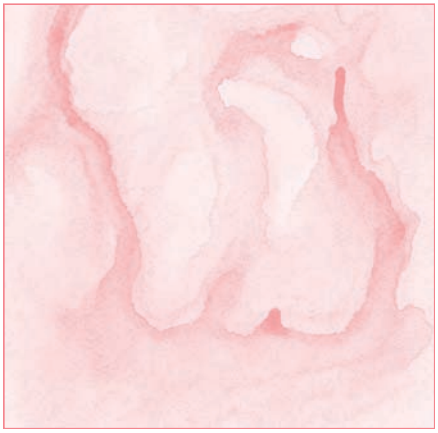
REVIEW



Clinician- and Patient-reported Endpoints in CNS Orphan Drug Clinical Trials: ISCTM Position Paper on Best Practices for Endpoint Selection, Validation, Training, and Standardization

by JOAN BUSNER, PhD; GAHAN PANDINA, PhD; SILVIA ZARAGOZA DOMINGO PhD; ANNA-KARIN BERGER PhD; MARIA T. ACOSTA, MD; NAHOME FISSEHA, PharmD; JOSEPH HORRIGAN, MD; JELENA IVKOVIC, MD; WILLIAM JACOBSON, PhD; DENNIS REVICKI, PhD; and VICTORIA VILLALTA-GIL, PhD, MSc

All authors are members of the ISCTM Working Group for Rare Disease/Orphan Drug Development; Drs. Busner and Pandina are Co-Chairs. Dr. Busner is with Signant Health in Blue Bell, Pennsylvania, and the Department of Psychiatry, Virginia Commonwealth University School of Medicine in Richmond, Virginia. Dr. Pandina is with Janssen Pharmaceuticals in Titusville, New Jersey. Dr. Domingo is with Neuorpsyncro in Barcelona, Spain. Dr. Berger is with Lundbeck in Copenhagen, Denmark. Dr. Acosta is with National Human Genome Research Institute, National Institutes of Health, in Bethesda, Maryland. Dr. Fisseha is with AbbVie Pharmaceuticals in North Chicago, Illinois. Dr. Horrigan is with AMO Pharma Limited and the Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine, in Durham, North Carolina. Dr. Ivkovic was with Lundbeck in Copenhagen, Denmark at the time this was written, but she is now with Zealand Pharma in Soborg, Denmark. Dr. Jacobson is with Harmony Biosciences in Plymouth Meeting, Pennsylvania. Dr. Revicki was with Evidera in Bethesda, Maryland. Dr. Villalta-Gil is with VeraSci in Durham, North Carolina.

Innov Clin Neurosci. 2021;18(10–12):15–22.

ABSTRACT

Objective: The International Society of CNS Clinical Trials Methodology (ISCTM) Working Group on Rare Disease/Orphan Drug Development is dedicated to improving and streamlining trials to best develop new treatments for rare diseases. The rarity of these disorders requires a drug development strategy that differs from those of nonrare conditions. Rare disease drug development programs are challenged with small sample sizes, heterogeneous clinical presentations, and few, if any, off-the-shelf endpoints. When disease-specific clinical endpoints exist, they might not be validated and are typically not well known or broadly used in clinical practice. This paper aims to provide an overview of the special issues surrounding endpoints in rare disease drug development, with guidance, practical applications, and discussion. **Discussion:** The paper covers regulatory considerations in endpoint selection; identification of relevant measurement domains; methods of quantifying clinical meaningfulness; incorporation of patient- and clinician-reported outcomes; considerations for global clinician- and patient-rated clinical assessments; cognition assessment challenges in rare diseases; translation considerations; training, standardization, and calibration of assessors; and endpoint quality assurance. Additionally, it provides guidance and resources for those involved in drug development for rare diseases. **Conclusion:** In keeping with the mission of ISCTM and the rare disease/orphan drug development working group, this article is designed to encourage thoughtful consideration and provide insight and guidance to promote and further efforts in in central nervous system (CNS) rare disease drug development efforts.

KEYWORDS: Endpoints, outcomes, COAs, eCOAs, orphan drug development, rare disease, assessment, measurement, CGI, PGI, orphan disease, International Society of CNS Clinical Trials Methodology (ISCTM), position paper

There are as many as 7,000 rare diseases, which are defined by the United States (US) Food and Drug Administration (FDA) as diseases occurring in less than 200,000 people in the US. Rare diseases are estimated to affect between 25 to 30 million people in the US and between 263 to 446 million people worldwide.[1] Largely ignored by pharmaceutical companies in the past, rare diseases are often areas of significant unmet clinical need, with affected individuals facing a reality of few or no approved treatments. Since 1983, with the advent of the FDA Orphan Drug Act,[2] more than 400 drugs and biologics have been approved and marketed for rare diseases. As of 2021, approximately one-third of all new drugs approved each year are for the treatment of rare diseases.

The International Society of CNS Clinical Trials Methodology (ISCTM) Working Group on Rare Disease/Orphan Drug Development is devoted to advancing best practices in clinical trial methodology for central nervous system (CNS) rare diseases. We recognize that the rarity of these disorders requires a drug development strategy that differs from those of nonrare conditions. Rare disease drug development programs are challenged with small sample sizes, heterogeneous clinical presentations, and few, if any, off-the-shelf endpoints. When these disease-specific clinical endpoints exist, they might not be validated and are usually not well known or broadly used in clinical practice.

Developing endpoints that meet clinical, scientific, and regulatory best practices has

FUNDING: No funding was provided for this study.

DISCLOSURES: Dr. Busner is a full-time employee of Signant Health and may hold stock and/or stock options. Dr. Pandina is a full-time employee of Janssen Research & Development, LLC and a stockholder of Johnson & Johnson. Dr. Fisseha is an employee of AbbVie and may hold stock and/or stock options. Dr. Horrigan is an employee of AMO Pharma Ltd. Dr. Jacobson is an employee of Harmony Biosciences.

CORRESPONDENCE: Joan Busner, PhD; Email: Joan.busner@signanthealth.com

become a central challenge in the development of novel treatments for these underserved populations.

SCIENTIFIC PRINCIPLES OF ENDPOINT DEVELOPMENT

Endpoints used in orphan drug development research are meant to be valid and reliable measures of the clinical benefit of the drug under investigation. Clinical benefit is defined by the FDA as a positive, clinically meaningful effect of an intervention (e.g., a positive effect on how an individual feels, functions, or survives).[3] Further details on scientific principles of endpoint development are available in Appendix 1, which can be accessed at https://innovationscns.com/wp-content/uploads/Busner-ISCTM-Position-Paper-Supplement.docx.

For rare diseases, natural history (the course a disease takes in the absence of intervention in individuals with the disease) data can guide selection of clinical endpoints from extant assessment measures used in clinical practice, novel clinical rating scales, and/or other endpoints.[4] The mechanism of action of the drug (i.e., the preclinical profile) can also aid in selection of the relevant clinical assessments. Endpoints will differ, for example, if the drug candidate is intended to improve an associated symptom versus having a broad, disease-modifying effect on the phenotype or specific biomarkers for a molecular targeted intervention, as in gene therapy interventions.

It is common in orphan drug development to use measures developed in other clinical populations to measure certain aspects of the targeted disease. Syndrome-specific rating scales that assess the entire phenotype are often used, particularly during Phase II development. Existing measures might be acceptable, but they typically do not assess all aspects of the disease in question and might miss key items of clinical relevance. Measures can also be adapted to meet the needs of the new population or population segments. Symptoms that are functionally important to one disease might derive from different biologic systems or might be peripheral manifestations of a systemic problem. For example, in Rett Syndrome, hypotonia, repetitive behavior, balance problems, breathing complications, and anxiety are all important features, but might not be subsumed under the same biologic system. In this case, a more specialized instrument might need to be developed, with an agreed-to set of critical items.

REGULATORY FRAMEWORKS AND APPROACHES FOR DEVELOPING AND VALIDATING ENDPOINTS FOR RARE DISEASES

Pediatric guidances and orphan drug designation. Orphan indication designation has a variety of regulatory advantages in a drug development program and has been successful in bringing more drugs to market. Advantages include greater regulatory feedback on the research approach, the potential for regulatory approval after one rather than two controlled pivotal trials (due to design optimization/agreement), and potential reduction of fees and length of clinical program/development paths, among others. In addition, incentives to study rare diseases can support a strategy for drug companies to begin innovative product development, such as in common genetic underpinnings associated with multiple rare diseases.

Because many rare CNS diseases affect pediatric populations, however, the regulatory advantages of orphan designation might be muted or less clear; the development program might still be faced with methodologic challenges that make it difficult to demonstrate efficacy. These include developmental variations and wider heterogeneity in disease expression across different age groups, as well as a lack of validated pediatric endpoints. This makes it difficult to identify validated endpoints that serve the potentially wide spectrum of age and heterogeneous disease expression.

For some companies, the additional work of developing age-appropriate pediatric endpoints might be viewed as impractical and unfeasible, with disproportionate time and money required to recoup research investment. Examples of difficulties encountered include determination of need and management of different age-band scale versions in analysis sets and need and management of parent versus child input when a child's cognitive maturity precludes accurate self-understanding and/or symptom recall.

Conversely, orphan drug designation might result in fewer pediatric studies, as sponsors are exempted from pediatric studies due to the limited clinical population; this exemption has received criticism from advocacy groups, such as the Treatment Action Group and the Elizabeth Glaser Pediatric AIDS Foundation.[5]

Different regulatory pathways for endpoint development. There are two main pathways to develop and evaluate the measurement properties of endpoints and assessment tools within the regulatory environment for novel drugs or indications.

The first regulatory pathway is to select a primary endpoint in the course of a clinical development program or new indication for an existing drug, in agreement with the regulatory authority (context of use [COU]). The development of an endpoint in the COU of an upcoming trial, rather than developing a new tool entirely, is the most common path used globally. This pathway is often the simplest and fastest method, relying on gold standard clinical measures familiar to the field. When no measure exists, however, as is often true for rare diseases, sponsors might attempt to modify an existing measure used for an associated condition, working in collaboration with regulators, patient groups, clinical experts, and other stakeholders. This can allow the program to move ahead without an extensive time investment in endpoint development.

The second regulatory pathway is to prequalify an endpoint prior to the start of a clinical development program (i.e., via an endpoint qualification program). Both the FDA and European Medicines Agency (EMA), as well as other agencies, have a process by which a sponsor may go through qualification of an endpoint. The FDA has the Clinical Outcome Assessment (COA) Qualification Program, and the EMA has the Qualification of Novel Methodologies for Medicine Development process. While these programs might result in a more robust, psychometrically sound primary endpoint, it might require years of time and substantial investment to conduct the necessary research. Given the limited subject pool in rare diseases, this can affect the sponsor's ability to conduct such studies, particularly when few or no therapies are available and when patient participation in endpoint development studies can affect later participation in clinical trials.

The research principles for endpoint development apply to existing tools used in a new context for a disease (e.g., the rare disease), as well as COU for developing new evaluation tools of any type, even if inspired by existing instruments. It is important to mention here that, in the instance of the FDA, the COA group plays a consultative role, but the final decision about the appropriateness of potential outcome measures rests with the actual review division (e.g., Psychiatry Products, Neurology Group 1, Neurology Group 2). This point is often missed by individuals with limited regulatory experience.

Global regulation and regulatory guidance documents. While the scientific principles of endpoint development for clinical development programs are universal, some regional differences exist between regulatory authorities, such as the FDA, EMA, and the Japanese Pharmaceuticals and Medical Devices Agency (PMDA), in requirements and approaches. There are also pilot programs emerging that are designed to encourage the use of novel technologies or approaches in endpoint development to facilitate better science. The Innovative Science and Technology Approaches for New Drugs (ISTAND) program from the FDA's Center for Drug Evaluation and Research (CDER) is one of these programs.[6]

COU endpoint development approach. The processes for endpoint development and evaluation typically start with a consultation meeting, often in the context of an investigational new drug (IND) application (FDA) or clinical trial application (CTA; EMA), with the regulatory agencies, with a predefined agenda. This meeting occurs prior to the first proof of concept study in the identified patient population. The FDA scientific experts on endpoint development may or may not be involved in this initial meeting or at follow-up meetings.

Inclusion of regulators as key stakeholders from the start of the project helps the final acceptance of the COAs to be used in a drug development program, provided the COA is eventually validated in the COU. Regulators, as stakeholders, can undertake different roles and provide guidance from the regulatory perspective in the different steps of COA development. Representation from different agencies in the US and European Union (EU) might be desirable to capture all perspectives and increase external validity outside the country or countries where the drug is planned to be first marketed.

Clinical experts in the targeted disorder are often key in proposing relevant outcomes of interest and possible corresponding clinical instruments to be used. The clinician recommendations are typically reviewed by COA experts, who are able to weigh in with respect to existing instruments, content validity, psychometric validation, and the regulatory environment, among other topics. Patient organization and advocacy groups are key in defining clinical meaningfulness from the patient's perspective, and input from these stakeholders has afforded increasing regulatory importance (e.g., FDA Patient-Focused Drug Development Guidance).[7]

FDA COA qualification approach. The FDA has a specific office focused on supporting the qualification of new tools for use in drug development research, as noted above. In the past, few researchers or sponsors followed this path, in part due to the uncertainty of outcome and the anticipated lengthy development and evaluation process required to establish evidence. More recently, however, a number of instruments have been submitted for this qualification and validation process, including new clinical outcome assessments and previously developed assessment tools that already have sufficient evidence to be accepted as sufficiently valid and reliable endpoints. Submissions are public and can guide the path for other researchers and drug developers. A number of new, publicly assessable COAs address rare indications. An example is Duchenne muscular dystrophy, for which the FDA has presented COA and an accepted COA video approach.[8,9]

The FDA COA qualification process has three main steps, beginning with a letter of intent (LOI). The LOI describes the need and presents existing preliminary evidence. If the LOI is accepted, there is the presentation of a qualification plan (QP), which is negotiated and agreed upon by the applicant and the agency. Once the QP is established, the research is conducted and summarized in a full qualification package (FQP). The FQP, which contains all relevant core and supportive data for evaluation, is then submitted for review by the agency. Qualification represents the determination that the drug development tool (DDT), within a specific COU, can be relied upon to have a specific interpretation and application in drug development and regulatory review. The full process for FDA endpoint qualification is summarized in a recent guidance document.[10]

Information for each endpoint that goes through the qualification process becomes public at certain fixed stages. Once qualified, the primary endpoint that has successfully completed the process can generally be included in an IND application, new drug application (NDA), or biologics license application (BLA) submission without the need for the FDA to reconsider and reconfirm its suitability. As mentioned earlier, while this path might be suitable for more common disorders, the complexity, time, and resources required might make this unsuitable for use in many orphan CNS drug development programs.

Companies exploring endpoints for rare diseases can access the FDA's compendium of prior approved endpoints in CNS disease. The FDA COA Compendium[11] is part of the FDA's efforts to foster patient-focused drug development. The COA Compendium is intended to facilitate communication and to provide clarity and transparency to drug developers and researchers by collating and summarizing COA information for many different diseases and conditions into a single resource. Consulting the COA Compendium can be a helpful starting point when considering a COA for use in clinical trials. The COA Compendium serves as a reference of past development programs, but new programs do not need to limit their strategy to what is in the compendium.

CLINICAL ENDPOINTS FRAMEWORK

The development of a new endpoint/clinical measure in common diseases is a well-established process. Typically, this requires the developer to identify all relevant and clinically important domains of the disease with representative items and to test and provide evidence supporting validity and reliability. For rare diseases, the classic gold standard approach might not be possible. Treatment sensitivity (and responsiveness) might be difficult to demonstrate in an indication where no treatments exist. In addition, the rarity of available affected individuals and a lack of knowledge of the full spectrum of disease might complicate endpoint development.

It is often a challenge to collect sufficient data to complete qualitative and psychometric studies in rare disorders. One approach is to combine samples from different studies to increase sample size. For example, clinical trial data can be combined with disease registry studies. Disease registry studies, however, are often heterogeneous in terms of disease severity and other characteristics. For some disorders, patients and their families might participate in multiple studies, resulting in some duplication across samples.

In 2018, as part of a series of public workshops on patient-focused drug development, the FDA issued specific guidance on selecting, developing, and modifying fit-for-purpose COAs to measure patient experience in clinical trials (Guidance 3 and Annexes).[12] This FDA guidance document provides a clear, stepwise process and clarifies concepts involved in the process of tool validation in connection with labeling claims.

COA selection good practices: identification of relevant domains (key and distal). Existing handbooks for the selection and use of clinical outcomes recommend starting with a literature review to ensure deep understanding of the natural history of the disease. This includes acute versus chronic symptoms, severity of the disease (mild, moderate, severe), fluctuations in clinical presentation, developmental course (for those diseases beginning in childhood), and progression rate (in neurodegenerative disorders). The Core Outcome Measures in Effectiveness Trials (COMET) initiative offers information on existing outcome studies.[13] The review of pre-existing works and contact with consolidated groups can be good starting points. Existing core outcome sets (COS) are reviewed by expert panels, which might include researchers and/or patients. This can help identify key domains for given disorders.

Once domains are identified, additional qualitative research with patients and proxies might be needed to help determine clinical meaningfulness of outcomes and ensure the content validity of the outcomes selected. For each therapeutic area, the best option should be evaluated and considered in the context of identifying and filling existing gaps in the assessment of potential treatment benefits.

The low incidence and prevalence of rare disorders make it especially difficult to effectively recruit a representative sample of patients or caregivers for qualitative purposes. Sponsors are encouraged to utilize a representative sample of patients via multiple methods (e.g., social networks, patient advocacy groups, disease-specific foundations, professional societies, existing national clinical trials networks) from representative countries to increase the external validity of the results.

Clear documentation and definition of endpoint selection process in study protocol. How a study endpoint or measure is established/selected, including which stakeholders were involved in its selection, must be clearly documented. Sponsors must provide all available reference data and background information on the measure. Often, an existing measure from another study is modified to fit the new therapeutic indication (in some cases significantly). The psychometric properties for the population will need to be assured in the new COU. A Delphi process can sometimes be used to establish the clinical outcome measures, with iterative versions of scales reviewed and evaluated by experts in the field, including multiple stakeholders.

Patient/caregiver- and clinician-reported measures and other endpoints: relevance in patient-centric approaches and orphan diseases. It is critical to have focused, efficient, disorder-specific, clinician-completed rating scales; functional/performance-based assessments; and, where feasible, clinical biomarkers. In addition, patient/caregiver outcome measures play an important role in orphan drug development. Many resources are available for the development of psychometrically patient/caregiver outcome sound measures. The FDA provides comprehensive guidance documents on the development of patient/caregiver reported outcomes.[14]

For pediatric populations, the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) conducted a task force focused on pediatric patient-reported outcomes (PROs) recommendations and good COA development practices and issued a reference document in 2013.[15]

Other groups, such as the International Rare Diseases Research Consortium (IRDiRC), have also issued position statements highlighting the importance of developing patient-centered outcome measures (PCOMs) in rare disease research to reflect the unmet needs of patients.

The use of item banks and individual patient-specific adaptive questionnaires are discussed in Appendix 1.

Global measures of clinical status—the global impressions scale. ISPOR has issued a guidance document for clinician-reported outcome assessments of treatment benefit.[16] This document identifies three types of clinician-reported outcomes: readings (clinician judgments of presence/absence, such as diagnosis or hospitalization needs), ratings (clinician judgment on a scale having at least 3 levels), and global assessments, deemed unique from readings and ratings in that the specific concepts forming the judgment are not specially defined.

The first global assessment scales to appear in the drug development literature were the Clinical Global Impressions of Severity (CGI-S) and Improvement (CGI-I), which were developed in the 1970s as a component of the Early Clinical Drug Evaluation Unit (ECDEU) assessments for use in psychopharmacology research.[17] These scales have been used broadly in clinical research since their publication, and their validity and utility have been extensively reviewed.[18,19] These scales are now included as secondary outcomes in virtually all psychiatric studies, many neurology studies, and an increasing number of rare disease orphan drug studies.[20–22] On occasion, CGI-S and CGI-I are included as primary outcomes (e.g., Angelman syndrome).[23]

In psychiatry, the CGI-S and CGI-I are included frequently in labeling and journal publications and have become widely familiar to prescribers, academicians, investigators, journal editors, and regulators. In their original and most common forms, CGI-S and CGI-I consist of seven-point unstructured anchors and are commonly used to define responders and remitters (e.g., change from baseline to end of treatment), and explore between group(s) change. These assessment tools and anchor points are also used to support responsiveness and calculations of minimal clinical important change and clinical responder thresholds for other COAs.

Controversies with the global impressions scale. Historically, the value of the CGI-S and CGI-I has been their reliance on the experienced clinician to make a global, integrated assessment over and above what might be seen on any created efficacy outcome scale. There is controversy, however, as to how directive the CGI-S and CGI-I should be. Although the CGI-S and CGI-I were intended to allow the experienced clinician latitude in conceptualizing overall severity and change,[16,17,19] guidance surrounding the individual anchors has ranged from none, as in the original, to generic conceptualizations[18,19] and highly specified algorithmic anchor definitions, requiring exact number and type of symptoms to equate for a severity level or clinical change.[25,26] Rather than specifying symptom combinations, some trials have included multiple CGI-S and CGI-I subscales, each associated with a particular symptom domain of interest (e.g., motor function, behavior, and sleep for Angelman Syndrome,[23] repetitive behavior and receptive language difficulties for autism[27]), which have been criticized for seeming to belie the concept of global impressions and for leaving the question of how to interpret drug response on only some, but not all, symptoms of a given disorder unanswered.[28]

Patient global assessment. The FDA has recognized the value of understanding the voice of the patient and stressed the importance of incorporating patient-reported outcomes in its patient-focused drug development initiative.[29,30] This initiative aims to substantially increase the role of the patient in the regulatory process (to explore change between group(s) and within the individual). The CGI assessments now include the Patient Global Impressions of Severity (PGI-S) and the Patient Global Impressions of Improvement (PGI-I, sometimes referred to as the Patient Global Impressions of Change, PGI-C). These relatively recent additions allow for the assessment of aspects of illness that are most problematic to the patient, as well as aspects of treatment perceived as most helpful.

Importantly, PGI-S and PGI-C data may or may not correlate with the CGI-S and CGI-I data or with what clinicians view as being important in a clinical study. Given these potential differences, it is important to consider whether PGI should trump CGI with respect to which aspects of illness or treatment change result in additional drug development. An understanding of the regulatory position on this crucial and somewhat novel aspect of drug development is important.

New positions by the COA group at the FDA—differing views. The COA group at the FDA recently discussed the benefit of changing the traditional seven-point CGI assessment scale to a simplified four- or five-point scale. Their rationale is that seven-category response instruments might not include clinically distinct response options and could require collapsing across levels, potentially resulting in a form of measurement error. A specific example of a four-point scale that they have proposed for CGI-S is as follows:

"Please choose the response below that best describes the severity of the patient's <OVERALL STATUS/ETC.> over the past (specify appropriate recall period here):

□ None, □ Mild, □ Moderate, □ Severe"

Such changes, newly recommended by the COA group and discussed in a series of documents in the patient-focused drug development guidance,[7] can add clarity to the voice of the patient in drug development when applied to a PGI, but can also raise some questions when applied to a CGI. Will anything be lost by making such changes? What will comparability across historical studies be? Will this change help or hurt signal detection?

At present, there is no hard and fast definition of what would constitute a moderate or a marked change. Would rigid definitions need to be put into place or pre-specified? For example, would something along the lines of two of XXX signs plus three of XXX signs with no functional impairment be defined as a "marked" change by the FDA? If so, would this jeopardize the basis of the CGI? How much of expert clinical "impressions" would be lost if the rater was scripted into a rigid algorithm? It remains an open question as to whether this would help or hinder CGI signal detection. Available information suggests that there is clinical and research value in maintaining a seven-point scale for CGI (and PGI) to preserve the utility of the measure in assessing change or outcome over time.

ENDPOINT AND DRUG DEVELOPMENT—SPECIAL CONCEPTS/TOPICS

Understanding the minimally clinical important difference (MCID). For any measure, but particularly for newer measures where little data are available, it is important to understand the clinical relevance of the measure and the relevance of change over time. This is often established using the MCID approach, where the threshold for clinically detectable change is identified. This is true of both patient/caregiver outcomes and clinician outcomes. The MCID is defined as the smallest difference (or change) that is considered meaningful for patients or their clinicians.[31] If the differences in mean baseline to endpoint change scores between an active treatment and placebo group exceed the MCID, these observed differences are deemed clinically meaningful. The FDA has focused on clinically meaningful within-individual responder thresholds, which are based on criteria often different from the MCID, suggesting that not all MCIDs are clinically meaningful.

Age/developmental level. When developing or choosing a scale, investigators should consider the developmental nature of the disease. The impact of the disease or symptoms might be different across age groups, thus requiring the scale to account for abilities at different ages. Different age-appropriate scales might need to be developed (e.g., alternative versions of the same instrument addressed to different age-bands in relation to cognitive developmental maturity, understanding/comprehension of item wording, and language skills), as could different versions that take into account age-related disease severity progression. Alternatively, adult scales might need to be validated in the COU of pediatric populations. At times, cutoff scores from established diagnostic samples, if available, can be used for inclusion or endpoint interpretation or sample stratification. In addition, the developmental impact of certain conditions might require the measurement of longitudinal rather than cross-sectional data observations to better assess the impact of the slope the disease and/or the intervention have in the developmental trajectory of a given individual.

Illness phase/stage. Disease progression can be modeled based on natural history data from cohort or registry studies. Well-matched, prospective, observational, natural history studies (e.g., COAs and biomarkers) can serve to decrease the number of patients needed in a placebo arm of a randomized clinical trial. This can be particularly useful in rare indications, when patients are difficult to recruit and when allocation to the placebo arm might be unethical (e.g., treatment is crucial but standard of care is not allowed in the clinical trial context). In some cases, endpoints only can be determined with natural history studies or retrospective data collection based in objective, subjective, and sometimes non-systematic data collection. Many times, it is necessary to consider the development of a multidomain endpoint, given the complex nature of the developmental impact of the condition and the heterogeneity of its manifestations across age groups.[32]

In-clinic versus remote assessment. Many factors can affect the ability to conduct in-person study visits with participants and caregivers. These include family-related factors, such as high-visit frequency creating an attendance burden, particularly for lengthy study visits, scheduling or transportation issues, or core features of disorders, such as physical, emotional, or behavioral problems that make clinic visits more challenging. Other factors are situational, ranging from simple proximity to actively recruiting study centers to the challenging COVID-19 pandemic. Electronic patient- and caregiver-reported outcomes, as well as telemedicine approaches, might enable greater flexibility in protocols when in-person visits are not feasible. There is a large body of literature noting the relative equivalence of paper versus electronic rating scales and growing evidence on in-person versus remote assessment equivalency for some clinician-rated outcomes.[33] There might be particular applicability of remote assessments to the orphan drug development arena, where sites are far-spread and travel is more difficult.

These electronic and remote procedures, however, have challenges: the technology is expensive to develop and maintain, and privacy requirements necessitate intensive planning and documentation to assure data security. Some study procedures, such as lab tests, physical exams, and collecting vital signs, cannot easily be done remotely or might require additional staffing. To ensure best quality endpoint data, as well as patient safety, the field must strive to facilitate and standardize a means of handling remote visits, particularly in the case of rare diseases, where the total population might be very low in frequency.

Endpoint translations quality in multinational clinic trials. Guidelines for how to create, translate, and validate existing rating scales have been established and utilized for decades.[34] This process includes 1) forward translation (by two native speakers, with knowledge of healthcare terms and linguistic and cultural knowledge); 2) independent, blinded, back-translation; 3) pilot testing and cognitive debriefing in nonexpert native language speakers; and 4) partial and/or full psychometric testing in a sample target population. Regulatory authorities accept linguistic validation (Steps 1–3), but the psychometric evaluation evidence is not necessary for international clinical trials. Also, depending on the type of COA, the process will have nuances in some of the steps. Step 2 can be iterated upon several times to produce a linguistic and cultural match of the original. Steps 3 and 4 cannot be done for all countries, particularly in scales that have been widely used and previously translated into multiple languages. This is a time-consuming and expensive process, but necessary to assure adequate validity of the scale (conceptual equivalence) across cultures. Expert groups have developed networks and streamlined processes for developing validated translations, and gold standard company procedures are available.[35,36] Given the limited number of individuals affected by a rare disease, aspects of the translation and validation process might need to be modified or reduced. Numerous countries might need to be included, with only a few participants from each country. Thus, it might not be feasible to incorporate all aspects of the validation process. When only English speaking countries are involved, the process of localization is needed to adapt the original English, usually from the US, terminology and vocabulary into to other linguistic communities, such as the United Kingdom, Canada, Australia, or South Africa.

The need for new translations should be considered early in the clinical trial process, as part of site selection, to provide sufficient time to identify available language versions of needed scales or to assure adequate development time should new translations be needed. These needs can also be factored into the study budget. For novel scales, translation should be considered during scale development; it might be useful to try at least one simple translation before finalization of the primary scale. It is also possible to evaluate translatability during instrument development to avoid known issues with the language of subsequent translations. This allows for consideration of language and concepts that may be more amenable to translation, as medical/clinical language, particularly around physical and emotional concepts, which can vary widely across cultures.

Measuring cognitive outcomes. Measures of cognitive function are often employed in orphan drug clinical trials. They can be used to measure short-term changes in cognition (hours or days) or long-term changes. In children and adolescents, where improved cognitive performance is expected as part of normal development, unless progressive cognitive impairment is a part of the natural course of the disease, normative data are often used as a comparator. Many diseases targeted for orphan drug development are associated with cognitive deficits or delays, either as a direct result of the disorder or due to comorbidities or polypharmacy. There is a base-rate problem, however, in that for many clinical populations, there are no large, well-controlled reference cohorts or registry databases to provide information on what progress can be expected in those receiving treatment-as-usual. Improvements in cognitive skills have been used as outcome measurements in clinical trials, but it is important to carefully select a trial duration that allows enough time to produce the expected change in the cognitive domain the intervention is targeted to modify. In general, complex cognitive domains involving composite scores, such as intelligence quotient (IQ), are more difficult to change than unitary domains, such as processing speed or reaction time. Also, composite scores can hide effects on single domains. The use of alternative forms of tests for test-retest reliability should also be considered to avoid learning effects and to capture a true change (improvement or worsening) across study visits. Incorporation of validated nonverbal tests might be required for some disorders. Agreement with regulators should be sought when an existing measure is applied to a new condition.

Implementation of rating scales in rare disease trials—rater training and calibration challenges. Rater training and calibration programs are generally designed to ensure that site investigators and personnel in charge of rating study patients understand the conventions, have a consistent understanding of scoring, and are proficient in administering the diagnostic instruments, cognitive testing, and measures of either the primary or other key clinician-reported outcome measures in the study. Rater training and calibration for orphan drug trials are associated with unique challenges. As has been discussed in previous sections of this article, new disease-specific outcome measures are often created to meet the needs of a rare disease trial. Such alterations often result in few to no investigators with scale familiarity or proficiency. In addition, heterogeneity of clinical presentation might necessitate customized administration modifications of outcome measures within a trial to ensure measurement validity. Common examples include modifications of administration to children who are ambulatory versus those who are nonambulatory, or those who require a feeding tube versus those able to eat unassisted.

It is not uncommon for the severity of the underlying disease to fluctuate and change during the course of a clinical trial; thus, investigators will need to be trained to reliably assess different levels of severity if a treatment change is to be detected.

Obtaining sufficient subject numbers in rare disorders often requires multinational trials with multiple investigators, each enrolling only one or two patients. Idiosyncratic rating practices across raters and countries, as well as "rater drift" caused by infrequent assessments, pose additional threats to standardization and adherence to trained conventions. An additional challenge in CNS rare disease training is the not uncommon overlap of medical specialties treating the same disorder, with adult and child psychiatrists, adult and child neurologists, and pediatricians frequently serving as investigators for the same trial. Discipline-wide differences in approach, experience, and rating practices must be addressed and harmonized to best ensure standardized, reproducible efficacy and safety ratings. A final challenge is the management of excessive placebo response that can emerge in pediatric rare disease studies due to expectancy effects.

Rare disease training recommendations. Comprehensive rater training for all study raters is highly recommended. This includes group consensus building exercises, description of study/scale conventions, use of video examples (where feasible), vignettes representing severity levels, practicum experiential exercises, scoring exercises to help establish inter-rater reliability, and study certification, such that raters are not permitted to rate in a study until they have achieved acceptable levels of scoring concordance and administration proficiency. Raters should be trained in means of assuring neutrality in interactions and data gathering. To help ensure continued calibration and minimize rater drift, we recommend continued evaluation and support of raters with retraining or recalibration plans.

Tools, such as audio or video monitoring of patient interviews with expert review and feedback for raters, as well as individual and group rater periodic recalibrations, are recommended to help ensure compliance with trained conventions. Challenges exist in some countries for capture and transfer of audio and video images, and solutions, such as housing personal data in a local-country server or obscuring facial features, must sometimes be designed in the service of ensuring rating accuracy throughout the study.

Examination of aberrant rating score patterns across visits, which suggest a potential misunderstanding of trained conventions, has also been used to help identify raters who might benefit from additional training and oversight.

Patient and/or caregiver training via videos and printed materials provide another route for assuring high quality data. Raters, patients, and caregivers might benefit from targeted education around inadvertent response biases and the natural phenomenon of placebo response. Patients and caregivers can be taught the value of reporting symptoms objectively, as that is the best means of helping a study achieve its goals. This can be of particular importance in rare disease studies, where patients might have been followed by one physician and study center for many years. It is important for patients and caregivers to understand that all reports, whether of benefit, stasis, or worsening, are paramount and in no way will jeopardize the relationship with, or indicate ingratitude to, the physician or study site.

CONCLUSION

Rare diseases represent an important, underserved population in clinical trials. Emerging science has led to a greater ability to develop targeted treatments, but endpoint development often lags behind. Endpoints for rare disorders often do not exist. Thus, novel, sensitive endpoints are a critical aspect in the development of new medicines in these populations. Heterogeneity, limited resources, and small patient samples complicate the conduct of drug development research in rare conditions. This article offers suggestions on approaches and concepts for tackling these endpoint-related challenges when embarking on an orphan drug development program.

To access Appendix 1, please visit https://innovationscns.com/wp-content/uploads/Busner-ISCTM-Position-Paper-Supplement.docx.

ACKNOWLEDGMENTS

This work is dedicated to the memory of our esteemed colleague and friend, co-author Dennis Revicki, PhD, who passed away on May 9, 2021.

REFERENCES

1. Nguengang Wakap S, Lambert DM, Olry A, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet*. 2020;28(2):165–173.
2. United States Food and Drug Administration. Orphan Drug Act–relevant excerpts. https://www.fda.gov/industry/designating-orphan-product-drugs-and-biological-products/orphan-drug-act-relevant-excerpts. Accessed December 20, 2021.
3. FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource [Internet]. Silver Spring (MD): Food and Drug Administration (US); 2016–. Glossary. 2016 Jan 28 [Updated 2021 Jan 25]. https://www.ncbi.nlm.nih.gov/books/NBK338448/. Co-published by National Institutes of Health (US), Bethesda, MD. Accessed December 20, 2021.
4. United States Food and Drug Administration. Rare diseases: natural history studies for drug development. 2019. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/rare-diseases-natural-history-studies-drug-development. Accessed December 20, 2021.
5. Treatment Action Group. Ensuring treatment for children with orphan diseases: ending exemptions from the pediatric research equity act (PREA). 2019. https://www.treatmentactiongroup.org/wp-content/uploads/2021/07/prea_brief_2021.pdf. Accessed December 20, 2021.
6. United States Food and Drug Administration. Innovative Science and Technology Approaches for New Drugs (ISTAND) pilot program. 2021. https://www.fda.gov/drugs/drug-development-tool-ddt-qualification-programs/innovative-science-and-technology-approaches-new-drugs-istand-pilot-program. Accessed December 20, 2021.
7. United States Food and Drug Administration. Patient focused drug development guidance

series for enhancing the incorporation of the patient's voice in medical product development and regulatory decision making. 2020. https://www.fda.gov/drugs/development-approval-process-drugs/fda-patient-focused-drug-development-guidance-series-enhancing-incorporation-patients-voice-medical. Accessed December 20, 2021.

8. United States Food and Drug Administration. Duchenne muscular dystrophy and related dystrophinopathies: developing drugs for treatment guidance for industry. 2018. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/duchenne-muscular-dystrophy-and-related-dystrophinopathies-developing-drugs-treatment-guidance. Accessed December 20, 2021.
9. United States Food and Drug Administration. Clinical Outcome Assessments (COA) Qualification Program submissions. 2021. https://www.fda.gov/drugs/clinical-outcome-assessment-coa-qualification-program/clinical-outcome-assessments-coa-qualification-program-submissions. Accessed December 20, 2021.
10. United States Food and Drug Administration. Qualification process for drug development tools guidance for industry and FDA staff. 2020. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/qualification-process-drug-development-tools-guidance-industry-and-fda-staff. Accessed December 20, 2021.
11. United States Food and Drug Administration. Clinical Outcome Assessment (COA) Compendium. 2019. https://www.fda.gov/media/130138/download. Accessed December 20, 2021.
12. United States Food and Drug Administration. Patient focused drug development guidance 3: discussion document: select, develop or modify fit-for-purpose clinical outcome assessments. 2018. https://www.fda.gov/media/116277/download. Accessed December 20, 2021.
13. Core Outcome Measures in Effectiveness Trials Initiative. Advanced search. https://www.comet-initiative.org/Studies. Accessed December 20, 2021.
14. United States Food and Drug Administration. Patient-reported outcome measures: use in medical product development to support labeling claims guidance for industry. 2009. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-reported-outcome-measures-use-medical-product-development-support-labeling-claims. Accessed December 20, 2021.
15. Matza LS, Patrick DL, Riley AW, et al. Pediatric patient-reported outcome instruments for research to support medical product labeling: report of the ISPOR PRO good research practices for the assessment of children and adolescents task force. *Value Health*. 2013;16(4):461–479.
16. Powers JH III, Patrick DL, Walton MK, et al. Clinician-reported outcome (ClinRO) assessments of treatment benefit: report of the ISPOR Clinical Outcome Assessment Emerging Good Practices Task Force. *Value Health*. 2017;20(1):2–14.
17. Guy W. *ECDEU Assessment Manual for Psychopharmacology*. Rockville, MD: US Department of Health, Education, and Welfare Public Health Service Alcohol, Drug Abuse, and Mental Health Administration; 1976.
18. Busner J, Targum S. The clinical global impressions scale: applying a research tool in clinical practice. *Psychiatry (Edgmont)*. 2007;4(7):28–37.
19. Busner J, Targum SD, Miller DS. The Clinical Global Impressions scale: errors in understanding and use. *Compr Psychiatr*. 2009;50(3):257–262.
20. Papapetropoulous S, Lee MS, Boyer S, et al. A Phase 2, randomized, double-blind, placebo-controlled trial of CX-8998, a selective modulator of the T-type calcium channel in inadequately treated moderate to severe essential tremor: T-CALM study design and methodology for efficacy endpoint and digital biomarker selection. *Front Neurol*. 2019;10:597.
21. Uchio Y, Enomoto H, Ishida M, et al. Safety and efficacy of duloxetine in Japanese patients with chronic knee pain due to osteoarthritis: an open-label, long-term, Phase III extension study. *J Pain Res*. 2018;11:1391–1403.
22. Yalcin I, Viktrup L. Comparison of physician and patient assessments of incontinence severity and improvement. *Int Urogynecol J Pelvic Floor Dysfunct*. 2007;18(11):1291–1295.
23. Kolevzon A, Ventola P, Keary CJ, et al. Development of an adapted Clinical Global Impression scale for use in Angelman syndrome. *J Neurodev Disord*. 2021;13(1):3.
24. Haro JM, Kamath SA, Ochoa S, et al. The SOHO Study Group. The Clinical Global Impression–Schizophrenia scale: a simple instrument to measure the diversity of symptoms present in schizophrenia. *Acta Psychiatr Scand*. 2003:107(Suppl 416):16–23.
25. Neul JL, Glaze DG, Percy AK, et al. Improving treatment trial outcomes for Rett syndrome. *J Child Neurol*. 2015;30(13):1743–1748.
26. Nissenkorn A, Borgohain R, Micheli R, et al. Development of global rating instruments for pediatric patients with ataxia telangiectasia. *Eur J Paediatr Neurol*. 2016;20(1):140–146.
27. Butter EM, Mulick JA. Brief assessment of the core symptoms of pervasive developmental disorders: preliminary development of the Ohio autism clinical impressions scale. Poster presented at NCDEU. 2006, Boca Raton, FL.
28. Scahill LS. Uncommon use of common measures in sulforaphane trial. *Proc Natl Acad Sci U S A*. 2015;112(4):E349.
29. Perfetto EM, Burke L, Oehrlein EM, et al. Patient-focused drug development: a new direction for collaboration. *Med Care*. 2015;53(1):9–17.
30. Chalasani M, Vaidya P, Mullin T. Enhancing the incorporation of the patient's voice in drug development and evaluation. Re*s Involv Engagem*. 2018;4:10.
31. Revicki D, Hays R, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol*. 2008;61(2):102–109.
32. United States Food and Drug Administration. Individualized endpoints in pediatric rare disease trials: a clinical perspective. 2019. https://www.fda.gov/media/133753/download. Accessed December 20, 2021.
33. Matcham F, Barattieri di San Pietro C, Bulgari V, et al. Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-centre prospective cohort study protocol. *BMC Psychiatry*. 2019;19:72.
34. Sousa VD, Rojjanasrirat W. Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. *J Eval Clin Pract*. 2011;17(2):268–274.
35. Wild D, Grove A, Martin M, et al. Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value Health*. 2005;8(2):94–104.
36. Acquadro C, Patrick DL, Eremenko S, et al. Emerging good practices for translatability assessment (TA) of patient-reported outcome (PRO) measures. *J Patient Rep Outcomes*. 2017;2(1):8. ICNS