an automated process or algorithm, a COA depends on the implementation, interpretation, and/or reporting from a patient, a clinician, or an observer. This report focuses on ClinROs. A ClinRO is an evaluation in which a member of the investigative team with appropriate professional training is the rater. A ClinRO assessment is any COA in which the individual determining the rating must have some specific professional training to properly form a judgment. Note that although all ClinRO assessments are COAs, not all COAs are ClinRO assessments. Appropriate development of ClinRO assessments and other COAs is consistent with the current trend toward a more patient-centric approach to drug development. Increased assessment validity and reliability can help to better define relevant treatment benefit, decrease variability resulting in a decrease in the number of patients needed to demonstrate treatment benefit, provide better information to make decisions regarding benefits versus harms for regulatory review, improve decision making in clinical practice, and justify payment for new interventions.

Background to the ISPOR Task Force

Since 2009, ISPOR has published eight ISPOR PRO Task Force Reports based on addressing aspects of the development and application of patient reported outcomes. (See the ISPOR Good Practices for Outcomes Research Index http://www.ispor.org/workpaper/practices_index.asp) These ISPOR PRO Task Force Reports are consistent with the US Food and Drug Administration's guidance for industry, "Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims" that described how the FDA would evaluate the adequacy and appropriateness of PRO measures used as effectiveness end points in clinical trials.

With the FDA's evolution toward the review and qualification of clinical outcome assessments (COAs), defined as any reported assessment used to support primary or secondary endpoints to document treatment benefit, several ISPOR PRO Task Force members decided to address a gap in the guidance focused on a specific type of COA — clinician reported outcome (ClinRO). A proposal was submitted, and in January 2013, the ISPOR Health Science Policy Council recommended formation of an ISPOR Clinical Outcomes Assessment – ClinRO Emerging Good Practices for Outcomes Research Task Force. The ISPOR Board of Directors subsequently approved task force formation.

Task Force members and primary reviewers were selected to represent a diverse range of perspectives: government / regulatory (US FDA and the European Medicines Agency (EMA)), academia, research organizations, and the pharmaceutical industry. The task force leadership group was comprised of experts in PRO and other assessments development, psychometrics, clinical trial data collection, and regulatory affairs. An international perspective was gained through task force membership, presentations at the ISPOR Annual Meetings and European Congresses, as well as through two rounds of reviews by the 650+ person ISPOR PRO / COA Review Group.

The task force met approximately every five weeks by teleconference to develop an outline, discuss content issues that arose and revised drafts. In addition, task force members met in person at the ISPOR International Meetings and European Congress. In the course of task force deliberations and in response to specific comments and suggestions from reviewers, it became clear that a foundation on COAs was needed before the ClinRO material could be discussed. The task force developed the introductory material, and *Clinical outcome assessments: A conceptual foundation – Report of the ISPOR Clinical Outcomes Assessment Emerging Good Practices Task Force* was published in September 2015.

The second report delves into this one COA, the specific aspects of ClinRO assessments and makes good measure practices recommendations. Like the first report, preliminary findings and recommendations were presented at the ISPOR Annual International Meeting and European Congress. Comments received during these presentations, as well as the written comments received during the two review rounds were addressed in subsequent drafts of the report. All comments were considered. Most were substantive

and constructive. Both ISPOR COA Task Force Reports are expert consensus guidance recommendations publications.

The ISPOR Clinical Outcomes Assessment Task Force developed two task force reports. The first report, "Clinical Outcome Assessments: A Conceptual Foundation—Report of the ISPOR Clinical Outcomes Assessment Emerging Good Practices Task Force" [2], covered the basic tenets, including definitions. It provides the background for this second report on emerging good measurement practices for ClinRO assessment development and evaluation.

Because this is a new field, we consider these recommendations to be emerging. They may evolve as more research is conducted. For convenience, we will refer to emerging good measurement practices as good measurement practices in this report. Although some of the practices are discussed in other publications [3], this report describes considerations for how to implement and operationalize good measurement practices specifically as they apply to ClinRO assessments.

The report's recommendations are intended to guide the development and evaluation of ClinRO assessments used to support conclusions regarding treatment benefit during the medical product review process and to support the deliberations concerning the use of those medical products in patients.

The goal is to develop ClinRO assessments that are fit for their intended purpose in the most efficient manner possible. The explicit objective in following good measurement practices is to minimize error in measurement of the concept of interest and achieve clarity in the interpretation of study results when evaluating treatment benefit. A number of ClinRO assessment examples are included in this report.

The results generated by ClinRO assessments are used in end point definitions that are specific to a clinical study protocol. A recent study showed that more than one quarter of denials of the first review approval for new drug applications for new molecular entities submitted to the US Food and Drug Administration (FDA) were based on a lack of clarity on the clinical meaningfulness of end points (13.2% of applications) and a lack of consistency across multiple study end points (in another 13.2% of applications). This could reflect a lack of effect with the interventions or a lack of appropriate development and/or clarity on end points [4].

These data highlight the importance of good measurement practices and proper development of outcome assessments for use as end points in clinical trials to more efficiently and accurately measure treatment effects. In addition, clarity on treatment benefits is important beyond regulatory approval. It provides justification for the uptake of new interventions into clinical practice and provides evidence to payers and health technology assessment bodies for reimbursement on the basis of the clearly demonstrated added value of new interventions [5].

The task force report's descriptions of COAs, ClinRO assessments, and recommended good measurement practices for their development and evaluation are closely aligned with those

in the US FDA PRO Guidance [6], the Guidance on the Qualification of Drug Development Tools [7], and the FDA Roadmap to Patient-Focused Outcome Measurement in Clinical Trials [8]. We summarize the good measurement practices presented in this report in a checklist—the Clinical Outcome Assessment Review of Listed Items (Table 1).

Treatment Benefit

Measurements of the treatment benefits of medical interventions rely on specified outcome assessments evaluating how patients feel, function, or survive. This assessment is used in defining the trial end points and in comparing patient groups within the study. Confirmatory controlled clinical trials (e.g., phase 3 trials) are designed to evaluate the study end point for patients who received the investigational intervention compared with those who received a comparator intervention, often a placebo or an active control.

These rigorous studies are the usual source of evidence demonstrating treatment effects (differences between the test and control groups) that support conclusions of treatment benefit (a meaningful effect on how patients feel, function, or survive). Therefore, differences in study end point results between treatment and control groups need to show or be confidently interpreted as indicating a meaningful effect on how patients feel, function, or survive.

In contrast to a PRO assessment in which patients can *directly* report their feelings and function, a ClinRO assessment is usually *indirectly* related to how patients feel or function in their daily lives. Clinicians cannot know exactly how patients feel or experience feelings or symptoms. Moreover, in daily life, it is unusual for clinicians to observe patients outside the clinic setting. Situations in which a ClinRO assessment is practical and when it is not possible for patients to directly evaluate symptoms and functioning in daily life directly, and/or when clinical judgment is required to make the assessment, a ClinRO assessment may be used to make observations related to patient functioning or signs [2].

This report does not address issues related to the *direct* assessment of survival (all-cause and cause-specific). Assessment of survival is discussed only in the context in which a ClinRO is used as an *indirect* assessment to predict future survival. A detailed description of direct and indirect evidence of treatment benefit is provided elsewhere [2].

Constructing End Points

All COAs, including ClinRO assessments, can be used as measurements to construct end points, but they are *not* end points in and of themselves. An end point is defined by *how* the COA is 1) used to quantitatively evaluate the effect of an intervention, 2) analyzed to determine the differences between test and control groups in the specific setting of a clinical trial, and 3) interpreted to reflect how those differences affect how patients feel, function, or survive.

Because many efficacy end points in treatment trials are defined using a ClinRO assessment as the COA, a ClinRO assessment should be 1) well defined in terms of the concept of

interest, 2) reliable, 3) able to detect change, 4) interpretable in a clinical trial in the population tested, and 5) representative of how patients survive, feel, or function.

A ClinRO assessment can have multiple subcomponents that affect clinician judgments. Although biomarkers are not ClinRO assessments, they can be included in a ClinRO assessment as part of the information clinicians use to form opinions or decisions on the basis of judgments. ClinRO assessments do not include situations in which biomarkers are the *sole* deciding factor for the clinical decision. For instance, decisions to administer blood transfusions on the basis of defined hemoglobin concentrations *plus* clinician-judged signs and symptoms are ClinRO assessments. In contrast, decisions to transfuse *solely* on the basis of reaching laboratory-defined hemoglobin concentrations *without* consideration of other factors are not ClinRO assessments.

A ClinRO assessment can be combined (composite or coprimary or secondary end point) with patient-reported assessments (e.g., pain intensity), other symptoms or patient function, or other types of outcome assessments (e.g., biomarkers). These various end points could differ in *appropriateness or meaning* for different disorders, different stages of the same disorder, or with different treatment objectives. The construction of the end point *must* be meaningful for patients in terms of how they feel, function, or survive, and not merely designed to show statistically significant differences between the test and control groups.

Furthermore, the *same* COA (including a ClinRO assessment) can be used in *different* ways to construct different end points. For instance, in skin infections, the ClinRO assessment might be clinician measurements of skin lesion area. The end point using this ClinRO assessment could be defined in several ways as 1) group differences on the amount of time to total resolution of skin lesions; 2) amount of time to decrease size of skin lesion area by a specified amount (e.g., a 20% decrease in lesion area); 3) difference in proportions of study participants who reach a “responder” criterion (e.g., total resolution or decrease in lesion area by a specified amount) by a specific time point (e.g., 3–4 days postrandomization); or 4) difference in average size of skin lesion area at a specific time point [9].

Types of ClinRO Assessments

We have categorized ClinRO assessments into three main types: 1) readings, 2) ratings, and 3) clinician global assessments (Table 2). ClinRO assessments are generated from procedures that result in readings or ratings of some form of information. The patient may be passive (e.g., sitting or lying), active (e.g., performing an activity, such as breathing deeply when clinicians are auscultating breath sounds), or questioned with regard to their symptoms or activities. This is then followed by the clinicians’ interpretation of these responses (unlike PROs in which there is no clinician interpretation).

Readings—Readings are clearly defined results that can be observed and reported (characterized) in a dichotomous manner (e.g., yes/no and presence/absence formats). Readings are most often dichotomized on the basis of clinicians’ judgment process, often using an assessment protocol with appropriate clinician expertise. Examples include the presence or absence of clinician-identified radiographic vertebral fractures or clinician-identified swollen joints [9].

End points based on dichotomous actions, such as prescribing additional medications, changing medications, or administering procedures, are defined and informed by clinician judgment, and thus are ClinRO COAs. Although end points such as hospitalizations are easily counted (yes/no) and appear dichotomous, the ClinRO assessment is the *decision to hospitalize patients* that depends on many different variables that form clinician judgments (e.g., various components of the patient's health status, bed availability, and social or administrative reasons unrelated to the illness under study).

The *reasons* leading to hospitalization related to the patients' health status are what is important, rather than solely the event of hospitalization itself. Where patients reside during the measurement is an indirect assessment of health status. Other examples include surgical interventions, administering oxygen or blood transfusions, discharge from hospital, and transfers to and from intensive care units. If the dichotomous action is based on well-defined variables that are readings or ratings (defined later), the ClinRO is defined by those assessments.

Ratings—Ratings are categorical (either ordered or not) or continuous measures, in which the assessment has at least three possible levels that generate scores representing the concept(s) of interest. Examples of categorical ClinRO assessments include the Ash-worth Spasticity Scale [10,11], the Aronchick Scale in bowel preparation for colonoscopy [12], and the Brief Psychiatric Rating Scale in mental disorders [13]. An example of a continuous ClinRO assessment is the size/area of skin lesions [14].

A rating type of ClinRO assessment may be dichotomized in an end point so that assessments higher or lower than certain thresholds are considered a "success" or "failure" on the basis of a defined responder criterion. Even though the ClinRO assessment itself is a rating, the end point as constructed for the trial is based on a dichotomous result. For example, in dermatologic indications, some studies use a multicategory scalar rating dichotomized to responder criteria defined by a two-grade or larger change to "clear" or "almost clear" from an initially more severe grade.

Clinician global assessments—Clinician global assessments (CGAs), including clinician global impressions and clinical global impressions of change, are assessments based on a clinician's overall judgment(s) of 1) the patient's total health status or 2) an aspect(s) of their health status for which the variable(s) assessed (e.g., which signs of disease should clinicians observe and assess) are not consistently defined or are undefined.

CGAs form a unique category of ClinRO assessments distinct from readings and ratings because the concepts that they assess are not explicitly defined. CGAs usually do *not* define 1) the specific variables assessed in performing the ClinRO assessment, 2) the methods of assessment, 3) how "response" on a particular variable is defined, or 4) how the variables are combined or weighted to arrive at the overall ClinRO assessment. Although often used in clinical practice and as supplemental/exploratory information in clinical trials, ClinRO CGAs are inadequate in a clinical trial setting as the primary or sole basis for evaluating the causal treatment effects of interventions.

Clinical practice is based on the use of interventions that 1) are intended solely to enhance the well-being of an individual patient and 2) clinician decisions are based on a reasonable probability of benefit [15]. In contrast, clinical research is designed to test a hypothesis about an intervention, permit conclusions to be drawn, and thereby develop or contribute to generalizable knowledge about the intervention [15]. In the setting of evaluating causal treatment effects of interventions, clarity and uniformity on what is assessed with regard to the intervention's benefits are needed. Clarity on outcome assessments, in turn, can inform clinical practice, and clinical practice can provide hypotheses to test in clinical trials.

Challenges in Development and Interpretation of ClinRO CGAs

Poorly measured variables or undefined/inconsistently defined or combined measures make it extremely difficult, if not impossible, to evaluate the properties of a CGA and to interpret the study result as a meaningful treatment benefit. Furthermore, added measurement error due to the lack of standardization can make it more challenging to detect important changes in patients' health status using CGAs. CGAs are often used as a "safety net" to capture unanticipated adverse effects or benefits of treatment in an exploratory way, but they are not useful as the primary or sole measure to define treatment benefit.

Given the lack of specification of the aspects of feel, function, concept of interest, and method of assessment for ClinRO CGAs, demonstration of the relationship between the CGA and the aspects of survival, feeling, and functioning is challenging, and in most cases, not possible. When the content of the ClinRO assessment is undefined or inconsistently defined, establishing content validity is inherently problematic.

Some studies of antimicrobials for the treatment of hospital-acquired pneumonia have exemplified the issues of unclearly defined global assessments [16]. Recent studies have shown that noninferiority on this global "clinical response" ClinRO assessment can be insensitive to increased mortality [17]. Similar problems in meaning and interpretation of the study results have been seen with studies of urinary tract infection [18] and heart failure [19].

The lack of specific definitions contributes to the often-weak reliability, validity, and ability to detect change, as well as interpretability issues with ClinRO CGAs. One study showed that ClinRO CGAs are affected by 1) variability between clinicians (because of differing criteria between clinicians and within individual clinician's judgments for the state of patients' health status or the "success" of a treatment); 2) the influence of recall bias when clinicians are asked to compare global assessments at specific times after trial initiation with a baseline global assessment; 3) poor reproducibility; and 4) response shift, if clinicians compare patients' health status with the previous visit instead of baseline [20]. Finally, studies have shown low inter-rater and intra-rater reliability when assessing signs of disease, such as tumor size on physical examination [21].

A study that compared a ClinRO rating assessment and CGA demonstrated the challenges in evaluating CGAs. The end point in a randomized controlled noninferiority trial comparing a test and control antibiotic in acute exacerbations of chronic bronchitis was complete resolution or improvement in clinician-judged patient symptoms [22]. The primary end point

in the trial was a CGA, that is, clinician-judged symptoms (specific symptoms or how much improvement counting as “response” not defined) improved such that no further antibiotics were prescribed. The case report form also included specific symptoms and asked clinicians to specifically capture symptom improvement on each variable (a ClinRO rating assessment on specific symptoms). A comparison of the CGA and the ClinRO rating assessment showed an overall kappa correlation of 0.63. The explanation for the discordance was unclear specifically because the variables used to determine the CGA were not defined.

Although the level of discordance might be interpreted as a “moderate” correlation, almost half of the participants in whom treatment would have been considered as treatment “failure” on the ClinRO rating assessment of specific symptoms were defined as treatment “successes” on the CGA [22]. It is often assumed that CGAs represent a more “comprehensive” assessment of patients’ health status. One could hypothesize that if this were true then the CGA would result in more “failures” than the ClinRO rating assessment, unless the ClinRO rating assessment was missing important symptoms.

Furthermore, the misclassification was differential between the test and control groups (more misclassification in one group compared with the other) showing that randomization does not address postrandomization events such as classification of outcomes. Last, the discordance between the CGA and the ClinRO rating assessment resulted in a change in the results of the study from +1.0% (95% confidence interval −3.8% to +5.7%) in favor of the test drug with the CGA to −2.2% (95% confidence interval −8.3% to +3.4%) in favor of the control drug, with the ClinRO rating assessment showing the potential to alter conclusions even with what appears to be moderate amounts of discordance.

A related challenge with ClinRO global assessments is that they may be influenced by assessments that are not ClinRO assessments. For example, if there is specified or inadvertent inclusion of a biomarker(s), this can influence clinician decision making and the overall end point definitions in uncertain ways. For instance, in the study comparing a CGA and ClinRO rating assessments, the authors hypothesized that clinicians may be incorporating biomarkers and signs such as body temperature preferentially over patient symptoms [22]. A ClinRO global assessment based on the vague “improvement in signs, symptoms, laboratory values, and radiography” may be strongly dependent on the biomarkers of laboratory values and radiography. These results may 1) override clinician judgments based on signs or symptoms; 2) not reflect patient benefit; and 3) lead to assessments varying between and within clinicians.

Furthermore, some ClinRO CGAs are used in an attempt to combine disparate outcomes that assess distinct aspects of survival, feelings, and functioning, as well as different concepts of interest (from other contexts of use) into a single outcome assessment. Unfortunately, this decreases clarity, confuses interpretation, and compounds uncertainty. For instance, if a single trial pools different stages of disease or different diseases in which different distinct disease manifestations predominate, separate well-defined outcome assessments may be needed because of the different aspects of feeling and functioning and the concepts of interest in these separate contexts (e.g., cough in pneumonia and dysuria in urinary tract

infections). Using a single CGA can mask the ill-defined amalgamation of the different aspects of feeling and functioning.

Use of a single nonspecifically described ClinRO CGA across various contexts of use does not address the issues of the distinct aspects of survival, feeling, and functioning nor the concepts of interest. This leads to problematic interpretation of the trial results, both overall and for the differing contexts of use.

Historically, the purported large effect sizes needed to overcome variability in measurement to demonstrate differences with ClinRO CGAs have been used as a justification for inclusion in superiority trials. The approach was based on the idea that if large effects are demonstrated, they must be clinically meaningful because of the magnitude of the effect. This, however, does not address the lack of clearly defined outcomes to understand the direct benefit to patients. Moreover, if biomarkers or minor changes in signs of disease (“hidden” surrogate end points) are driving observed differences on an end point, those differences between groups may not indicate an important treatment effect.

In noninferiority trials, CGAs are problematic because the variability in assessments and increased variance may drive study results toward minimizing differences when meaningful differences in treatment effects may exist. Demonstrating an absence of statistical differences between treatments on an unclear end point does not rule out a meaningful effect size in favor of either the test or control intervention or absence of clinical meaningful difference to patients, as in hospital-acquired pneumonia as noted earlier.

Some patient assessments or batteries of assessments include combinations of PROs, ClinROs, ObsROs, PerfOs, survival, and/or biomarkers. This discussion applies to the ClinRO elements of those assessments. Many of the good measurement practices explained here apply to all COAs developed using an iterative, stepwise combination of qualitative and quantitative methods [6,7].

Good Measurement Practices

General principles of good measurement practices for ClinRO assessments do not substantially differ from those for other COAs. There are, however, differences in the methods and approaches, as well as certain areas requiring increased attention. The principles presented here represent the current state of knowledge. Nevertheless, research on ClinRO assessments is still developing and additional research may reveal more precise differences and similarities between ClinRO assessments and other types of COAs.

Good measurement practices in developing and evaluating ClinRO assessments include 1) defining the context of use; 2) identifying the concept(s) of interest; 3) identifying and evaluating the relationship between the concept(s) of interest and the important aspects of feelings (e.g., symptoms), and function intended as the treatment benefit; 4) documenting content validity; 5) evaluating other measurement properties beyond content validity; 6) defining study objectives and end point(s) objectives, defining study end points, and placing study end points within the hierarchy of end points; 7) establishing interpretability for

clinical trial results; and 8) evaluating operational considerations when using ClinRO assessments to define end points in clinical trials (Table 1).

The first step in an initial ClinRO assessment evaluation is to recognize the intrinsic characteristics of the measure and what it is intended to represent. An important early element of this appraisal is to describe the intended benefit as an effect on a clearly identified, inherently meaningful aspect of how patients feel or function in their typical lives in relation to the clinical trial context.

There is no one-size-fits-all approach to outcome assessment development, including a ClinRO assessment. All assessments are imperfect in that they encompass the true measurement and some associated error. The goal is to reduce error in measuring the concept of interest and to achieve clarity in interpretation when evaluating treatment benefit. Although "validity" means that the ClinRO assessment measures what it purports to measure, we have avoided the terms "validity" and "validated" in this report because it implies that an outcome assessment is fit for purpose in all situations if it is declared "validated." Rather, we explore here the development and *evaluation* of ClinRO assessment measurement properties, including validity, for a specific context of use.

Good Measurement Practice 1: Define the Context of Use

The term context of use refers to the comprehensive descriptions that fully and clearly delineate the setting, manner, and purpose of use of the ClinRO assessment [2]. When used in clinical trials for development of medical interventions, context of use includes specifying *how* the ClinRO assessment is intended for use as an end point in clinical trials. As stated previously, ClinRO assessments are not in and of themselves end points.

The context of use description includes a variety of study attributes setting the boundaries within which a ClinRO assessment is appropriate for the intended use and considered adequate. When defining the context of use, it is important to consider *all* study features that can affect the validity and reliability of the assessments as discussed previously (Table 3) [2]. Inherent in the context of use is a clear definition of the disease and study population that determines the study entry criteria. Only *after* the disease definition and targeted study population are defined can appropriate outcomes be identified.

There is an inherent difference between COAs used in the context of diagnosis (enrollment criteria) compared with COAs used to define end points in treatment clinical trials. The purpose of diagnosis is to identify persons with and without disease or at risk of developing a disease. *The purpose of an outcome assessment used to define an endpoint in a clinical trial is to evaluate treatment benefit.*

A COA used for diagnosis is often not useful as an outcome assessment in treatment trials. For example, chest x-rays that help identify patients with pneumonia are not useful outcome assessments in pneumonia treatment studies because chest x-rays 1) do not represent how patients feel or function, 2) are slowly responsive to change over time, and 3) do not correlate with patient symptoms. Components of diagnosis, however, are used in prevention studies to define occurrence of disease. Even in the setting of prevention and disease

detection, disease often is defined by a composite of symptoms, signs, and other diagnostic laboratory criteria (not laboratory tests alone).

Similarly, assessment of the bulging of the tympanic membrane in acute otitis media may be useful in diagnosing bacterial versus viral disease; this evaluation is, however, not useful for measuring treatment benefit. It does not reflect a treatment effect on patient symptoms or functioning in their daily lives as evaluated using an Observer Reported Outcome assessment as shown in randomized trials [23,24].

A ClinRO assessment (or other COA) that is an appropriate outcome assessment in one context of use may or may not be adequate in a different context. For example, disease manifestations may differ between early and later stages of disease or other distinguishing aspects of disease populations. This may influence the choice of 1) primary concepts of interest, 2) the type of COA chosen (including whether a ClinRO assessment is the most appropriate or not), and 3) the type(s) of reporter to perform assessments. A PRO assessment may be the most appropriate COA for itchiness in atopic dermatitis in older children who can self-report, whereas a ClinRO or ObsRO assessment is a better COA for measuring clinical signs or behaviors in young patients who cannot self-report.

A given instrument can be used across different contexts of use if empirical evidence supports that the content validity—including the concepts, domains, and relative importance of items and domains providing the conceptual framework—remains constant among the contexts of use. The instrument may need to be used along with other outcome assessments in different contexts of use, for example, a PRO or ClinRO alone in early stage or less severe disease with no mortality, but a PRO or ClinRO along with mortality and complications in later or more severe stages of disease when death and disease complications may occur more often.

Nevertheless, defining an overall end point, the hierarchy of end points, and interpretations of scores may or may not differ between contexts of use. For instance, the content validity of an instrument to measure influenza symptoms was empirically shown to remain the same between hospitalized and nonhospitalized patients, obviating the need to develop an entirely new instrument for each context of use [25].

Good Measurement Practice 2: Identify the Concept(s) of Interest

The chosen concept of interest should align with key study objectives and intended treatment benefit. The concept of interest will determine how the treatment benefit is described, especially in the setting of regulated product labeling. Clarity on the intended treatment benefit of how patients feel, function, or survive that the concept of interest represents will help explain treatment benefit more clearly to patients and clinicians, payers, and other stakeholders. When the relationship between the concept of interest and treatment benefit is clear and adequately understood, the relevant aspects of patients' feeling, function, or survival may be used to describe the benefit provided to patients from interventions.

The concept evaluated by a ClinRO assessment forms the basis of the interpretation of findings from a clinical trial. The process of deciding what to measure relies on input from

clinicians, as well as patients, to identify the meaningful health aspect of how patients feel or function and what concept of interest may be both related to the meaningful health aspect and measurable by the clinician in the intended context of use. Specification of the concept of interest is often an iterative process as the relationship of the concept of interest to the meaningful health aspect is examined. Further consideration of the context of use may result in modifications in the concept of interest.

When inadequate attention is given to input from patients, clinicians, regulatory agencies, payers, or sponsors, the concept(s) of interest may be 1) inconsistent with the trial objectives and 2) challenging for interpretation of the clinical trial findings. There may be a lack of empirical evidence that the outcome assessments used as end points actually provided evidence of a meaningful benefit for patients.

To identify an adequate ClinRO assessment in a specific context of use, we recommend starting with identifying the meaningful aspect of how patients feel, function, or survive that is hypothesized as the intervention's treatment benefit. The next step is to identify the concept of interest that could be observed by a clinician that purports to assess the meaningful health aspect, and then, when possible, evaluate the relationship between the ClinRO assessment and the selected meaningful aspect of health as described in Good Measurement Practice 3 later.

For example, in acute bacterial skin infections, the meaningful health aspects that define treatment benefit are decreased symptoms, such as pain, and improved daily activities, such as bathing, dressing, or other activity related to the location of the skin infection. A ClinRO assessment measuring the concept of skin lesion size may be chosen on the premise that skin lesion size reductions reflect improved pain and daily activities [9]. Interpretation of study results and a conclusion of treatment benefit would depend on knowing the relationship between lesion size and how patients feel and function in the patient population studied.

Good Measurement Practice 3: Consider and Evaluate the Relationship between Treatment Benefit and the Concept of Interest Assessed by the ClinRO Assessment

As stated previously, most ClinRO assessments are *indirect* assessments of treatment benefit. A ClinRO assessment may be based on clinician evaluation of resolution or improvement of observable disease signs, or on a clinician's interpretation of what a patient says about "how he or she is doing" on some defined list of activities, which are indirect measures of how patients feel or function. For example, splenomegaly (enlargement of the spleen) in myelofibrosis or bulging of the tympanic membrane in acute otitis media are ClinRO assessments because they require professional judgment to perform the assessments, but these ClinRO assessments do not directly measure nor necessarily reflect how patients feel or function.

The indirect relationship does not eliminate the ClinRO assessment's potential to be a valuable outcome evaluation as long as there is an understanding of the relationship between the ClinRO assessment's measured concept of interest and the meaningful health aspect of patient feeling, functioning, or survival. If an understanding of that relationship does not

exist, an evaluation and explanation of how the ClinRO assessment relates to patients' lives in terms of how they survive, feel, or function is needed.

For example, improvement in the ClinRO assessment of clinicians' physical examination of lung sounds in pneumonia may or may not reflect patient benefit on symptoms such as cough and dyspnea. Assessment of biological activity has utility in earlier phases of evaluating medical interventions and selecting candidates for further study. They are not sufficient by themselves in demonstrating confirmatory treatment benefits on how patients feel, function, or survive.

COAs that measure concepts related to the intervention's mechanism of action do not necessarily represent how patients survive, feel, or function. Inhibiting growth of a microorganism with an antimicrobial agent or physiological changes caused by an intervention may explain *how* an intervention achieves a biological effect. Nevertheless, it may or may not reflect whether that mechanism *actually results in treatment benefit* in the context of use, that is, how patients feel, function, or survive.

Indirect measurements of treatment benefit, like most ClinRO assessments, fall along a continuum of "indirectness" (i.e., how closely the concept of interest is related to treatment benefit). Some indirect assessments are more closely related to real-life experiences than are others. For example, redness of the skin in psoriasis is closely related to patient symptoms. In contrast, although spleen size is an important physiologic and diagnostic measure for myelofibrosis, and may be related to some important symptoms, such as pain and dyspnea, it is not closely related to symptoms of night sweats and itching that are direct measures of treatment benefit (Fig. 1).

A rationale should be developed describing why a particular concept of interest, as measured by a specific indirect assessment, is expected to reflect a treatment benefit. Importantly, the relationships between measured effects on the basis of ClinRO assessment and the intended treatment benefits in terms of how patients function, feel, or survive should be described and evaluated empirically. As stated previously, patient and clinician input is needed to appraise the relationship between the concept of interest measured by a ClinRO assessment and the treatment benefit.

The relationship between a ClinRO assessment and a treatment benefit includes the timing of the expected benefit as well. A ClinRO assessment can represent a concurrent treatment benefit or a benefit expected to occur in the future. For example, a nurse's assessment of a neonate's crying behavior during a heel prick is likely to represent children's experienced *concurrent* pain. Conversely, a ClinRO assessment of clinician-judged signs of tumor shrinkage may be hypothesized to reflect a *future* treatment benefit of decreased mortality (increased survival). Such relationships between a ClinRO assessment and future treatment benefit are, however, rarely known. Increased survival needs to be directly demonstrated in order to conclude that a ClinRO reflects improved survival in a given context of use.

If improvement or resolution in the ClinRO assessment in the clinical trial's context of use is demonstrated to relate to how patients survive, feel, or function, the ClinRO assessment is appropriate as a measure of treatment benefit. The determination of appropriateness is made

in light of the specific context of use. Developing a hypothetical rationale on the basis of physiological or epidemiological evidence is a useful and necessary first step.

A theoretical rationale alone, however, does not demonstrate the relationship between an indirect assessment and how patients survive, feel, or function. Claiming that a ClinRO assessment represents treatment benefit because it is part of the disease pathophysiology is not sufficient. For instance, epidemiological evidence may demonstrate associations or correlations between tumor shrinkage and mortality as part of the natural history of treated or untreated disease independent of therapy. Nonetheless, it does not mean that an intervention's effect on tumor shrinkage after treatment will necessarily lead to a treatment effect on mortality in clinical trials [26]. Another example is that a treatment effect on vein appearance in superficial venous incompetence (varicose veins) measured by a ClinRO assessment may or may not relate to treatment effects of decreased symptoms of heaviness, achiness, throbbing, and itching.

Therefore, in practical terms, the relationship between direct assessments of treatment benefit and indirect outcome assessments should be clearly described and empirically tested. Ideally, treatment effects on direct measures of patient benefit compared with treatment effects on the ClinRO will demonstrate this relationship.

Some previously used ClinRO instruments have been shown to be unrelated and inadequate to support conclusions of treatment benefit. For example, the changes in the ClinRO global assessment of "clinical response" did not reflect treatment effects on patient symptoms in acute bacterial sinusitis [27]. Changes in tympanic membrane bulging did not reflect treatment effects on patient pain and function in acute otitis media [23,24].

Once the concept of interest has been selected and is known to relate to how patients feel, function, or survive, the content validity of a measure selected or developed to assess the concept of interest may be evaluated as described in the next section.

Good Measurement Practice 4: Document Content Validity of the ClinRO Assessment

Content validity denotes the extent to which a ClinRO assessment actually measures the concept of interest in the context of use. For PRO assessments used to directly measure treatment benefit, establishing content validity requires evidence that the instrument's items capture the way patients from the target population understand and express the aspects of feeling or functioning in the context of use of the clinical trial [6,28,29].

For ClinRO assessments, there are analogous tenets of content validity, that is, evidence that the ClinRO assessment captures the way clinical investigators understand and measure the concept of interest, in the clinical trial context of use. The ClinRO assessment is usually indirectly measuring a meaningful aspect of how patients feel or function. The ClinRO assessment's content validity should be documented *after* the relationship between the ClinRO assessment and how patients survive, function, or feel has been established (Table 4).

Although literature reviews can support the choice of the ClinRO assessment's concept of interest, clinician input is essential to establish the appropriateness and comprehensiveness

of the specific ClinRO assessment's content for measuring a particular concept of interest in a particular context of use. Previous use of a ClinRO assessment or previous publications using the ClinRO assessment alone do not necessarily ensure content validity of a ClinRO assessment especially in a new context of use.

The types of clinicians interviewed should be aligned with the types of clinicians who will perform the assessments in the clinical trial setting. Professional groups in gastroenterology, respiratory disease, rheumatic diseases, and other therapeutic areas have outlined recommended outcome assessments for certain diseases [30]. Some professional groups, however, have recommended diagnostic criteria for certain diseases that are not necessarily adequate for use as outcome assessments in clinical trials.

Clinician input can be obtained through interviews, either singly or in groups. Cognitive interviews, preferably performed with a separate independent group of clinicians, can help determine the completeness of concepts captured by a ClinRO assessment and how well they are understood. Interviews should continue until sufficient input is obtained to ensure a comprehensive and appropriate understanding of the concept of interest. Evaluating key concepts of the assessment until saturation is achieved can accomplish this goal.

Often, consensus will be necessary. This can be achieved using various methodologies, such as the Delphi technique or a nominal group approach [31]. The principles of interviews are in accord with those outlined for development of PRO assessments [6].

In multinational trial settings, evaluation of differences in expertise and training, as well as the variations in clinical practice at study locations, can inform the content validity of the ClinRO assessment. When multiple languages are needed for the study, interviews can also assess whether the phrasing in the translated ClinRO assessment has the same content in all regions.

As with PRO assessments, multifaceted items (single items with multiple components, e.g., redness, thickness, and scaling of skin lesions in psoriasis combined into a single item) pose particular challenges. They often lack clarity in how to combine different components. Separating the elements of multifaceted items, so that *each* item measures a *separate* element, reduces ambiguity for clinicians. It provides clarity on evaluating consistency of treatment effects across items. International guidance on clinical trials points out the need for clarity and evaluation of each component of composite assessments [32].

When ClinRO assessments have multiple components, the content validity of *each* component may be evaluated *separately* to understand how it is contributing to the content validity of the aggregate [32]. An overall assessment composed of multiple parts is only as strong as its weakest link. For example, in the concept of "gait quality" in children with cerebral palsy, clinicians might look at several aspects of gait, such as heel strike, foot roll, push off, and supination/neutral/pronation. Each could be rated on a scale of 1 to 4, and the several subscores combined in a stated way to yield a gait quality score. Each component has a different concept and together they address the concept of gait quality.

Therefore, proper evaluation of all components is necessary. If a COA comprises multiple types of assessments (e.g., PRO and ClinRO assessments), the content validity of each assessment should be evaluated *separately*. In addition, patient input is necessary for ascertaining patient understanding of what they are being asked to do if the ClinRO assessment involves a physical examination or patient responses of any kind.

For ClinRO readings or rating assessments, evidence should be provided on 1) the definitions of the concepts clinicians will be asked to assess, 2) the operations clinicians will perform, 3) the variables assessed (e.g., the observable signs of disease clinicians should assess), 4) how judgments will be made, 5) justification of the recall period (if any), 6) tests conducted to assess the domain and/or items in the instrument, 7) the range of clinician ratings relative to the clinical characteristics of study patients, and 8) the extent to which raters use the categories of the severity rating. In addition, the type and amount of specific training required to make assessments should be described in detail. (Smaller scale studies may be an option to help evaluate these ClinRO assessment characteristics before larger scale psychometric testing.)

Investigators may examine content validity across different contexts of use using qualitative data that address whether new content is required, and whether the relative importance of the items and domains is consistent across contexts. In different contexts of use, qualitative research may reveal potentially differing concepts of importance. In some situations, content may be the same, but there may be differences in the intensity of symptoms, for example, between periods of exacerbations of a disease and intra-exacerbation periods. This situation would not require an entirely new assessment, but may necessitate changes in scoring or interpretation of the results obtained with an assessment.

An illustrative ClinRO rating assessment is used for acute bacterial skin and skin structure infections (ABSSSI). The COA in current clinical trials is a ClinRO assessment of skin lesion size. It is used in an end point evaluating the change in skin lesion area from baseline to 48 hours after randomization [33]. Clinicians should be provided with specific instructions on 1) what aspects of the skin lesions they are to measure (e.g., redness alone or redness plus induration), 2) who on the study team should perform measurements, 3) whether the same person should make the assessments at baseline and at outcome, 4) how skin lesion assessments should be performed (e.g., body position of patient and measuring longest length and width), 5) the materials used to make measurements (e.g., a flexible ruler), and 6) the specific time points for assessment during the study [9]. Furthermore, any specific training needed to perform skin lesion measurements should be detailed.

Good Measurement Practice 5: Evaluating Other Measurement Properties after Content Validity Is Established

Once content validity is established, the evaluation of the psychometric and measurement properties of ClinRO assessments should follow the same basic principles as for PRO assessments [6]. The principles underlying “well-defined and reliable” measurement apply to ClinRO assessments as they do for PRO assessments [6]. These evaluations include 1) reliability (internal consistency and test-retest), 2) construct and known-groups validity, and 3) responsiveness/ability to detect change [3,34–36].

In PRO assessments, patients are their own reference standards for their health status. Although the general principles remain the same, several aspects of evaluation of ClinRO assessments call for increased attention because measures are performed or interpreted by someone other than the patient. For instance, inter-rater and intra-rater reliability/agreement has increased importance for ClinRO assessments. In COAs measured by someone other than the patient, there can be variability in measurements from one rater to another or from patient to patient on the basis of the performance of the measurement itself or the interpretation of the results with the same rater [21,37].

There has been considerable discussion regarding the terminology and evaluation of reliability/agreement [36,38], which is beyond the scope of this report. Measurements can vary within and between raters because of 1) differences in the mean values obtained (called “elevation” by Cronbach and Gleser [39]), 2) differences in the range of values obtained (SD or spread), and 3) the tendency of measures to vary together or separately over an entire range of measurement (scatter) [39,40].

Several types of statistical tests are available to evaluate reliability/agreement. The appropriate test depends on the type of measurement. Simple correlations, such as the Pearson and Spearman correlation coefficients may lack the ability to evaluate measures on all three types of variability (elevation, spread, and scatter). Methods such as the intraclass correlation coefficient and graphical methods to evaluate the limits of agreement (Bland-Altman plots) may be helpful [41,42].

Defining an “optimal” correlation may depend on the measures evaluated, the context of use, and the amount of variability that could result in clinically meaningful changes in the interpretation of results. Nonetheless, it is always good practice to state a priori the amount of acceptable variability before performing such analyses [41]. If the analysis of inter-rater and intra-rater agreement is not consistent with the prespecified definition of an acceptable measurement, then it may be necessary to retrain the clinicians performing the measurements (see Good Measurement Practice 8 later) or to re-evaluate the ClinRO assessment itself.

Good Measurement Practice 6: Using ClinRO Assessments to Define the Study End Point and End Point Positioning Aligned with the Study Objectives

The results generated by ClinRO assessments are used in end point definitions that are specific to a clinical study protocol. They need to be consistent with the study objectives and reflect treatment benefit. Investigators should distinguish between the ClinRO assessment itself and the way the ClinRO assessment is used to define an end point(s) in clinical trials. Considerations in using ClinRO assessments (or other COAs) as end points include 1) definition of the end point (either with the ClinRO assessment alone or in combination with other outcome assessments) and 2) analysis of the results including defining the timing of analysis and methods of statistical evaluation.

The positioning of study end points establishes a hierarchy of all the end points chosen and their specific contribution to the claimed treatment benefit(s) [6,43]. A lack of clarity in the ClinRO assessment’s role in the hierarchy of end points can lead to challenges in

interpretation of study results. For example, in a recent acute heart failure trial, the primary end point was a PRO assessment of acute dyspnea. Nevertheless, a ClinRO CGA based on clinicians' judgments on administration of additional medication for "worsening heart failure" was used to impute missing values on the patient-reported dyspnea assessment. In this case, the ClinRO assessment results impacted the overall results of the PRO assessment. This, in turn, resulted in an impact on the primary outcome. This impact was neither obvious nor explained in the description of the primary end point, raising challenges in interpretation of study results [19].

End points based on counting dichotomous events or actions can be based on ClinRO readings or rating assessments, such as clinicians' decisions to administer additional medication. If decisions are based on ClinRO CGAs with undefined or inconsistently defined variables, these end points can appear deceptively "objective" (and are often mistakenly called "hard" end points).

End points based on an event driven by clinician judgments are actually subjective—they vary between and within clinicians. They are often difficult to interpret. These end points may be proposed because of the apparent ability to count the number of events without interpretation, for example, additional prescriptions. Such end points, however, provide a false sense of objectivity because *the basis for clinical decision making* that resulted in the countable event is not clear. This leads to difficulty in interpretation of how the end point relates to the intended aspect of feeling, function, or survival.

Unless the variables assessed and the criteria for evaluation are clearly defined, end points based on clinician-judged overall "improvement" in a patient's health status are not interpretable. The relationships between those variables used in the ClinRO assessment and how patients feel, function, and/or survive must be clearly explained. Some end points, such as transfers of patients out of intensive care units, are intended to indicate improvement in health status. Nevertheless, extraneous factors, such as bed or staff availability in a given institution, may obscure the relationship between patients' health status and the measured actions (bed transfer). Furthermore, such actions may be based on different decision making in varying geographic areas on the basis of local practice patterns.

A ClinRO assessment can be used as part of a more complex multicomponent end point. This type of end point may include more than one ClinRO assessment, survival, other type(s) of COA (PRO, ObsRO, and/or PerfO assessments), and/or a biomarker(s) (e.g., the Alzheimer's Disease Assessment Scale—Cognitive) [44]. It can be difficult to determine how to interpret such multi-component end points. Moreover, how these elements are combined can be complicated. Regardless, ensuring the proper development and meaning of *each* component is critical.

Another example is the American College of Rheumatology (ACR) Scale for clinical trial end points in rheumatoid arthritis (ACR20, ACR50, and ACR70). This scale includes several types of ClinRO and PRO assessments and biomarkers [30]. In general, the clarity of multicomponent end points may be increased, if there is an explanation of *how* each component separately and in combination reflects patient treatment benefits.

Good Measurement Practice 7: Establishing Interpretability for Clinical Trial Results

Interpreting the results of a clinical trial involves evaluating whether those observed differences are meaningful to patients in their daily lives. The ability to detect change and the interpretability of clinical trial results depends on adequate attention to each of the good measurement practices outlined in the previous sections.

Even when there is a qualitative relationship between the ClinRO assessment and the intended treatment benefit, it cannot be ensured that any specific amount of measured difference between groups reflects a meaningful difference in health status to the patient or that the same amount of change along different parts of a scale has the same meaning (importance) to patients.

The ability to detect change and the test-retest reliability are ClinRO assessment measurement properties that provide the basis for statistical effect sizes. They will inform sample size calculations to show differences that achieve statistical significance. Observing statistically significant differences between treatment groups, however, does not ensure that the differences are meaningful to patients.

The methods used to evaluate end points including ClinRO assessments are similar to the analysis methods of end points using other types of COAs. There are two main types of analysis methods for interpretation of clinical trial results: 1) those based on individual patient change and 2) those based on group-level change.

The amount of change that is meaningful (important) to individual patients is often used to determine a *responder definition*. The responder definition may be used to propose a clinical trial analysis of differences in proportions at a specific time point or time for individual patients who meet the responder criteria.

The amount of change that is meaningful to patients may vary 1) depending on the context of use even within a single disease and 2) for the same outcome assessments when used across various diseases. For example, meaningfulness of incremental change in a ClinRO assessment may differ between the extremes of the rating scales' range for different disease stages. A given amount of change in a score may be meaningful for patients with more severe forms of a disease, but not for patients with a mild form of a disease or vice versa for a particular ClinRO assessment and disease.

All types of patients in a trial may not receive uniform treatment benefit from investigational interventions. For example, different magnitudes of treatment effects of the test intervention compared with the control may be observed on the basis of baseline characteristics of patients or disease states. For some patients, the size of the effect will be meaningful, whereas for others the study may show no effect or a trivial effect that is not meaningful.

Measurement with other COAs that have a well-understood interpretation of meaningfulness can provide anchors for comparison with ClinRO assessments. They can establish the amount of change in the ClinRO assessment that has identifiable meaning to patients and can be confidently used to define responder criteria for use in future clinical trials.

Exploration of group-level data is often useful to aid interpretation by analyzing patient subgroups on the basis of baseline demographic or disease-related factors and on the basis of a range of responder definition criteria. For example, graphic displays of the cumulative distribution of change can be useful to visualize the size and variability in response compared between test and control groups for the chosen end point.

When ClinRO rating assessments used as end points are analyzed as the difference in average score between groups, an understanding of the size of change (or difference) that is meaningful to individual patients can also be important. A responder definition used to explore clinical trial results aids interpretation of the clinical meaningfulness of the study findings because clinicians and patients may think of results in terms of differences between the numbers of patients rather than differences between rating scores.

Good Measurement Practice 8: Evaluating Operational Considerations Implemented in Clinical Trials

Protocol standardization of ClinRO assessment administration across clinicians and centers in instruction, training, and execution is crucial to reduce variability and increase reliability for treatment effects evaluation. Instruction and training should be consistent with the procedures for administration used in developing the ClinRO assessment. This is especially important in multicenter and cross-national trials, in which implementation may differ across various sites [45–47]. Developers should plan and ensure adequate training of investigators and site staff to mitigate sources of variability and measurement error.

The appropriate amount of clinician training in implementation of ClinRO assessments is context-specific. Some amount of training and quality assurance is needed in almost every situation. The training materials are a component of evaluating the adequacy of ClinRO assessment development and should be documented with other elements of ClinRO assessment development. A comprehensive user manual for standardized implementation, data capture, analysis, and interpretation of ClinRO assessments should be included. In addition, training materials in the local language(s) of participating clinicians should be provided.

Furthermore, developing criteria to qualify participating clinicians as correctly completing ClinRO assessments may be helpful. Depending on the context of the study, periodic clinician requalification may be helpful as well. Finally, instituting a means of quality assurance during data collection can be advisable.

Elements of study design can play a key role in the performance of ClinRO assessments. Special consideration should be given to aspects of study design that can influence patient cooperation or clinician judgment during the performance of ClinRO assessments. Traditional methods for decreasing systematic error, such as masking/blinding of patients, clinicians, and other outcome assessors to treatment assignment, are extremely important because of the clinician judgment influencing ClinROs and study outcomes.

Applying Good Measurement Practices

ClinRO Assessment Evaluation and Development of a New ClinRO Assessment—Skin Lesion Area in ABSSSI

Good measurement practices in developing and evaluating ClinRO assessments include 1) defining the context of use; 2) identifying the concept(s) of interest; 3) identifying and evaluating the relationship between the concept(s) of interest and the important aspects of feelings (e.g., symptoms), and function intended as the treatment benefit; 4) documenting content validity; 5) evaluating other measurement properties beyond content validity; 6) using ClinRO assessments to define study end points and end point positioning (e.g., the hierarchy of multiple end points); 7) establishing interpretability for clinical trial results; and 8) evaluating operational considerations when using ClinRO assessments to define end points in clinical trials.

Evaluation as a ClinRO Assessment

Measuring skin lesion area is a ClinRO rating assessment because it is a scalar continuous measure performed using a flexible ruler at baseline and during study by clinicians (judgment required to perform measurement). The context of use (good measurement practice 1) is adult patients with ABSSSI who can self-report on their own symptoms. The concept of interest (good measurement practice 2) is skin lesion area.

As per good measurement practice 3, optimally, the concept of interest should have been chosen for its ability to reflect the patient benefit on the meaningful health aspect of patients' lives that represents treatment benefit in this disease. The treatment benefit for this disease is improved patient function (improved ability to perform activities of daily living) and decreased pain.

In this example, however, the concept of interest (skin lesion area) was chosen on the basis of the demonstration of treatment effects on reduction of skin lesion area from historical studies, which allowed for justification of a noninferiority margin for trials in ABSSSI with noninferiority hypotheses. These studies did not demonstrate a clear relationship between the concept of interest and how patients feel or function [9,48]. In general, it is better to choose the meaningful benefit for patients first and then choose a ClinRO that reflects that benefit, rather than choosing a ClinRO solely on its ability to demonstrate treatment differences between the test and control groups.

Nevertheless, studies conducted more recently demonstrated high correlations between two separate PRO assessments (Visual Analogue Scale and Faces Rating Scale) measuring patient pain. Patient pain was related to decreases in skin lesion area over time, despite poor correlations of skin lesion size and pain at baseline [14]. These trials were also noninferiority studies. Without trials demonstrating differences between the test and control groups (superiority trials), it was not possible to show that the correlations between decreases in pain and changes in skin lesion area were due to treatment effects (good measurement practice 3).

Building on these findings, further studies were done to support the ClinRO assessment of skin lesion area. The content validity of the ClinRO rating assessment (good measurement practice 4) was supported by interviews with clinician and patients showing that signs of disease such as redness, swelling, and induration were important concepts in ABSSSI [49]. Evaluation of the measurement properties of the ClinRO rating assessment showed high inter-rater reliability of skin lesion area measurements over time (good measurement practice 5) [14].

The ClinRO rating assessment of skin lesion area was used to define an end point for clinical trials in ABSSSI of a responder criteria of 20% or more decrease in lesion area from baseline (good measurement practice 6) [9,48]. This ClinRO rating assessment has been used in clinical trials to date [33,50,51]. Finally, investigators were trained as part of clinical trial start-up on how to perform and capture measurements (good measurement practice 7) [14].

Of note, the ClinRO rating assessment of skin lesion area as the primary end point replaced in US trials the previous poorly defined CGA of clinician-judged improvement in patient signs and symptoms such that no further antimicrobial therapy is needed [9,48]. This CGA is still used as a secondary end point in US trials and as a primary end point in European trials. A PRO assessment directly measuring patient benefit on a comprehensive list of symptoms and patient function is under development [52].

Conclusions

ClinRO assessments are commonly used in end points that form the basis for review and approval of medical interventions. Applying the good measurement practices outlined here can increase efficiency of clinical trials while providing clarity on treatment benefit for patients.

Considerations regarding ClinRO assessments (and all outcome assessments) as end points in trials should occur early in the development process. Trial designers should consider the context of use of the proposed study. This informs the aspects of survival, feeling, and functioning in patients' daily lives on which the intervention is hypothesized to have a treatment benefit. Trial designers should decide what concept(s) of interest is informative about that treatment benefit and appropriate for the context of use, and how to best assess the concept(s) of interest using all-cause survival, COAs (PRO, ClinRO, ObsRO, or PerfO assessments), and/or biomarkers.

Trial designers should decide whether a ClinRO assessment is the best choice for evaluating the chosen concept of interest in the specific context of use. The choices related to the meaningful aspect(s) of survival, feeling, and functioning; context of use; concept(s) of interest; and COA characteristics and measurement properties are best conceived as an iterative rather than strictly linear process.

For PRO assessments, the concept of interest and aspect of survival, feeling, and functioning (e.g., pain) are usually identical. In contrast, in ClinRO assessments, the concept of interest is usually an indirect assessment of the aspect of survival, feeling, and functioning.

Therefore, trial designers should consider the relationship between the meaningful aspect(s) of survival, feeling, and functioning and the concept of interest before developing, modifying, or choosing a ClinRO assessment for use in clinical trials.

Once these relationships and the concept(s) to be measured by the ClinRO assessment are clear, trial designers can proceed with documenting content validity of the ClinRO assessment. After evidence demonstrates that the ClinRO assessment measures the concept of interest in the context of use, other measurement properties can be evaluated including construct validity, reliability and ability to detect change, as well as information on interpreting what is a meaningful amount of change in a given context of use.

A ClinRO assessment can be defined as a dichotomous reading, a scalar or categorical rating, or a CGA. CGAs, however, should be avoided as primary end points because of the issues discussed in this report. A ClinRO assessment can be used to define end points in many different ways. Trial designers should consider how *each* end point fits within the hierarchy of study objectives and how outcome assessments, singly or in combination, can be used to provide confirmatory information about treatment benefit. Standardized training and procedures for administration and use of ClinRO assessments for the specific context of a trial should accompany implementation of ClinRO assessments in clinical trials.

Following these emerging good practice recommendations in the choice of a ClinRO assessment as an end point in a trial and in the development of a ClinRO assessment will aid trial designers, regulators, policymakers, payers, and ultimately patients by providing interpretable evidence of treatment benefit. Future work will elucidate when ClinRO assessments are best to use and optimizing methods for their development and implementation.

Acknowledgments

We gratefully acknowledge the following ISPOR member reviewers who contributed their time and expertise through submission of written comments: Michael Adena, Patricia Alegre, Rene Allard, Vasudha Bal, Katy Benjamin, Steven Blum, Michael Carter, David Cella, Susan M. Dallabrida, Rahul Dhanda, Ozlem Equils, Sonya Eremenco, Francis Fatoye, Henok Getachew, Greg Gogates, Parul Gupta, Michael Hagan, Manthan Janodia, Joanna Leśniowska, Wee Hwee Lin, Trudy Mallinson, Linda Gore Martin, Nneka Onwudiwe, Janos Pitter, K.C. Ramanth, Farhan Abdul Rauf, Etta Vinik, and Tom Willgoss. Their generous feedback improved the manuscript and made it an expert consensus ISPOR Task Force Report. Finally, we are especially grateful to ISPOR staff, including Elizabeth Molsen, for helping us get these task force reports done—from beginning to end —and to Kelly Lenahan for her assistance in producing this report. Special thanks to Electra Papadopoulos, Center for Drug Evaluation and Research, Food and Drug Administration; Mira Pavlovic, Haute Autorite de Sante; Trevor Richter, Canadian Agency for Drugs and Technologies in Health; and Segolene Ayme, European Union Committee of Experts for Rare Diseases for their comments.

References

1. US Food and Drug Administration. [Accessed January 3, 2017] Clinical outcome assessment (COA): glossary of terms. 2016. Available from: http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm370262.htm
2. Walton MK, Powers JH III, 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; 18:741–52. [PubMed: 26409600]
3. Aaronson N, Alonso J, Burnam A, et al. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res. 2002; 11:193–205. [PubMed: 12074258]

4. Sacks LV, Shamsuddin HH, Yasinskaya YI, et al. Scientific and regulatory reasons for delay and denial of FDA approval of initial applications for new drugs, 2000–2012. JAMA. 2014; 311:378–84. [PubMed: 24449316]
5. Ermisch M, Bucsics A, Vella Bonanno P, et al. Payers' views of the changes arising through the possible adoption of adaptive pathways. Front Pharmacol. 2016; 7:305. [PubMed: 27733828]
6. US Food and Drug Administration. [Accessed January 3, 2017] Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims. 2009. Available from: http://www.fda.gov/downloads/Drugs/Guidances/UCM193282.pdf
7. US Food and Drug Administration. [Accessed January 3, 2017] Clinical Outcome Assessment Qualification Program. 2014. Available from: http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm230597.pdf
8. US Food and Drug Administration. [Accessed January 3, 2017] Roadmap to patient-focused outcome measurement in clinical trials. 2014. Available from: http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/UCM370174.pdf
9. Talbot GH, Powers JH, Fleming TR, et al. Progress on developing endpoints for registrational clinical trials of community-acquired bacterial pneumonia and acute bacterial skin and skin structure infections: update from the Biomarkers Consortium of the Foundation for the National Institutes of Health. Clin Infect Dis. 2012; 55:1114–21. [PubMed: 22744885]
10. Ashworth B. Preliminary trial of carisoprodol in multiple sclerosis. Practitioner. 1964; 192:540–2. [PubMed: 14143329]
11. Bohannon RW, Smith MB. Interrater reliability of a modified Ashworth scale of muscle spasticity. Phys Ther. 1987; 67:206–7. [PubMed: 3809245]
12. Aronchick CA, Lipshutz WH, Wright SH, et al. A novel tableted purgative for colonoscopic preparation: efficacy and safety comparisons with Colyte and Fleet Phospho-Soda. Gastrointest Endosc. 2000; 52:346–52. [PubMed: 10968848]
13. Mortimer AM. Symptom rating scales and outcome in schizophrenia. Br J Psychiatry Suppl. 2007; 50:s7–14. [PubMed: 18019038]
14. Powers JH III, Das AF, De Anda C, Prokocimer P. Clinician-reported lesion measurements in skin infection trials: definitions, reliability, and association with patient-reported pain. Contemp Clin Trials. 2016; 50:265–72. [PubMed: 27530088]
15. US Department of Health and Human Services. [Accessed January 3, 2017] Ethical principles and guidelines for the protection of human subjects of research: the Belmont Report. 1979. Available from: http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html
16. Rubinstein E, Cammarata S, Oliphant T, Wunderink R. Linezolid Nosocomial Pneumonia Study Group. Linezolid (PNU-100766) versus vancomycin in the treatment of hospitalized patients with nosocomial pneumonia: a randomized, double-blind, multicenter study. Clin Infect Dis. 2001; 32:402–12. [PubMed: 11170948]
17. US Food and Drug Administration. [Accessed January 3, 2017] Anti-Infective Drugs Advisory Committee: televancin for hospital and ventilator associated pneumonia. 2012. Available from: http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/Anti-InfectiveDrugsAdvisoryCommittee/UCM329482.pdf
18. US Food and Drug Administration. [Accessed January 3, 2017] Review of ceftalozone-tazobactam for the indication of complicated urinary tract infections. 2015. Available from: http://www.accessdata.fda.gov/drugsatfda_docs/nda/2014/206829Orig1s000StatR.pdf
19. US Food and Drug Administration. [Accessed January 3, 2017] Cardiovascular and Renal Drugs Advisory Committee review of biologics licence application 125468 serelaxin White Oak, MD. 2014. Available from: http://www.fda.gov/advisorycommittees/calendar/ucm388348.htm
20. Kamper SJ, Maher CG, Mackay G. Global rating of change scales: a review of strengths and weaknesses and considerations for design. J Man Manip Ther. 2009; 17:163–70. [PubMed: 20046623]
21. Moertel CG, Hanley JA. The effect of measuring error on the results of therapeutic trials in advanced cancer. Cancer. 1976; 38:388–94. [PubMed: 947531]

22. Cooper, CK., Rochester, G., Komo, S., Powers, JH. Misclassification bias in clinicians' assessments (CA) of patient outcomes and patient symptoms (PS) in antimicrobial therapy of acute exacerbations of chronic bronchitis (AECB). Presented at: 45th Interscience Conference on Antimicrobial Agents and Chemotherapy, American Society for Microbiology; San Francisco, CA. 2005;
23. US Food and Drug Administration. [Accessed January 3, 2017] Issues in the design of clinical trials for systemic antibacterial drugs for the treatment of acute otitis media: a (public workshop). 2011. Available from: http://www.fda.gov/downloads/Drugs/NewsEvents/UCM273907.pdf
24. Hoberman A, Paradise JL, Rockette HE, et al. Treatment of acute otitis media in children under 2 years of age. N Engl J Med. 2011; 364:105–15. [PubMed: 21226576]
25. Powers JH, Guerrero ML, Leidy NK, et al. Development of the Flu-PRO: a patient-reported outcome (PRO) instrument to evaluate symptoms of influenza. BMC Infect Dis. 2016; 16:1. [PubMed: 26729246]
26. Institute of Medicine. [Accessed January 3, 2017] Evaluation of biomarkers and surrogate endpoints in chronic disease. 2010. Available from: http://www.iom.edu/Reports/2010/Evaluation-of-Biomarkers-and-Surrogate-Endpointsin-Chronic-Disease.aspx
27. Hadley JA, Mosges R, Desrosiers M, et al. Moxifloxacin five-day therapy versus placebo in acute bacterial rhinosinusitis. Laryngoscope. 2010; 120:1057–62. [PubMed: 20422704]
28. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2—assessing respondent understanding. Value Health. 2011; 14:978–88. [PubMed: 22152166]
29. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: Part 1—eliciting concepts for a new PRO instrument. Value Health. 2011; 14:967–77. [PubMed: 22152165]
30. Felson DT, Anderson JJ, Boers M, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum. 1993; 36:729–40. [PubMed: 8507213]
31. Jones J, Hunter D. Consensus methods for medical and health services research. BMJ. 1995; 311:376–80. [PubMed: 7640549]
32. [Accessed January 3, 2017] International Conference on Harmonisation of Technical Requirement for Registration of Pharmaceuticals for Human Use. Statistical principles for clinical trials E9. 1998. Available from: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf
33. Prokocimer P, De Anda C, Fang E, et al. Tedizolid phosphate vs linezolid for treatment of acute bacterial skin and skin structure infections: the ESTABLISH-1 randomized trial. JAMA. 2013; 309:559–69. [PubMed: 23403680]
34. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. Health Technol Assess. 2009; 13:iii, ix–x, 1–177.
35. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol. 2010; 10:22. [PubMed: 20298572]
36. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010; 63:737–45. [PubMed: 20494804]
37. Cochrane AL, Chapman PJ, Oldham PD. Observers' errors in taking medical histories. Lancet. 1951; 1:1007–9. [PubMed: 14825885]
38. Hernaez R. Reliability and agreement studies: a guide for clinical investigators. Gut. 2015; 64:1018–27. [PubMed: 25873640]
39. Cronbach LJ, Gleser GC. Assessing similarity between profiles. Psychol Bull. 1953; 50:456–73. [PubMed: 13112334]

40. Beckstead JW. Agreement, reliability, and bias in measurement: commentary on Bland and Altman (1986; 2010). Int J Nurs Stud. 2011; 48:134–5. [PubMed: 20850746]
41. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1:307–10. [PubMed: 2868172]
42. Sedgwick P. Limits of agreement (Bland-Altman method). BMJ. 2013; 346:f1630. [PubMed: 23502707]
43. Patrick DL, Burke LB, Powers JH, et al. Patient-reported outcomes to support medical product labeling claims: FDA perspective. Value Health. 2007; 10(Suppl 2):S125–37. [PubMed: 17995471]
44. Pena-Casanova J. Alzheimer's Disease Assessment Scale—cognitive in clinical practice. Int Psychogeriatr. 1997; 9(Suppl 1):105–14. [PubMed: 9447433]
45. Baldereschi M, Amato MP, Nencini P, et al. Cross-national interrater agreement on the clinical diagnostic criteria for dementia. WHO-PRA Age-Associated Dementia Working Group, WHO-Program for Research on Aging, Health of Elderly Program. Neurology. 1994; 44:239–42. [PubMed: 8309565]
46. Dopfner M, Steinhausen HC, Coghill D, et al. Cross-cultural reliability and validity of ADHD assessed by the ADHD Rating Scale in a pan-European study. Eur Child Adolesc Psychiatry. 2006; 15(Suppl 1):I46–55. [PubMed: 17177016]
47. Hanssen-Bauer K, Gowers S, Aalen OO, et al. Cross-national reliability of clinician-rated outcome measures in child and adolescent mental health services. Adm Policy Ment Health. 2007; 34:513–8. [PubMed: 17710527]
48. Talbot GH, Powers JH, Hoffmann SC. Developing outcomes assessments as endpoints for registrational clinical trials of antibacterial drugs: 2015 update from the Biomarkers Consortium of the Foundation for the National Institutes of Health. Clin Infect Dis. 2016; 62:603–7. [PubMed: 26668337]
49. Powers, JH., Howard, K., Saretsky, T., et al. Development of a patient-reported outcome instrument (SKINFECT-PRO) to standardize and qualify symptoms of acute bacterial skin and skin structure infection (ABSSSI). Presented at: 20th Annual Meeting of the International Society for Pharmacoeconomics and Patient Outcomes Research (ISPOR); Philadelphia, PA. 2015;
50. Boucher HW, Wilcox M, Talbot GH, et al. Once-weekly dalbavancin versus daily conventional therapy for skin infection. N Engl J Med. 2014; 370:2169–79. [PubMed: 24897082]
51. Corey GR, Good S, Jiang H, et al. Single-dose oritavancin versus 7–10 days of vancomycin in the treatment of gram-positive acute bacterial skin and skin structure infections: the SOLO II noninferiority study. Clin Infect Dis. 2015; 60:254–62. [PubMed: 25294250]
52. Powers, JH., Portalupi, S., Devine, J., et al. Acute bacterial skin and skin structure infections (ABSSSI): development of a new patient reported outcome (PRO). Presented at: 19th Annual Meeting of the International Society for Pharmacoeconomics and Patient Outcomes Research (ISPOR); Montreal, Canada. 2014;

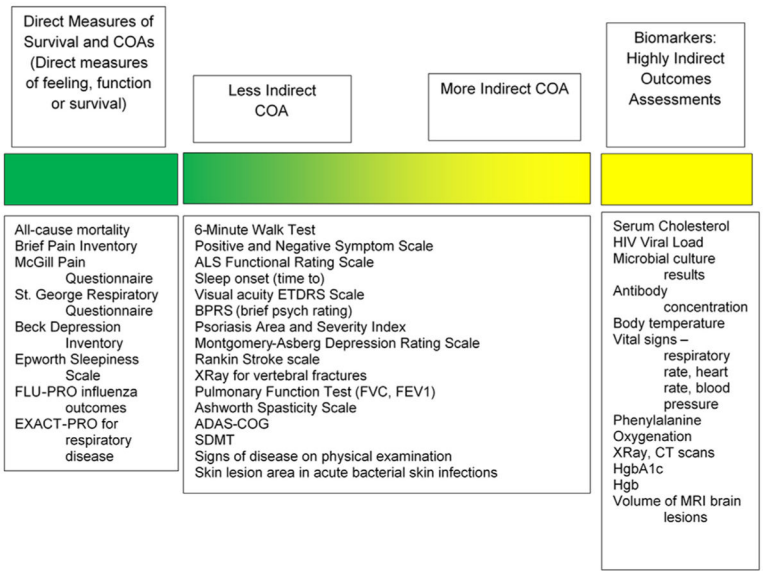

Fig. 1.

Direct to indirect assessments of treatment benefit. ADAS-COG, Alzheimer's Disease Assessment Scale—Cognitive; ALS, amyotrophic lateral sclerosis; BPRS, Brief Psychiatric Rating Scale; COA, clinical outcome assessment; CT, computed tomography; ETDRS, Early Treatment Diabetic Retinopathy Study; EXACT-PRO, the EXAcerbaCTtions of chronic pulmonary disease tool; FVC, forced vital capacity; FEV1, forced expiratory volume in 1; FLU-PRO, InFLUenza Patient-Reported Outcome; HgbA1c, glycated hemoglobin; HIV, human immunodeficiency virus; MRI, magnetic resonance imaging; SDMT, Symbol Digit Modalities Test.

Table 1

ISPOR Good Measurement Practices Checklist for COAs: ClinRO Assessment Development and Evaluation.

✓ Consider and define the context of use for the ClinRO assessment(s)
✓ Choose the intended treatment benefit on how patients feel, function, or survive that the ClinRO assessment is meant to reflect
✓ Identify the concept(s) of interest to be measured by the ClinRO assessment(s)
✓ Consider and evaluate the relationship between treatment benefit and the concept of interest assessed by the ClinRO assessment
✓ Document the content validity of the ClinRO assessment
✓ Evaluate other measurement properties of the ClinRO assessment once content validity is established
✓ Establish place of the ClinRO assessment to define study end points and study objectives and place in the hierarchy of end points
✓ Establish the interpretability of trial results
✓ Evaluate operational considerations for implementation of ClinRO assessments in clinical trials

ClinRO, clinician-reported outcome; COA, clinical outcome assessment.

Table 2

Definitions and types of ClinRO assessments.

ClinRO assessment: Evaluation of patients' health status reported by trained health care professionals (e.g., physicians, nurses, midwives, and therapists) and requiring specialized professional training to make the assessment
ClinRO reading: Evaluation in which the observed characteristics are reported in a dichotomous (e.g., yes/no) form (e.g., presence or absence of fractures, causal attribution of death) *
ClinRO rating: Evaluation in which the characteristics observed and reported have at least three possible categories or levels that generate scores representing the concept(s) of interest (e.g., Unified Parkinson's Disease Rating Scale, the Aronchick Scale in bowel preparation for colonoscopy, and the Brief Psychiatric Rating Scale in mental disorders)
ClinRO clinician global assessments (clinician global impressions and clinical global impressions of change): Evaluation based on clinicians' overall judgment(s) of the patient's total health status, or aspects of their health status for which the variables assessed are poorly defined or undefined.

CGA, clinician global assessment; ClinRO, clinician-reported outcome.

* An assessment that is dichotomized on the basis of a general and unstated clinician judgment process in which the variables measured are poorly defined or undefined (e.g., decision to prescribe additional medication on the basis of poorly defined "signs and symptoms") is considered a CGA.

Table 3

Considerations in defining the context of use of ClinRO assessments.

Disease definition (including, if appropriate)
• Pathogenesis
• Disease subtype
Targeted subpopulations
• Patients' demographic characteristics
• Disease severity
• History of previous treatment
• Culture and language
Clinical trial design and objectives
• End point model
• End point definitions
• Analysis plan
• Targeted labeling
Study setting
• Inpatient vs. outpatient
• Geographic location
• Clinical practice variation

ClinRO, clinician-reported outcome.

Table 4

Elements to document the content validity of a ClinRO assessment intended to provide evidence of treatment benefit.

- Definition of the concept of interest for measurement of the ClinRO assessment reading, rating, or global impression (i.e., conceptual framework of the instrument). For indirect measures of how patients survive, feel, or function, the documented relationship between the concept of interest and the treatment benefit should be established a priori.
- The following elements should be specific to the clinical trial context of use:
  - **Summary of literature review**
    - Research strategy relevant to the concept of interest
    - Published and nonpublished data
  - **Summary of concept elicitation methods and results, including evidence of saturation from concept elicitation with clinicians**
    - Protocol, interview guide, analysis plan
    - Transcripts
  - **Origin and derivation of concepts measured in the ClinRO assessment**
    - Rationale for omitting concepts
  - **Understandability of the instructions to implement or perform the assessment and to score or interpret the outcomes (research with clinicians)**
    - Protocol, interview guide, analysis plan
    - Transcripts
  - **Understandability of patient instructions, if any are associated with assessment**
    - Protocol, interview guide, analysis plan for research with patients
    - Transcripts
  - **Comprehensiveness of the ClinRO assessment qualitative and quantitative methods and results**
    - Protocol, analysis plan
    - Rasch or IRT on item locations if data available
    - Differential item functioning if applicable
  - **Integrity of the ClinRO assessment in other targeted languages**
    - Translatability assessment
    - Cultural adaptation methods
  - **Integrity of the measure across modes of administration**
    - Demonstration of measurement properties
- User manual
- Key references

ClinRO, clinician-reported outcome; IRT, item response theory.