Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments

Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Center for Devices and Radiological Health (CDRH)

October 2025
Administrative/Procedural

Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments

Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders

Additional copies are available from:
Office of Communications, Division of Drug Information
Center for Drug Evaluation and Research
Food and Drug Administration
10001 New Hampshire Ave., Hillandale Bldg., 4th Floor
Silver Spring, MD 20993-0002
Phone: 855-543-3784 or 301-796-3400; Fax: 301-431-6353; Email: druginfo@fda.hhs.gov
https://www.fda.gov/drugs/guidance-compliance-regulatory-information/guidances-drugs
and/or
Office of Communication, Outreach, and Development
Center for Biologics Evaluation and Research
Food and Drug Administration
Phone: 800-835-4709 or 240-402-8010; Email: industry.biologics@fda.hhs.gov
https://www.fda.gov/vaccines-blood-biologics/guidance-compliance-regulatory-information-biologics/biologics-guidances
and/or
Office of Policy
Center for Devices and Radiological Health
Food and Drug Administration
10903 New Hampshire Ave., Bldg. 66, Room 5431
Silver Spring, MD 20993-0002
Email: CDRH-Guidance@fda.hhs.gov
https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/guidance-documents-medical-devices-and-radiation-emitting-products

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Center for Devices and Radiological Health (CDRH)

October 2025
Administrative Procedural

TABLE OF CONTENTS

Contains Nonbinding Recommendations

Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders[1]

This guidance represents the current thinking of the Food and Drug Administration (FDA or Agency) on this topic. It does not establish any rights for any person and is not binding on FDA or the public. You can use an alternative approach if it satisfies the requirements of the applicable statutes and regulations. To discuss an alternative approach, contact the FDA office responsible for this guidance as listed on the title page.

I. INTRODUCTION

A. Overview of FDA Guidances on Patient-Focused Drug Development

This guidance (Guidance 3) is the third in a series of four methodological patient-focused drug development (PFDD) guidance documents[2] that describe how stakeholders (patients, caregivers, researchers, medical product developers, and others) can collect and submit patient experience data[3] and other relevant information from patients, caregivers, and clinicians to be used for medical product[4] development and regulatory decision-making. The topics that each guidance document addresses are described below.

---

[1]This guidance has been prepared by the Center for Drug Evaluation and Research, in cooperation with the Center for Biologics Evaluation and Research and the Center for Devices and Radiological Health, at the Food and Drug Administration.

[2] The four guidance documents fulfill FDA commitments under section I.J.1 associated with the sixth authorization of the Prescription Drug User Fee Act (PDUFA VI) under Title I of the FDA Reauthorization Act of 2017 (FDARA). The projected time frames for public workshops and guidance publication reflect FDA's published plan aligning the PDUFA VI commitments with some of the guidance requirements under section 3002 of the 21st Century Cures Act (available at https://www.fda.gov/downloads/forindustry/userfees/prescriptiondruguserfee/ucm563618.pdf).

[3] "Patient experience data" is defined for purposes of this guidance in Title III, section 3001 of the 21st Century Cures Act, as amended by section 605 of FDARA, section 569C(b) of the Food Drug and Cosmetic (FD&C) Act, (21 U.S.C. 360bbb-8c(c)) to include data that "(1) are collected by any persons (including patients, family members and caregivers of patients, patient advocacy organizations, disease research foundations, researchers and drug manufacturers); and (2) are intended to provide information about patients' experiences with a disease or condition, including (A) the 'impact (including physical and psychosocial impacts) of such disease or condition or a related therapy or clinical investigation on patients' lives; and (B) patient preferences with respect to treatment of such disease or condition."

[4] For purposes of this guidance a "medical product" refers to a drug (as defined in section 201 of the Federal Food, Drug, and Cosmetic (FD&C) Act (21 U.S.C. 321)) intended for human use, a device (as defined in such section 201) intended for human use, or a biological product (as defined in section 351 of the Public Health Service Act (42 U.S.C. 262)).

- Methods to collect patient experience data that are accurate and representative of the intended patient population (Guidance 1)

- Approaches to identifying what is most important to patients with respect to their experience as it relates to burden of disease or condition and burden of treatment (Guidance 2)

- Approaches to selecting, developing, modifying, and evaluating clinical outcome assessments (COAs) to measure outcomes of importance to patients in clinical trials (Guidance 3)

- Methods, standards, and technologies for collecting and analyzing COA data for regulatory decision-making, including selecting the COA-based endpoint and determining clinically meaningful treatment effects on that endpoint (Guidance 4)

Please refer to **Guidance 1, Guidance 2, Guidance 4,** and other FDA guidances[5] for additional information on patient experience data.

In conducting research that involves accessing patient experience data or directly engaging with patients, it is important to carefully consider Federal, State, and local laws and institutional polices for protecting human subjects and reporting adverse events. For additional information about human subjects' protection, refer to **section IV.A.2 of Guidance 1**.

FDA encourages stakeholders to interact early with FDA and obtain feedback from the relevant FDA review division when considering collection of patient experience data related to the burden of disease and the potential benefits, burdens, and harms of treatment.[6] FDA recommends that stakeholders engage early with patients and other appropriate subject matter experts (e.g., qualitative researchers, clinical and disease experts, survey methodologists, statisticians, psychometricians, patient preference researchers) when designing and implementing studies to evaluate the burden of disease and treatment, and perspectives on treatment benefits and risks.

In general, FDA's guidance documents do not establish legally enforceable responsibilities. Instead, guidances describe the Agency's current thinking on a topic and should be viewed only as recommendations, unless specific regulatory or statutory requirements are cited. The use of

---

[5] See FDA guidance for industry *Patient Preference Information—Voluntary Submission, Review in Premarket Approval Applications, Humanitarian Device Exemption Applications, and De Novo Requests, and Inclusion in Decision Summaries and Device Labeling* (August 2016), or subsequent guidances in the PFDD series, when available. Also see *Principles for Selecting, Developing, Modifying, and Adapting Patient-Reported Outcome Instruments for Use in Medical Device Evaluation* (January 2022). We update guidances periodically. For the most recent version of a guidance, check the FDA guidance web page at https://www.fda.gov/regulatory-information/search-fda-guidance-documents.

[6] In addition to the general considerations discussed in this guidance, a study may need to meet specific statutory and regulatory requirements governing the collection, processing, retention, and submission of data to the FDA to support regulatory decisions regarding a marketed or proposed medical product. This guidance focuses on more general considerations that apply to many types of studies, and you should consult with the review division and applicable regulations and guidance regarding any other applicable requirements.

the word *should* in Agency guidances means that something is suggested or recommended, but not required.

B. Purpose and Scope of PFDD Guidance 3

This document provides guidance that is generally applicable to COAs,[7] including patient-reported outcome (PRO), observer-reported outcome (ObsRO), clinician-reported outcome (ClinRO), and performance-based outcome (PerfO) measures.[8] Appendices A, B, C, and D include additional considerations for each type of COA, respectively, and multiple illustrations of conceptual frameworks.

This guidance is intended to help sponsors use high quality measures[9] of patients' health in medical product development programs. Ensuring high quality measurement is important for several reasons: measuring what matters to patients; being clear about what was measured; appropriately evaluating the effectiveness, tolerability, and safety of medical products. Findings from high quality measures may help support regulatory decision-making in a variety of contexts. For example, findings based on a well-defined and reliable COA-based endpoint[10] in an appropriately designed and conducted investigation may be used to support a claim in medical product labeling if the claim is consistent with the findings and the COA's documented measurement capabilities.[11]

The overall structure of this guidance is:

- Overview of COAs in clinical trials:
    - Describe four types of COAs
    - The role of COAs in evaluating clinical benefit

[7] For medical device submissions, the recommendations in this guidance should be implemented consistent with the least burdensome principles outlined in the guidance for industry and FDA staff *The Least Burdensome Provisions: Concept and Principles* (February 2019).

[8] For brevity in this guidance, the terms "PRO," "PerfO," "ClinRO," and "ObsRO" are used interchangeably with "PRO measure," "PerfO measure," "ClinRO measure," and "ObsRO measure".

[9] A measure is a means to capture data (e.g., a questionnaire) that includes clearly defined methods or procedures; instructions for administration or responding; a standard format for data collection; a well-documented method for scoring; and a method and/or criteria for interpreting results.

[10] Constructing COA-based endpoints is addressed in PFDD Guidance 4.

[11] The considerations addressed in this guidance may be relevant to a variety of regulatory decisions that require a benefit-risk assessment, including but not limited to: drug approval decisions under the standards in section 505(d) of the FD&C Act and regulations in 21 CFR part 314; device approval decisions under the standards in sections 513(a)(2) and 515(d) and regulations in 21 CFR part 814; device classification decisions under the standards in sections 513(a)(2) and 513(f) and regulations in 21 CFR parts 807 and 860; investigational new drug and investigational device exemption applications under sections 21 CFR parts 312 and 812; REMS and PMR requirements under sections 505-1 and 505(o)(3) and device post-approval requirements under 21 CFR part 814 subpart E; labeling decisions under 21 CFR parts 201, 801, and 809. Necessarily, this guidance does not attempt to capture all of the regulatory standards that might apply to a sponsor's intended plan of study; sponsors should consult the relevant review division(s) as necessary to discuss their study plans and are responsible for satisfying applicable requirements.

  - Specify what a COA assesses (the concept of interest)
  - Specify the purpose and context of the COA (the context of use)
  - Determine whether a COA has sufficient evidence to support its context of use, i.e., is fit-for-purpose[12]

- A general process, referred to as a Roadmap to patient-focused outcome measurement, to help guide the selection, modification, or development of a COA
- A discussion of components of a well-supported rationale to justify the COA's ability to assess the concept of interest for a specified context of use.

This guidance is informed by developments that have occurred in research and applications of COAs to derive clinical trial endpoints. Examples of these developments include the following:

- Patients and caregivers have been increasingly integrated as stakeholders in the development and evaluation of medical products.
- Several best-practice publications have described recommendations for developing and evaluating COAs, as well as analyzing and reporting COA data. Readers are directed to relevant publications throughout this guidance.
- The growing need for FDA guidance regarding all types of COAs has motivated the broader scope of this PFDD guidance series.
- The framework discussed in this guidance for development of well-constructed measures is based on developing evidence-based rationales. Several publications have described the development of evidence-based rationales (American Educational Research Association et al. 2014; Kane 2013; Weinfurt 2021, 2022). This validity framework is helpful for discussing the broad range of COAs addressed by this guidance and helps to clarify evidence that may be useful to support the rationale for using a particular COA.

This guidance distinguishes an endpoint from the COA, and the score produced by that COA. The COA includes any instructions, administration materials, content, formatting, and scoring rules. A COA score refers to any numeric or rated values generated by a COA through a standardized process. For example, a score could refer to:

- A rating assigned by a patient (PRO), clinician (ClinRO) or observer (ObsRO) describing the patient's functioning
- The result from a performance test (PerfO), such as grip strength measured in kilograms

[12] *BEST (Biomarkers, EndpointS, and other Tools) Resource*. 2016. https://www.ncbi.nlm.nih.gov/books/NBK338448/

- A combination of item responses assumed to measure some concept
- A combination of scores from multiple domains[13] to reflect some larger concept

A COA might produce scores on different scales (e.g., raw score, transformed score) and/or multiple scores that each correspond to a different concept (e.g., subscale scores). In contrast to a COA score, an endpoint is a precisely defined variable intended to reflect an outcome of interest that is statistically analyzed to address a particular research question. For a COA-based endpoint, the COA score is part of the endpoint definition. A complete definition of an endpoint typically specifies the type of assessments made; the timing of those assessments; the assessment tools used; and possibly other details, as applicable, such as how multiple assessments within an individual are to be combined (see Guidance 4 for a discussion of COA-based endpoints). For example, an endpoint might be the patient's average score from a daily symptom measure assessed for 7 days prior to the Week 12 post-randomization appointment.

II. OVERVIEW OF COAs IN CLINICAL TRIALS

A. Types of COAs

A COA is a measure that is intended to describe or reflect how a patient feels or functions. COA scores can be used to support effectiveness, dose optimization, safety, and tolerability in the context of a clinical trial to determine the clinical benefit(s) and risks(s) of a medical product. There are four types of COAs and choosing which type(s) of COA to use is driven by the concept(s) of interest to be measured, the best source of that measurement (e.g., self-report, clinician report/rating), and the context in which it will be applied (the context of use). More than one type of COA can be used in a clinical trial to capture the patient experience and the status of the patient's disease or condition.

The following are the four types of COAs:

- **Patient-reported outcomes** (PROs; see Appendix A)
  - Reports come directly from the patient
  - Useful for assessment of symptoms (e.g., pain intensity, chest tightness), functioning, events, or other aspects of health from the patient's perspective
- **Observer-reported outcomes** (ObsROs; see Appendix B)
  - Reports come from someone other than the patient or a health professional (e.g., a parent or caregiver) who observes the patient in everyday life
  - Useful when patients such as young children cannot consistently or accurately report for themselves, or to assess observable aspects related to patients' health (e.g., signs, events, or behaviors)
- **Clinician-reported outcomes**[14] (ClinROs; see Appendix C)

[13] Some COAs may assess a single domain such that all items or tasks measure a single concept. Other COAs may assess multiple domains such that each domain is a sub-concept represented by a subset of the COA items or tasks.

[14] Although reports of particular clinical events, such as stroke or pulmonary exacerbation of a chronic lung disease, may be determined by the assessment of a health care provider and therefore considered ClinROs, measures of such events are not discussed further in this guidance.

  - — Reports come from a trained healthcare professional after observation of a patient's health condition
  - — Useful when clinical judgement or interpretation is needed for reports of observable signs, behaviors, or other manifestations related to a disease or condition
- **Performance outcomes** (PerfOs; see Appendix D)
  - — Reports come from an individual trained to administer a particular task in a standardized fashion, an electronic assessment, and/or other standardized quantification of a patient's performance on standardized tasks
  - — A measurement based on a standardized task(s) actively undertaken by a patient according to a set of standardized instructions
  - — Useful when one wants to assess level of functioning. May also be useful when there is concern that patients cannot recall their level of functioning in daily life with sufficient accuracy

Another type of measure—a proxy-reported outcome measure—is discouraged by FDA. FDA defines a proxy-reported measure as an assessment in which someone other than the patient reports on patient experiences as if the reporter were the patient (see Appendix B). FDA acknowledges that there are instances when it is impossible to collect valid and reliable self-report data from the patient. In these instances, it is recommended that an ObsRO be used to assess the patient's observable behavior(s) rather than a proxy-reported measure to report on the patient's experience.[15]

When an electronic mode of administration (e.g., web-based application or an app on a mobile device) is used to collect a PRO, ObsRO, ClinRO, or PerfO, the source of measurement is still considered to be the patient, observer, clinical rater, or standardized task assessment, respectively. However, if the source of measurement is from a digital health technology (DHT) itself (e.g., a mobile sensor), please refer to the FDA guidance for industry, investigators, and other stakeholders *Digital Health Technologies for Remote Data Acquisition in Clinical Investigations* (December 2023).

Sometimes scores from several types of measurement are combined into a single score to form a multi-component endpoint. Discussion of such endpoints is beyond the scope of this guidance.[16]

[15] Instances where the assistance of another person is necessary to accommodate people living with disabilities to make their own report without any interpretation or intervention by the assistant are still considered PROs (e.g., use of a 'read-aloud' protocol for patients with visual impairments, or an assistant marking forms for a patient with motor control difficulties).

[16] For discussion of COA-based multi-component endpoints, see the draft guidance for industry, FDA staff, and other stakeholders *Patient-Focused Drug Development: Incorporating Clinical Outcome Assessments Into Endpoints for Regulatory Decision-Making* (April 2023) (When final, this guidance will represent the FDA's current thinking on this topic.)(PFDD Guidance 4) and, for CDER and CBER decision-making, see the guidance for industry *Multiple Endpoints in Clinical Trials* (October 2022).

B. The Role of COAs in Evaluating Clinical Benefit for a Medical Product

Clinical benefit is defined as "a positive effect on how an individual feels, functions, or survives".[17,18] To provide clinical benefit, a medical product should affect a meaningful aspect of health (MAH), i.e., some aspect of feeling or functioning in daily life that is important to patients (Walton et al. 2015). In a clinical trial, we study the effect of a medical product on a MAH by estimating a treatment effect on an endpoint that is thought to reflect the MAH.

To precisely describe the role of a COA in a clinical study, sponsors should propose to FDA how they intend to interpret scores from a COA (i.e., what they believe the score measures), how scores will be used to reflect the MAH (e.g., to construct an endpoint), and the context in which scores will be used. In other words, the sponsor's proposal should explicitly reference the MAH, the concept of interest (COI; see section II.B.2) and the context of use (COU; see section II.B.3). For COAs with multiple domains and related scores, the domain of interest (and particular score) should be clearly stated.

Having established the COA measures something meaningful, the construction of the COA-based endpoint should preserve the meaningfulness. Using COAs to construct trial endpoints is discussed in PFDD Guidance 4, *Incorporating Clinical Outcome Assessments Into Endpoints for Regulatory Decision-Making* (April 2023).[19]

1. *Meaningful Aspect of Health (MAH)*

The MAH could be a narrow concept such as nausea intensity or a broader aspect such as lower limb-related function. The MAH and COI may be identical or very similar. Sometimes, as in the case of many PROs, the COA scores can be interpreted as direct measures[20] of the MAH. In other cases, the COI only reflects a specific aspect of a broad MAH. For example, for the MAH of lower limb-related function, a PerfO such as the 6-Minute Walk Test might be used to measure functional walking capacity, which is only one of the several aspects of lower limb-related function.

2. *The Concept of Interest for Measurement*

The concept of interest is what is specifically measured by a COA to help understand how a medical product affects a MAH. Depending on the intervention, the intent of treatment may be, for example, to improve a symptom(s) or a specific function (e.g., ambulation); delay or avoid further worsening of a symptom(s) or further loss of a specific function; prevent the onset of a symptom or a loss of a specific function; or restore a specific function. Sponsors might also want to assess whether aspects of how patients feel and/or function could be negatively impacted by receipt of the intervention (i.e., harms). Any aspects of health that might be affected, positively

[17] *BEST (Biomarkers, EndpointS, and other Tools) Resource*. 2016. https://www.ncbi.nlm.nih.gov/books/NBK338448/

[18] This guidance addresses assessments related to how patients feel or function, not how patients survive.

[19] When final, this guidance will represent the FDA's current thinking on this topic.

[20] Here "direct" means that the COA is intended to measure the MAH rather than something not identical to the MAH. "Direct" does not imply that the COA measures without error or without interpretation by the patient.

or negatively, by the medical product could be targeted for assessment. As described above (see section II.B.1), the concept of interest for measurement is sometimes a direct reflection of the MAH (e.g., nausea intensity) and other times is less directly related to the MAH depending on the chosen assessment strategy.

The identification of concepts of interest that are both appropriate for, and important to, a given target population in CDER and CBER decision-making is described in Guidance 2[21] of this series. Such identification may involve qualitative research with patients, caregivers and clinical experts. For some diseases or conditions, important concepts of interest might have already been identified and used in previous studies based on input from patients, caregivers, clinical experts, and other sources. In such cases, sponsors should reference and summarize the prior work done when justifying their choice of concept(s) of interest.

In a clinical trial, it is important to carefully select concepts that:

- Reflect an aspect of health that is important to patients, i.e., a MAH
- Have the potential to demonstrate a clinically meaningful effect of the investigational treatment within the time frame of the planned clinical trial, when measured adequately and incorporated into an endpoint

3. The Context of Use

The context of use should clearly specify the way COA scores will be used as the basis for an endpoint intended to reflect a specific MAH. The appropriateness of a COA is evaluated within the proposed context of use. During the course of a development program, some elements of the context of use will be established early on, such as the target population, and others (e.g., trial design, timing of assessments) might evolve, for example, through discussions with FDA and/or as different COAs are considered.

Context of use considerations may include the following:

- **Target Population**: Including a definition of the disease or condition; participant selection criteria for clinical trials (e.g., baseline symptom severity, comorbidities, patient demographics and cultures); and expected patient experiences or events during the trial (e.g., that some patients will require assistive devices).
- **Use of the COA**: Clinical trial objectives and how the COA will be used to support a COA-based endpoint intended to reflect a specific MAH (e.g., computing the mean COA score at 12 weeks).
- **COA Implementation**: Including the location where the COA is collected (e.g., inpatient hospital, outpatient clinic, home); how the COA will be collected (i.e., mode of administration, such as electronic data capture, paper form); and by whom (e.g., patient, study coordinator, investigator, parent/caregiver).

The interpretability of COA scores and COA-based endpoints depends on the points above as well as others, including:

[21] See the guidance for industry, FDA staff, and other stakeholders *Patient-Focused Drug Development: Methods to Identify What Is Important to Patients* (February 2022) (PFDD Guidance 2).

- **Clinical Trial Design**: The study design in which the COA is to be used, including the type of comparator group and whether those providing responses or participating in the tasks for the COA (patients, observers, clinicians, trained raters) are masked[22] with respect to treatment assignment and, in the case of some COAs, study visit.
- **Schedule** for administration(s) of the COA(s).

Discussions with FDA about the selected COA and its interpretation within a medical product development program may also include the medical product's mechanism of action and other topics. Some of these topics are discussed further in Guidance 4.

C. Determining Whether a COA Is Fit-for-Purpose

A COA is considered fit-for-purpose when "the level of validation associated with a medical product development tool is sufficient to support its context of use".[23] Whether a COA is fit-for-purpose is determined by the strength of the evidence in support of interpreting the COA scores as reflecting the concept of interest within the context of use. It is expected that both qualitative and quantitative sources of evidence may be needed to support a determination that a COA is fit-for-purpose.

Decisions about whether a COA is fit-for-purpose are based on two considerations:

1. The Concept of Interest and Context of Use Are Clearly Described

Sponsors should make clear how they intend to interpret the COA scores as measures of the concept of interest within the context of use. The statement should explicitly specify the concept of interest and the context of use in enough detail to describe clearly how the COA and its scores are intended to be used.

2. There Is Sufficient Evidence to Support a Clear Rationale for the Proposed Interpretation and Use of the COA

Regardless of whether sponsors propose to use an existing COA, a modified COA, or a newly developed COA, sponsors should present a well-supported rationale for why the proposed COA should be considered fit-for-purpose. The rationale is a set of reasons supported by evidence submitted to the FDA and/or cited (see section IV and Appendix E).

The rationale may have multiple components (see section IV, Table 1) and each component should be justified by one or more sources of evidence, including, for example, literature reviews; natural history studies; qualitative studies with patients, caregivers, or other stakeholders; and quantitative studies.

---

[22] Keeping study group assignment hidden from those involved in a study is commonly referred to as "blinding" or "masking." Those who do not know the assignment are referred to as "blinded" or "masked." The term "masked" is used in this guidance.

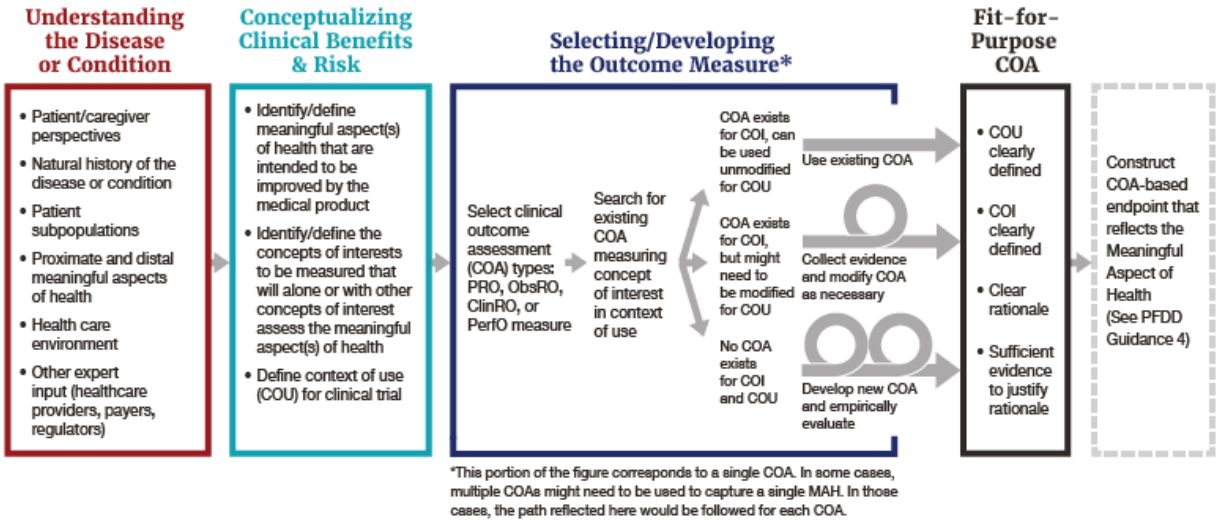
[23] *BEST (Biomarkers, EndpointS, and other Tools) Resource*. 2016. https://www.ncbi.nlm.nih.gov/books/NBK338448/

To determine whether sufficient justification has been provided for the rationale, FDA will review each part of the rationale and assess whether an appropriate type and amount of evidence has been presented. The evidence for a particular part of the rationale is weighed relative to the degree of uncertainty about that part. The greater the uncertainty, the greater the need for additional evidence to support that part of the rationale. In addition to the degree of uncertainty about each part of the rationale, FDA considers the context of use, and may consider the broader impact on the target population and medical product development of collecting additional evidence when determining whether a COA is fit-for-purpose. Section IV provides guidance about how to develop a clear rationale with supporting evidence.

III. A ROADMAP TO PATIENT-FOCUSED OUTCOME MEASUREMENT IN CLINICAL TRIALS

This section describes a general Roadmap to patient-focused outcome measurement in clinical trials (see Figure 1). Sponsors and COA developers are not required to use this approach, and it may not fit every development program, but it has worked well for several COAs in many contexts of use. FDA recommends sponsors seek FDA input as early as possible and throughout medical product development to ensure COAs are appropriate for the intended context of use.

Figure 1 Roadmap to Patient-Focused Outcome Measurement in Clinical Trials



For simplicity, the Roadmap in Figure 1 portrays decisions occurring in stepwise fashion. However, decisions about the concept of interest for measurement, the context of use, and the type of COA might be interrelated. For example, consider the MAH of pain intensity. If the target population is adolescents who are capable of self-report, then a PRO could be used to assess the concept of interest of pain intensity. But if the target population was children ≤ 2 years old, reliable self-report is not possible, which also means that a direct assessment of the child's pain intensity is not possible. Instead, the concept of interest might need to be observable pain behaviors assessed using an ObsRO.

A. Understanding the Disease or Condition

The first step involves considering the manifestations and natural history of the disease or condition; important patient subpopulations; heterogeneity among patients in how the disease or condition manifests; and the clinical environment in which patients with the condition seek care. It is critical at this stage to collect or cite patient and/or caregiver perspectives on the disease, its impacts, and therapeutic needs and priorities.

One important outcome of this step is understanding and summarizing the important signs, symptoms, and health impacts patients with the disease or condition might experience. Sponsors might find it useful to develop a disease model that represents how a disease or condition affects bodily structures and/or processes and all the resulting effects on the patients in terms of MAHs. In addition to input from clinical experts, qualitative data from patients and/or caregivers (e.g., from individual interviews or focus groups; Patrick et al. 2011a) play an essential role in ensuring that the MAHs associated with the disease or condition are identified and clearly described.

B. Conceptualizing Clinical Benefits and Risks

The next step involves considering which MAHs will be targeted by the medical product. This consideration leads to identifying the concept(s) of interest to be measured (see section II.C.1) and the context of use (see section II.C.2), including the population of interest, clinical trial design, and the trial objective and endpoints.

Often, a single disease or condition is associated with many MAHs. For example, a condition that causes chronic pain may also be associated with fatigue and impacts on physical and social functioning. To help focus a medical product development program intended to demonstrate effectiveness[24], sponsors should identify, where possible, the primary manifestations of a disease or condition (i.e., proximal or core aspects of a disease or condition) that are relevant and important to patients.[25] MAHs might represent the downstream (i.e., distal) impact of these core aspects on other aspects of how a patient feels or functions.

For example, when evaluating a treatment for the management of moderate to severe endometriosis-associated pain, it would be important to assess a core concept of interest such as dyspareunia, defined as pain with intercourse. In addition, the impact of moderate to severe endometriosis-associated pain severity on daily activities could also be assessed.

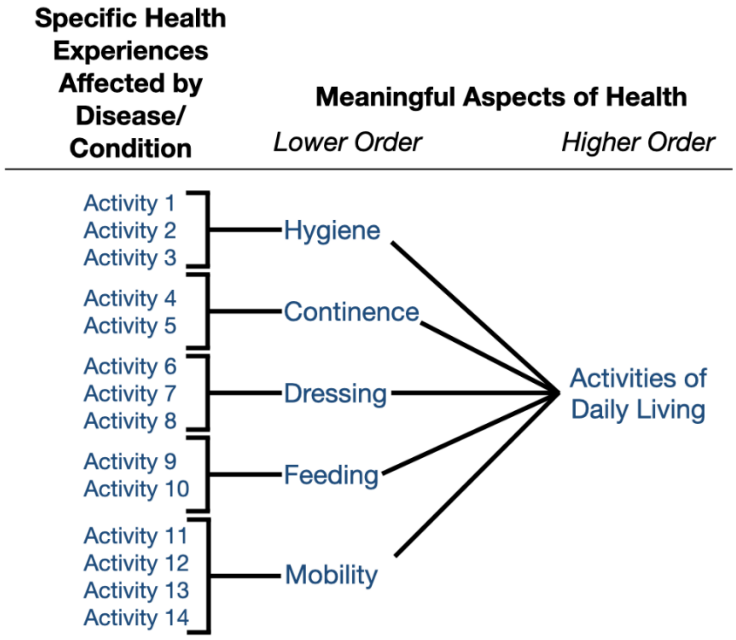
To communicate more complex MAHs to FDA, it might be helpful for sponsors to develop a representation of the MAH and the types of patient experiences it summarizes. Developing a working version of such a representation might also be helpful during the collection of qualitative and other data to inform an understanding of the MAHs. Figure 2 displays a representation for the hypothetical MAH of activities of daily living (ADLs). In the figure,

[24] A similar process may also be used for COAs to evaluate tolerability or safety.

[25] Alternatively, if a medical product is intended to address a secondary manifestation(s) of a disease (and not address the primary manifestation), sponsors should justify why the medical product would be of value to patients. Considerations for selecting COA-endpoints are found in PFDD Guidance 4.

specific health experiences of the patients (Activities 1-14) are conceptualized in terms of five different lower order MAHs—hygiene, continence, dressing, feeding, and mobility. For example, the activities collected under "mobility" might include getting in and out of bed, being able to stand from a sitting position, walking across a room, etc. These five lower-order MAHs together make up a higher order MAH known as ADLs. For relatively simple and narrow MAHs, such as pruritus (itch) severity, a simple definition might suffice without a more elaborate representation.

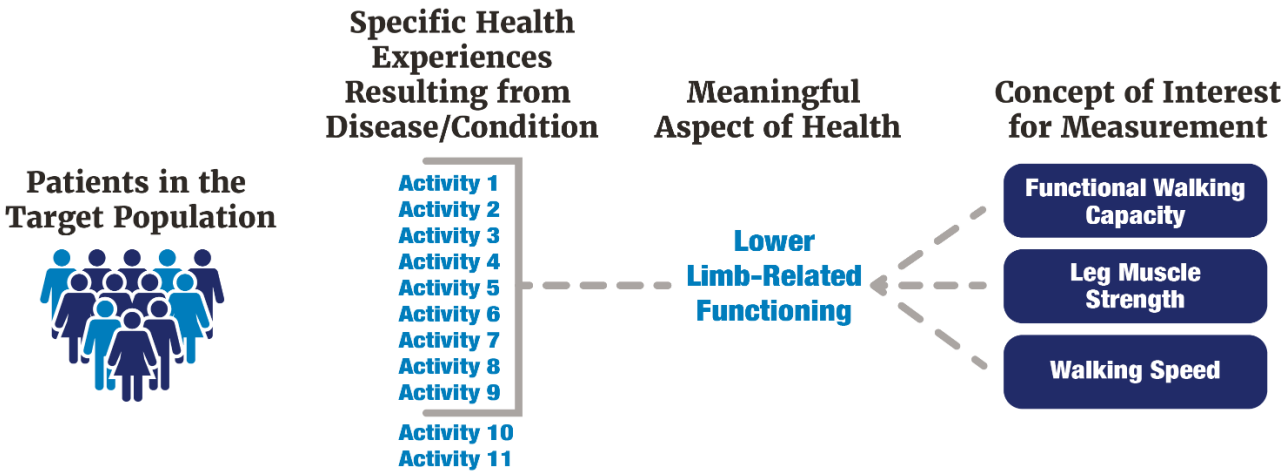
Figure 2 Example of Representing a More Complex Meaningful Aspect of Health: ADLs



After identifying the MAH(s) that is the focus of the assessment, sponsors can consider the specific concept(s) of interest for measurement which reflect that MAH(s). For a broad MAH which is related to multiple COIs, these relationships might be communicated more clearly by a graphical representation. Figure 3 shows a hypothetical example using the broad MAH of lower limb-related functioning and the three concepts of interest the sponsor intends to measure to make an inference about lower limb-related function (Walton et al. 2015). Viewing the figure from left to right the patients in the target population have a variety of specific health experiences that may be affected by their disease or condition. Activities 1 – 9 comprise the MAH and cover three different COIs for measurement. It is also useful at this point to develop initial thoughts about the corresponding COA-based endpoint in general terms (e.g., patient's status at a fixed time point, change in status, time-to-event).

In some situations, there may be health experiences that while important to the patient are not going to be assessed in a clinical trial, represented in the figure by Activities 10 and 11.

Figure 3 Example of Representing a Meaningful Aspect of Health and Its Corresponding Multiple Concepts of Interest for Measurement



C. Selecting/Developing the Outcome Measure

There are several steps involved in selecting, modifying, or, if necessary, developing a COA to measure the concept of interest.

1. *Selecting the COA Type*

Sponsors and measure developers should consider what type of COA is most appropriate for assessing the concept of interest in the context of use. Considerations for selecting a specific type of COA are discussed in section II.A and in Appendices A-D.

2. *Evaluating Existing and Available COAs Measuring the Concept of Interest in the Context of Use*

FDA recommends conducting a search to identify whether a COA already exists that measures the concept of interest in the intended context of use and is available for use. Existing COA measures for which there is already experience in the relevant context of use are generally preferred, particularly when measuring well-established concepts of interest (e.g., pain intensity). Sponsors can identify potential measures by searching the scientific literature; repositories of measures, including item banks[26] comprising previously developed and tested items; and clinicaltrials.gov, summaries of prior FDA decisions, and other resources (FDA COA Qualification Program; FDA Medical Device Development Tools [MDDT]). Ultimately, Sponsors should ensure that there is sufficient evidence to support the use of such COAs within the intended context of use in the planned clinical trial. Sufficient evidence may vary where there is residual regulatory uncertainty such as in the context of a rare disease.

[26] For assessments developed using Item Response Theory, an item bank is a collection of items intended to measure the same concept. All the items in the item bank have been shown to fit well within an Item Response Theory model and any and all items can be used to estimate a score for a respondent on the concept of interest.

The following sections describe the potential outcomes of a search.

a. An Appropriate COA Exists for the Concept of Interest in the Same Context of Use: Use Existing COA

If a COA exists to assess the concept of interest in the same or similar context of use as intended in the sponsor's trial, the sponsor should: assess its fitness for purpose; provide the rationale for selection of the COA; and summarize the evidence that supports that rationale.[27]
There are times when an existing COA may not have all the evidence recommended to support its use because the COA is still under development, was developed a long time ago, or for other reasons. For example, some types of studies (such as qualitative) may not have been conducted for some of the target population (e.g., adolescents) or some documentation may not be available for some steps in the development. Sponsors should summarize existing information and evidence that supports the rationale for the use of the COA and assess how well the rationale is supported by the available information. In some instances, adequate evidence may be found in the literature or available clinical trial data, while in other instances, it may be necessary to collect additional evidence for the rationale before the COA can be considered fit-for-purpose. COAs being used in registries, natural history studies, or observational trials may or may not be fit-for-purpose in other contexts of use. Sponsors should ensure that there is sufficient evidence to support the use of such COAs within the intended context of use in the planned clinical trial.

b. A COA Exists for the Concept of Interest for a Different Context of Use: Collect Additional Evidence and Modify COA as Necessary

If a COA exists that assesses the concept of interest but was not developed for the sponsor's context of use (e.g., was not developed for the same target population), then the sponsor should evaluate whether the COA can be used in the different context of use and provide supporting evidence or explanations supporting the new context of use. Evidence presented in prior work on the COA may suffice to support the rationale for its use in the new context of use. Alternatively, if the existing evidence leaves too much uncertainty about the appropriateness of use in the new context of use, we recommend the collection of additional evidence. Depending on the situation, such evidence might include conducting cognitive interviews (also known as cognitive debriefing or testing) in the new target population to confirm content relevancy and understanding of the items and responses. Sponsors should discuss with FDA the type and amount of additional evidence that might be needed to support the rationale.

A sponsor may also consider modifications intended to improve the COA's ability to reflect the concept of interest or to improve data collection of the COA. Modifications could include, but are not limited to, changes to:

- Instructions/training materials
- Item/task content (e.g., omitting, adding, or modifying wording of items and/or response options; translating from one language to another; modifying the activity performed for a PerfO)

[27] Note that if a COA exists for the concept of interest in the same context of use, sponsors are not required to use that COA. For example, a sponsor could choose a less burdensome COA if it can be justified as fit-for-purpose.

- Order of the items/tasks
- Recall period
- Format of the measure or mode of administration (e.g., paper or electronic device)
- Method of scoring, including changes to the scoring algorithm

The sponsor should carefully consider the impact of the proposed modifications to an existing COA and whether the measure is subject to any copyright restrictions, if applicable. Depending on the alteration (and extent of alteration) of the COA, this could create a new measure and result in altering the measure's scores and/or their interpretation. Some modifications are unlikely to alter the scores or their interpretation (e.g., changing from multiple items on a page on a paper-based PRO measure to a single item per screen on a tablet; O'Donohoe et al. 2023), whereas other changes are likely to affect scores and their interpretation (e.g., changing the recall period from 1 day to 7 days or from 7 days to 1 day). In the latter case, the modification may, in effect, create a new measure. The amount and type of evidence (qualitative and/or quantitative) to support modifications of a COA will depend on the type of changes that are proposed and the way in which the new context of use differs from the one for which the COA was originally developed.

For COAs that may be administered in a different mode, empirical studies conducted across a range of settings generally support the comparability of measurement properties across different modes of administration (e.g., paper-based, tablet computer, or patient's own device; see O'Donohoe et al. 2023 for summary). Thus, additional supportive evidence for comparability across modes is unlikely to be necessary if (a) the changes are small to moderate (e.g., changes in item format or presentation, or change from paper-based to interactive voice response) and (b) sponsors followed best practices for migrating measures to different assessment platforms (Critical Path Institute ePRO Consortium 2018a, 2018b, Byrom et al. 2019; Eremenco et al. 2014; Mowlem et al. 2024, Romero et al. 2022).

References are available that address considerations for modifying a COA (e.g., Houts et al. 2022; Papadopoulos et al. 2019; Rothman et al. 2009; and O'Donohoe et al. 2023).

c. No COA Exists for the Concept of Interest: Develop a New COA and Empirically Evaluate

It is beyond the scope of this guidance to provide specific recommendations for developing all types of COAs, but helpful references that address measure development are provided at the end of this guidance (e.g., Cappelleri et al. 2014; de Vet et al. 2011; Fayers and Machin 2016; Streiner et al. 2015). There are general principles regarding the development process for any type of new COA:

- Involve patients and/or caregivers, and clinical experts, as collaborators when developing new COAs.
- Submit proposed qualitative study materials, including protocols, structured and/or semi-structured interview guides, and/or observation checklists for FDA review and comment prior to beginning the qualitative research.

- Clearly document all steps and data collected in the development process such as including an item/task tracking matrix that describes the history of the development and modification of all items/tasks, and transcripts from any qualitative interviews.
- Submit the proposed analysis plan (e.g., psychometric analysis plan) to develop and provide scientific-based evidence to support the rationale for interpreting COA scores as a measure of the concept of interest in the context of use (discussed in section IV) for FDA review and comment prior to conducting analyses.
- When collecting data from patients and/or caregivers, use study samples that reflect the clinical characteristics and demographics of the target patient population.
- Consider and evaluate potential limitations of the proposed COA. For example, could measurement of the concept of interest be affected by processes or concepts not part of the concept of interest (see section IV.E)?
- Create a user manual for the COA describing how to administer the measure. For most types of COAs, it is important to create training materials (e.g., for investigators and other study personnel, patients, observers, or clinicians) so that assessments are conducted in a consistent way.
- Document the method of scoring the COA, including how missing items/tasks should be handled. There should be clear justifications for the approach to scoring and addressing missing data.

When the sponsor is developing or significantly modifying a COA, in general, FDA does not recommend evaluating the COA for the first time in a registration[28] trial, because it may be too late to learn that the COA is not performing as it should, potentially jeopardizing the success of a medical product development program. Early phase trials conducted prior to the registration trial represent an opportune time to evaluate measurement properties of COAs, and sponsors are encouraged to include prospectively planned analyses to inform subsequent trials.[29] If this is not a feasible option, FDA recommends conducting a standalone observational study prior to the initiation of a registration trial(s) to aid in the development of fit-for-purpose COAs. Furthermore, using data from the observational study to evaluate a proposed COA prior to the registration trial may reduce the risk[30] of using a COA that may not perform as expected, and therefore may not be fit-for-purpose.

FDA encourages precompetitive collaboration when developing a new COA such that the COA can be shared among sponsors, researchers, and patient advocacy groups to promote efficiency and to maximize the returns on the efforts made by patients who cooperated in its development.

3. Special Considerations for Selecting or Developing COAs for Pediatric Populations

If the concept of interest can be reliably measured across the age spectrum of the trial patient population, we recommend using one version of a COA for all pediatric patients in a study.

---

[28] In this guidance, *registration trials* are used to stand for what different groups call pivotal trials, confirmatory trials, and clinical trials for marketing authorization.

[29] Sponsors should also use data from later clinical trials to confirm, to the extent possible, the measurement properties evaluated in earlier phase trials.

[30] The generalizability of results from an observational study to a registration trial will depend, in part, on the similarities between the patient populations.

Including multiple versions of a COA for patients with differing abilities in the same trial may introduce challenges such as scores that do not reflect the same concept of interest or scores affected by a large amount of measurement variability. In these cases, the same scores from different pediatric patients cannot be used and interpreted interchangeably. However, depending on the concept of interest, at times it may be necessary to use multiple versions of a COA, because assessment of the target concept may differ substantially across the developmental spectrum (e.g., gross motor functioning in infants and adolescents). It may also be necessary to use multiple types of COAs to measure a concept with pediatric patients in a valid way. For example, to measure the MAH of itching in pediatric patients across the full age range, a caregiver ObsRO (assessing the concept of interest of scratching behavior) could provide one version of measurement across all patients while a PRO (assessing the concept of interest of itch intensity) could provide patient-report for a subset of patients who could validly self-report.

When pediatric versions of PROs or PerfOs are feasible, they should be completed by the child independently, without any assistance from caregivers, investigators, or anyone else, to avoid influencing the child's responses. Computer-administration, including automated reading of items, using a touch screen, or games, may make it easier for children to self-report. Self-administration and self-report may not be suitable with younger children and therefore might call for alternative approaches, such as interviewer-administration by a trained interviewer and/or an ObsRO.

Young children and some children with neurodevelopmental differences may be limited in their understanding of certain response scales used in a PRO (e.g., a 0 to 10 numeric rating scale, more/less comparison, references to periods of time). Simplified response scales (e.g., scales with few and simple response options, broadly culturally acceptable and interpretable pictorial scales) should be considered for use. Supporting evidence for the suitability of a COA for specific pediatric populations should address the level of comprehension of vocabulary, language used (e.g., instructions), the target concept, and relevance of the recall period.

References are available that discuss measurement in pediatric patient populations (e.g., Arbuckle and Abetz-Webb 2013; Bevans et al. 2010; Matza et al. 2013; Papadopoulos et al. 2013). Also, refer to PFDD Guidance 2, section VI (Managing Barriers to Self-Report) for considerations on how to obtain input from pediatric patients, their caregivers, and treating health care providers.

4. COA Accessibility and Universal Design

A portion or all of the target population may benefit from accessibility features and universal design[31] considerations. Usability testing is recommended for accessibility features for a COA under certain situations (e.g., event-triggered data collection; see guidance for industry and FDA staff *Applying Human Factors and Usability Engineering to Medical Devices* (February 2016) for guidance on CDRH decision-making). The following resources are available to ensure the COA is accessible for patients with impairments (e.g., vision impairment/low vision, hearing impairment/deaf or hard of hearing):

[31] In the context of COAs, *universal design* is consideration for the design and composition of a COA so that it can be accessed, understood, and used to the greatest extent possible by all people, inclusive of people with disabilities.

- The World Wide Web Consortium (W3) has a Web Accessibility Initiative (WAI) with resources and recommendations for making electronically delivered material more accessible to people, see https://www.w3.org/TR/WCAG22/ and https://www.w3.org/TR/low-vision-needs/.
- Section 508, a U.S. Government website, has resources addressing universal design, including color universal design, creating accessible portable document formats (PDFs), and other topics: https://www.section508.gov/.

Options including assistive technology that may be used by participants, such as screen readers, can allow patients and/or their caregivers to provide reliable reports. Consider which modifications and/or assistive technologies might be useful to assist broad inclusion in COA development, evidence generation, and trials. While options that minimize the involvement of third parties are optimal, there may be cases where a patient requires an accommodation of assistance from another person, such as to 'read aloud' in cases of limited literacy or to mark a form on their behalf due to fine motor disability. In such situations every attempt should be made to standardize procedures and minimize intervention by third parties but still allow patients to provide direct reports.

D. Developing a Conceptual Framework

The Roadmap describes a recommended path sponsors can take to arrive at a fit-for-purpose COA and, ultimately, an endpoint constructed from scores on the COA. Throughout this path, it would be helpful to communicate with FDA by describing the COA-based endpoint approach (see Table 1) that includes:

- Each MAH targeted by the intervention
- The corresponding concept(s) of interest for measurement
- The type of COA being used to assess the concept of interest
- The name of the COA (and score if the COA can generate more than one score), and
- The COA-based endpoint

The COA-based endpoint approach might begin as a tentative representation with multiple options for measurement being investigated and/or multiple methods of constructing an endpoint, which then evolves into a final endpoint approach at the end of the Roadmap[32].

Table 1 Suggested Format for Communicating the COA Based Endpoint Approach

		COAs			
Meaningful Aspect of Health	**Concept of Interest for Measurement**	*Type*	*Name*	*Score*	**COA-based Endpoint**

[32] For guidance on constructing COA-based endpoints, see PFDD Guidance 4.

IV. DEVELOPING THE EVIDENCE TO SUPPORT THE CONCLUSION THAT A COA IS APPROPRIATE IN A PARTICULAR CONTEXT OF USE

Evidence collected in support of the use of a COA should support the rationale that explains how and why the specific COA is expected or intended to work. It is important for FDA to understand each part of a sponsor's rationale and the evidence being offered in support of each part. This understanding will help facilitate efficient conversations between FDA and sponsors or measure developers.

This section describes eight components (see Table 2) that should be considered for inclusion in the rationale and supporting evidence or justification section of submissions to FDA. The discussion below also includes possible sources of evidence to evaluate each component. Different trials and contexts of use might call for different rationale components and/or evidence to support a COA as fit-for-purpose. That is, some cases might require fewer components and other cases might require more components than those described here. Note that some types of studies might supply evidence to support more than one component. For example, a qualitative study using cognitive interviews involves asking patients how they understand items from a COA and arrive at their responses (Willis 2005, Willis 2015, and Patrick et al. 2011b) or asking patients and assessors how they interpret instructions for a PerfO. Data from such a study might be used to support components C, D, and E in Table 2.

Table 2 Eight Components Comprising an Evidence-Based Rationale for Proposing a COA as Fit-for-Purpose

A	The concept of interest should be assessed by [*COA type*] because . . .
B	The COA selected captures all the important parts of the concept of interest.
C	The COA is administered appropriately.
D	Respondents understand the instructions and items/tasks of the measure as intended by the measure developer.
E	The method of scoring responses to the COA is appropriate for assessing the concept of interest.
F	Scores from the COA are not overly influenced by processes/concepts that are not part of the concept of interest.
G	Scores from the COA are not overly influenced by measurement error.
H	Scores from the COA correspond to the meaningful aspect of health related to the concept of interest.

Note: Listed components are those that are likely but not necessarily needed in the rationale for a specific COA, concept of interest, and context of use. Each rationale should be tailored to the proposed interpretation and use. Each component should be accompanied by comprehensive supporting evidence and justification.

Prior to presenting the evidence-based rationale, sponsors should clearly describe the following:

- **Intended context of use**
- **Description and justification for the MAH:** Sponsors should report or cite data collected from patients and/or caregivers on the nature and importance of the MAH (see sections III.A and B)
- **Concept of interest:** If not identical to the MAH, then sponsors should explain and justify how measurement of the concept of interest helps us to understand the MAH. For

example, when there are several concepts of interests that together make up the MAH, the sponsor could explain how each concept of interest contributes and is necessary to fully support an inference about the MAH.

The evidence-based rationale is used to justify interpreting COA scores as measures of the concept of interest. Ensuring that the COA scores reflect the concept of interest is a necessary part of the endpoint rationale for any endpoint constructed using the COA scores. In other words, the rationale for interpreting the COA scores presented here becomes part of the larger endpoint rationale for a COA-based endpoint (The rationale for the endpoint is discussed in PFDD Guidance 4).

Appendix E provides a suggested format for summarizing the rationale and its corresponding evidence.

A. The Concept of Interest Should Be Assessed by [COA Type], Because . . .

The sponsor should provide a clear rationale for the type of COA (i.e., PRO, ObsRO, ClinRO, or PerfO) selected to assess the concept of interest. Considerations for selecting the specific type of COA are discussed in section II.A and Appendices A-D.

B. The COA Selected Captures All the Important Parts of the Concept of Interest

All important parts of the concept of interest should be covered by the chosen COA.[33] This includes the specific characteristic(s) of interest, such as frequency, intensity, or duration. For narrow and simple concepts that can be assessed with a single item (e.g., worst pain intensity over the past 24 hours), it is straightforward to see whether the item content covers the concept of interest. For more complex concepts of interest that include multiple parts (for example, physical function), all important parts should be reflected in the content of the COA. Similarly, the setup and tasks included in a PerfO should cover all important parts of the function being evaluated as the concept of interest.

The source of evidence for this part of the rationale depends on how closely the concept of interest reflects patients' lived experiences of their disease (Edgar et al. 2023; Weinfurt 2023):

- If the concept of interest closely reflects patients' daily lived experiences with their disease, then patients' input regarding the items/setup/tasks (e.g., through qualitative studies) is important.

- If the concept of interest less closely reflects patients' daily lived experiences with their disease, then the items/setup/tasks might not resemble patients' daily lives. In these cases,

[33] How well a measure reflects all important aspects of a concept of interest was previously referred to as *content validity*. The field of measurement, as reflected by the 2014 Standards for Psychological and Educational Testing, has moved from talking about different types of validity to specifying different sources of evidence. Validity is understood as a unitary concept and refers to the "degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association et al. 2014, p. 11), where tests in this case refer to COAs.

other expert input (e.g., clinicians, physiologists, psychologists) might be needed to judge how well the items/setup/tasks capture the concept of interest.

C. The COA is Administered Appropriately

This component is relevant for any COA in which someone other than the respondent administers the COA or assists the respondent in completing the COA (e.g., research personnel who administer the tasks as part of a PerfO). Support for this component could include a user manual that is clear and detailed, evidence of successful completion of a standardized training program by personnel at all sites, and intermittent observation of personnel throughout the clinical trial to ensure ongoing adherence to the protocol.

D. Respondents Understand the Instructions and Items/Tasks of the Measure as Intended by the Measure Developer

For PROs, ObsROs, and ClinROs, the most straightforward type of support for component D is in the form of cognitive interviews (Willis 2005, Willis 2015, and Patrick et al. 2011b). It is important in such studies to clearly document the intended meaning of each item, response option, and instruction so that the respondents' understandings can be compared to these intended meanings. For PerfOs, cognitive interviews with patients regarding task instructions combined with pilot testing tasks can confirm whether patients understand the task they are asked to do.

When collecting evidence in support of component D, sponsors should include perspectives of reporters who reflect the range of literacy and numeracy in the target population.

We also recommend that measure developers follow good practices in questionnaire design to avoid common pitfalls that could interfere with respondent understanding (e.g., avoiding double-barreled items, which ask about more than one thing within a single item) (see section IV in PFDD Guidance 2). Additionally, involving patients and/or caregivers and their treating health care providers in the creation of items for PROs and ObsROs can make it more likely that patients and/or caregivers will understand the items as intended.

E. The Method of Scoring Responses to the COA is Appropriate for Assessing the Concept of Interest

Every COA provides some way for responses to be recorded or coded as an observed score for a prompt. For example, a PRO that assesses current nausea intensity might allow patients to record their responses on a verbal rating scale with four adjectives (e.g., none, mild, moderate, severe), which the PRO's simple scoring rule converts into corresponding observed scores between 0, 1, 2, or 3. A walking test might record the distance (or time) a patient walks for a specified time (or distance), producing an observed score in meters (or seconds). In both examples, the appropriateness of the scoring depends on key assumptions, which are discussed below.

1. Responses to an Individual Item/Task

For an individual item/task, response options should be non-overlapping and differences among adjacent response categories should reflect true gradations in the concept of interest in the target population. The wording of the response options should be clear and concrete, and the instructions for making or recording the responses should be clearly understandable. Support for these considerations can come from cognitive interviews, demonstrating that respondents have no difficulty selecting an answer that matches their experience.

FDA generally does not recommend the use of a visual analog scale (VAS). There are known limitations with its administration (e.g., cannot be administered verbally or over the phone; electronic rendering may lead to different lengths of lines displayed during a single trial) and interpretability (e.g., higher rates of missing data or incomplete data) (Dworkin et al. 2005 and Hawker et al. 2011). Instead of a VAS, consider using verbal rating scales (e.g., *none*, *mild*, *moderate*, or *severe*) or numeric rating scales (e.g., 0 to 10 rating of pain intensity).

2. Rationale for Combining Responses to Multiple Items/Tasks

If multiple items/tasks are combined to generate a score on a COA, then the rationale for the method of scoring should be described and supported with evidence (Edwards et al. 2017). The approach for combining responses to multiple items/tasks is often expressed as a measurement model that relates responses to particular items/tasks to the score(s) thought to measure the concept of interest. There are different types of measurement models that might be appropriate for a COA. The rationale and justification for scoring a multi-item/task COA will depend upon the type of measurement model. Therefore, sponsors should specify the measurement model when supporting the method of scoring the COA.

Two of the more common models underlying the scoring of COAs are the reflective indicator and composite indicator models.[34]

- **Reflective Indicator Model[35]:** The reflective indicator model combines responses to items/tasks where all the responses reflect, or are caused by, a single aspect of the patient's health described by the concept of interest (Fayers and Machin 2016).[36] In Figure 3 to measure the MAH of lower limb-related mobility, a measure developer might create a PRO consisting of items corresponding to different activities that require varying degrees or features of lower limb-related mobility (e.g., walking a block, climbing one

[34]A third type, the causal indicator model (Bollen and Bauldry 2011; Fayers and Machin 2016), is seldom seen in COAs used for regulatory submissions, though it might be appropriate for some COAs. Sometimes both causal indicator and composite indicator models have also been referred to as formative measurement models (Fayers and Machin 2016).

[35] This is also known as the effect indicator model (Bollen and Bauldry 2011).

[36] Some PROs based on a reflective indicator model consist of multiple items assessing multiple domains. For such measures, if the multiple domains will be used to assess the concept(s) of interest, a rationale should be given supporting the conceptual distinctiveness of the different domains and psychometric analyses should be provided in support of the assumed dimensionality of the measure (e.g., demonstrating adequate fit of a confirmatory factor analysis model that includes the multiple domains).

flight of stairs). The patient's lower-limb related mobility is reflected by the responses to these items.

COAs based on a reflective indicator model might use different measurement frameworks (e.g., Classical Test Theory, Item Response Theory, Rasch Measurement Theory) for scoring (Petrillo et al. 2015). Sponsors or measure developers should be clear about the measurement framework and the corresponding psychometric model that is assumed (e.g., True Score Model, Partial Credit Model, Samejima's Graded Response Model, Polytomous Rasch Model) and provide statistical evidence in support of model assumptions and fit, as well as relevant model parameters.[37] Note that FDA does not endorse any particular psychometric modeling approach but will review the strength of evidence in support of a model's use in specific cases.

- **Composite Indicator Model[38]:** A composite indicator model assesses a concept of interest using multiple items that, taken together, define the concept of interest (e.g., a set of everyday tasks that are labeled basic activities of daily living; Bollen and Bauldry 2011). In Figure 2 the higher order MAH of basic activities of daily living might be defined by a sponsor or measure developer as the degree to which the patient is able to accomplish everyday tasks that are necessary to live independently. The item content or activities that define everyday tasks might be determined through a consensus process with patients and their caregivers and could result, for example, in items addressing personal hygiene, dressing oneself, toileting, eating, and ambulation. Note that although it is likely that some of the item responses will be associated with one another, it is not necessary, because it is not assumed that all the items are reflective of or caused by a single, underlying concept of interest as was the case for the reflective indicator model. Rather, these items, known as composite indicators, are like the ingredients of what is labeled basic activities of daily living.

  For COAs based on a composite indicator model, sponsors or measure developers should describe and justify the process for selecting the items that make up the measure (e.g., by citing a consensus process with patients and others). A rationale should also be given for the way in which responses to the multiple items are combined to arrive at a score for the COA, including providing a rationale for any score transformation (e.g., log or linear) or normalization (z-score). For example, one might justify taking the sum of the item responses (which implies they are all weighted equally) based on qualitative or quantitative evidence that patients felt that all the activities described by the items are equally important for daily, independent living.

A COA may have characteristics from both of the above models, or their scoring may be a blend.

[37] COAs based on psychometric models from Item Response Theory might sometimes be administered using computerized adaptive testing (CAT) procedures, whereby the next item administered to a respondent depends upon a running estimate of the respondent's status based on the respondent's responses to prior items (Wainer 2000). Guidance regarding CAT administration is provided in PFDD Guidance 4.

[38] It is important to distinguish between a composite indicator measurement model and a composite endpoint. The *composite endpoint* is defined as the occurrence or realization in a patient of any one of the specified components. For more discussion of composite endpoints, see the guidance for industry *Multiple Endpoints in Clinical Trials* (October 2022).

3. *Approach to Missing Item or Task Responses in Scoring the COA*[39]

The scoring algorithm should explicitly state the conditions under which a score can still be computed in the presence of missing item/task responses, e.g., specifying the minimum number of item/task responses to compute a score and/or how missing items are to be identified and scored. A copy of the scoring manual should be provided to FDA so that reviewers can verify and replicate the sponsor's proposals according to the published scoring rules. Any rules for handling missing item/task responses should be justified sufficiently (e.g., through a missing data simulation study). Some considerations for justifying the approach to missing item/task responses include the following:

- The reasons why item/task responses might be missing can have implications for how missing responses are handled (i.e., are they expected to be Missing at Random, Missing Completely at Random, or Missing Not at Random; see Cappelleri et al. 2014). Sponsors are encouraged to collect reasons for missing responses to inform the choice of method to handle missing responses.

- The measurement framework (Classical Test Theory, Item Response Theory, Rasch Measurement Theory).

- For scoring based on adding or averaging over responses, simple imputation using the mean of the respondent's observed item/task responses might produce biased results if there is a hierarchy of difficulty or severity among the items (Cappelleri et al. 2014; Fayers and Machin 2016). For example, consider a PRO designed to assess mobility that asked patients to rate how much difficulty they have doing increasingly challenging physical activities (e.g., getting up from a chair, walking across their bedroom, walking one block, walking several blocks, jogging several blocks). If a patient rated themself as having little difficulty doing the less challenging activities and was missing responses to the more challenging activities, it would be inappropriate to impute the missing responses with the patient's mean response from the less challenging items (e.g., little difficulty walking across the bedroom does not automatically imply little difficulty jogging several blocks).

F. Scores From the COA Are Not Overly Influenced by Processes/Concepts That Are Not Part of the Concept of Interest

In a well-designed measure, it is the concept of interest that predominantly affects a respondent's responses to items/tasks. Thus, sponsors or measure developers should consider the most likely interfering influences on responses to items/tasks and assess the presence and strength of those influences. FDA recognizes that factors besides the concept of interest might influence scores (e.g., patients' fatigue from trying to complete overly burdensome measures). The important question is whether those other influences are so strong that they obscure the measurement of the

[39] This discussion is limited to methods for scoring the COA in the presence of missing items/tasks. PFDD Guidance 4 discusses other types of missingness affecting the construction and analysis of COA-based endpoints (e.g., missing entire forms).

concept of interest and increase uncertainty about the interpretation of the scores in the context of the trial for the target population.

Component F is stated in a generic way in Table 2 and in the title of this section. Sponsors should tailor the description of component F to correspond to the most likely sources of interfering influence. Sponsors should consider each step involved in generating a score on the COA to identify additional factors not listed here that may influence score interpretation. What follows are considerations for some of the most frequently encountered sources of interfering influence.

1. *Item or Task Interpretation or Relevance Does Not Differ Substantially According to Respondents' Demographic Characteristics (Including Sex, Age, and Education Level) or Cultural/Linguistic Backgrounds.*

Sponsors and measure developers should consider whether there are demographic groups for whom items might be interpreted differently or tasks might have different relevance for measuring the concept of interest. Some differences can be avoided by including all relevant demographic groups in the qualitative and quantitative research conducted to develop the COA and, in the case of cultural adaptation or linguistic translation, using best practices for adaptation/translation (Eremenco et al. 2018; McKown et al. 2020; Wild et al. 2005).

If there are remaining concerns about differences in interpretation/relevance across demographic groups, further qualitative research (e.g., using cognitive interviews) can be done to explore potential differences. If such qualitative or other data indicate likely differences between groups in the interpretation or relevance of items/tasks, sponsors can consider conducting quantitative analyses (e.g., measurement invariance and/or tests of differential item functioning; Cheung et al. 2002; Holland 1993; Leitgöb 2022; Millsap 2012; Teresi et al. 2021) to clarify the presence and magnitude of any group differences. Results from such analyses could inform judgments about whether there is a concerning amount of difference between groups that makes it difficult to interpret the COA scores as measures of the concept of interest.

2. *Recollection Errors Do Not Overly Influence Assessment of the Concept of Interest. [PROs, ObsROs, and ClinROs]*

For COAs that involve a recall period (e.g., past 24 hours, past 7 days), sponsors should consider the appropriateness of the recall period to be used. FDA recommends a clearly specified recall period to help standardize reporting. The selected recall period should be short enough to minimize the measurement error and/or potential bias (i.e., systematic inflation or deflation of scores) due to recall error, while also minimizing respondent burden. The recall period should be shown to be suitable for the intended context of use (e.g., trial duration or frequency and saliency of events). For some concepts of interest and contexts of use, there is already sufficient published data and regulatory experience that a given recall period will not overly influence scores. For this reason, sponsors should discuss the planned recall period with FDA to understand the type and amount of evidence needed to support the use of the recall period.

Potential sources of evidence could include qualitative and/or quantitative studies. Qualitive studies with patients might provide information about patients' understanding of the recall period used; patients' explanations of how they arrived at their answers (and specifically what period they were thinking about or what strategy they used to recollect); or patients' opinions about how far in the past they can recall with accuracy.[40] Quantitative studies might provide information about the correspondence between measurements based on longer (e.g., 14-day) and shorter (e.g., 7-day) recall periods.

3. Respondent Fatigue or Burden Does Not Overly Influence Assessment of the Concept of Interest

Consider whether COAs may induce respondent fatigue and burden due to measure length or complexity. Respondents who feel fatigued or over-burdened during an assessment might not provide data reflective of the underlying disease or the impact of treatment. Evidence from cognitive interviews and/or usability testing may provide insight as to whether a COA might lead to fatigue and/or burden. Patient experience of burden might also be addressed by improving patients' motivation through explaining the reasons for and importance of any lengthy and/or complex assessments.

4. Additional Influences on Score Due to Implementation and/or Study Design

Even after a COA has been determined to be fit-for-purpose, sponsors should consider whether later implementation and/or study design decisions might influence scores' ability to reflect the concept of interest. Issues that might arise include the following:

- **Fatigue or burden** when a COA is included as part of a long battery of measures and/or is assessed very frequently. Approaches to minimizing patient fatigue and burden are discussed in PFDD Guidance 4.

- **Expectation bias**. Responses to a COA may be overly influenced by the respondent's (i.e., patient's, caregiver's, or clinician's) or administrator's (for PerfOs) expectations of how well the patient should be doing. Such expectations could be based on the patient's assignment to an experimental group in an unmasked trial and/or the duration the patient has been in the clinical trial (e.g., earlier versus later study visits), their sex, or their age. Minimizing the influence of biases, including expectation bias, is very important and can be done, for example, by conducting randomized, controlled, and double-masked trials. When masked trials are not feasible, sponsors should evaluate the potential for influence on the COA by expectation bias and provide evidence that it will not unduly affect the study results.

[40] Qualitative studies might also solicit patients' and/or caregivers' perspectives on the length of time that should be sampled to obtain a representative assessment of some concept of interest. For example, patient feedback might suggest that characterizing a patient's current level of sexual functioning should be based on the patient's experiences over the past 30 days. Whether these 30 days are assessed using a single COA with a 30-day recall or by summarizing multiple administrations of COA with a shorter recall (e.g., average of 4 assessments of a COA with 7-day recall) would depend on how accurately patients can recall their sexual functioning over different lengths of time.

- **Practice effects with PerfOs.** It is possible that patients' performance on the tasks of a PerfO could improve over time due to practice rather than to real improvements in the concept of interest. Strategies to minimize the influence of practice effects on trial results are discussed briefly in Appendix D and at greater length in PFDD Guidance 4.

G. Scores From the COA Are Not Overly Influenced by Measurement Error

Ultimately, a COA-based endpoint needs to be reliable enough to detect treatment effects of interest—an issue discussed in PFDD Guidance 4. But for PFDD Guidance 3, reliability is relevant because low reliability might make it difficult to interpret the COA scores as measures of the concept of interest. There will always be some amount of measurement error in scores. This component of the rationale says that measurement error does not dominate the scores. In other words, variation in the scores can still be interpreted as reflecting variation in the concept of interest, even though there may also be some variation in scores due to random measurement error between assessment occasions, different raters, different sets of items/tasks, etc. This can be evaluated by collecting data on the reliability of scores (Mokkink et al. 2023). When evaluating reliability, sponsors and measure developers should consider and evaluate the most likely sources of variability in scores within the context of use (see Table 3).

Table 3 Possible Sources of Variation Associated with Different Types of COAs

		Potential Relevance for COA Type			
Source of Variation	**Type of Evidence**	***PRO***	***ObsRO***	***ClinRO***	***PerfO***
. . . over time within clinically stable patients	Test-retest reliability	X	X[a]	X[a]	X
. . . across different raters	Inter-rater reliability		X[b]	X	X[c]
. . . within the same rater for the same patients (when the patients have not clinically changed)	Intra-rater reliability		X	X	X[c]
. . . across different but highly related or similar tasks for the same patients	Evaluation of score differences between related tasks or sets of tasks				X
. . . across different but highly related or similar items for the same patients	Evaluation of score differences between related items or sets of items	X	X	X	

[a]For ObsROs and ClinROs, variations over time in clinically stable patients might reflect some combination of variations across different raters and variations within the same rater over time.
[b]Might be relevant if two different caregivers (e.g., parent and preschool teacher) provide ratings on the same patient.
[c]Applies only if the PerfO requires a trained rater as part of the assessment process.

Note that test-retest reliability evidence is only relevant for diseases or conditions in which a patient's health status can remain stable for some period of time (e.g., 1 to 2 weeks). In situations in which it is challenging to identify patients who remain stable, sponsors should consider whether there might be alternative ways to demonstrate the reliability of the scores. Test-retest reliability should be evaluated in the absence of any systematic intervening effects other than natural variability among patients. Sponsors should specify one or more criteria to define stable patients. The interval between the test and retest should be long enough so that respondents are unlikely to recall their initial responses, but short enough that the patients' health status is stable over the interval. FDA recommends that, in most cases, intraclass correlation coefficients be calculated to estimate test-retest reliability (McGraw and Wong 1996; Qin et al. 2019). The sponsor should provide justification for the selected method.

For measures developed using Item Response Theory (IRT) modeling, an additional estimate of reliability can be generated based on the information function. The associated standard errors can provide another method of examining the variability and consistency of scores (Embretson and Reise 2000). Measures of internal consistency reliability (e.g., Cronbach's alpha) might also provide additional information. However, these estimates of reliability (IRT-based and internal consistency) alone may not be sufficient to support this component of the rationale.

During the development process of a COA, evidence of good reliability might be obtained earlier in the process (e.g., using a cross-sectional study design). This evidence, along with other supporting material, might be enough to justify the exploratory use of the COA in prospective trials (e.g., phase 2).

H. Scores From the COA Correspond to the Meaningful Aspect of Health Related to the Concept of Interest

As discussed in section II.B, we learn about the effect of a medical product on a MAH by estimating a treatment effect on a COA-based endpoint thought to reflect the MAH. Ultimately, an important part of the rationale for a COA-based endpoint (addressed in PFDD Guidance 4) is that the scores of the endpoint are related to the MAH. Typically, the scores from the COA make up a large part of a COA-based endpoint. Therefore, it is important that the scores from the COA correspond to the MAH related to the concept of interest that the COA measures.[41]

As noted in section II.B, some measures assess a concept of interest that directly reflects the MAH, such as patient-reported nausea intensity or pain intensity. For such measures, there might be little uncertainty that the scores correspond to the patient's experience. However, other measures might assess a concept of interest that is indirectly related to the MAH that the medical product is targeting.

[41] In cases where scores of the COA-based endpoint may be very different from the scores on the COA, it will be more important to show that the endpoint scores correspond to the MAH. For example, a MAH might be assessed using a multi-component endpoint constructed using scores from multiple COAs. In that case, the scores from any one COA may or may not correspond well to the MAH; but the multi-component endpoint scores should correspond well with the MAH.

Consider again the example of lower limb-related functioning (Walton et al. 2015). A PRO might be used to assess this MAH in a relatively direct way by asking the patient about the ease with which they have done a range of activities that require lower limb-related function. Although measurement error will influence scores on the PRO, it is generally thought that those scores are directly related to the lower limb-related activities in the lived experience of patients. However, if there was significant heterogeneity among patients' physical environments and/or wide heterogeneity in the lower limb-related activities that patients undertake, a sponsor might decide instead to assess patients in a standardized environment via a PerfO. Under standardized conditions, one is no longer directly assessing lower limb-related function outside the test environment. Instead, the concepts of interest being assessed are important subcomponents of lower limb-related function that are amenable to standardized assessment, but neither subcomponent is sufficient alone to support an inference about the patient's overall lower limb-related function (see Figure 3). Because of this, these concepts of interest could be considered indirect reflections of patients' lower limb-related functioning in their daily lives. In this example, the sponsor might decide to measure in the test environment walking capacity, leg muscle strength, and walking speed, which all contribute to patients' lower limb-related functioning in their daily lives (reflected by the dotted line in the conceptual framework shown in Figure 3). But in the rationale for the use of each measure, it would still be important to evaluate how well scores are related to the patients' lower limb-related activities in the lived experience of patients outside of the clinical trial context.

For measures such as these in which the relationship between the scores and the MAH is less direct, more uncertainty exists. Thus, sponsors and measure developers might seek additional evidence by investigating the relationship between scores on the COA and other variables that are expected to be more directly related to the patient's experience. This is known as convergent evidence (American Educational Research Association et al. 2014).[42] The other variables could include alternative measures of, or be related to, the concept of interest using different methods and/or sources (e.g., observer report or performance tests). For example, the sponsor might assess patient-reported lower limb-related functioning in daily life along with measures of walking capacity, leg muscle strength, and walking speed in a phase 2 trial or early feasibility study. The sponsor might predict a moderate correlation[43] between the PRO's scores and scores on each of the three performance measures and could test this expectation using this data.

FDA reminds sponsors that when prespecifying correlation coefficient cutoffs for the psychometric statistical analysis plan (SAP), it is important to take into consideration the *a priori* hypothesized relationship(s) among the concept(s) measured by any proposed reference measure(s) and the concept(s) measured by the proposed COA. When interpreting correlation coefficients, sponsors should consider the size of the corresponding coefficient of determination and how the magnitude of the correlation might be attenuated by the distribution of the variables (i.e., restriction of range) and/or unreliability of the variables.

[42] *Convergent evidence* was previously referred to as *convergent validity*. See Footnote 33.

[43] It would be reasonable in this example for a sponsor to expect a moderate, but not large, correlation in this case. In the example, the sponsor chose PerfOs rather than PROs out of concern for the heterogeneity in the patients' environments. That environmental heterogeneity is expected to reduce the magnitude of the relationship between patient-reported and performance tested assessments of lower limb mobility.

Sponsors and measure developers might also conduct empirical comparisons of scores for patient groups known to differ with respect to the MAH (i.e., known-groups evidence[44]). When a sponsor is collecting known-groups evidence, FDA does not recommend dividing COA scores into groups based on the distribution(s) of reference measure scores (e.g., tertiles, quartiles, medians, or quintiles), because the percentile cutoff values are arbitrary and may vary across samples. Additionally, patient groups created based on the distribution of reference measure scores may not represent clinically distinct groups. Sponsors should propose appropriate reference measures and justify the corresponding cutoff values that represent distinct levels of symptom severity and/or impairment. In addition, sponsors should provide details of the proposed model and the hypothesis tests that will be performed.

[44] The extent to which scores differed between groups known to differ on the concept of interest was previously referred to as *known groups validity*. . See Footnote 33.

REFERENCES

Please note that the citation of a scientific reference in this guidance does not constitute FDA's endorsement of approaches or methods presented in that reference for any particular study. Study designs are evaluated on a case by case basis under applicable legal standards.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. American Educational Research Association; 2014.

Arbuckle R, Abetz-Webb L. "Not Just Little Adults": Qualitative Methods to Support the Development of Pediatric Patient-Reported Outcomes. *Patient*. Sep 2013;6(3):143-159. doi:10.1007/s40271-013-0022-3

BEST (Biomarkers, EndpointS, and other Tools) Resource. 2016. https://www.ncbi.nlm.nih.gov/books/NBK338448/

Bevans KB, Riley AW, Moon J, Forrest CB. Conceptual and methodological advances in child-reported outcomes measurement. *Expert Rev Pharm Out*. Aug 2010;10(4):385-396. doi:10.1586/Erp.10.52

Bollen KA, Bauldry S. Three Cs in Measurement Models: Causal Indicators, Composite Indicators, and Covariates. *Psychol Methods*. Sep 2011;16(3):265-284. doi:10.1037/a0024448

Bowling A. Mode of questionnaire administration can have serious effects on data quality. *J Public Health*. Sep 2005;27(3):281-291. doi:10.1093/pubmed/fdi031

Braekman E, Charafeddine R, Demarest S, et al. Comparing web-based versus face-to-face and paper-and-pencil questionnaire data collected through two Belgian health surveys. *Int J Public Health*. Jan 2020;65(1):5-16. doi:10.1007/s00038-019-01327-9

Byrom B, Gwaltney C, Slagle A, Gnanasakthy A, Muehlhausen W. Measurement Equivalence of Patient-Reported Outcome Measures Migrated to Electronic Formats: A Review of Evidence and Recommendations for Clinical Trials and Bring Your Own Device. *Therapeutic Innovation & Regulatory Science*. Jul 2019;53(4):426-430. doi:10.1177/2168479018793369

Cappelleri JC, Zou KH, Bushmakin AG, Alvir JMJ, Alemayehu D, Symonds T. *Patient-Reported Outcomes: Measurement, Implementation and Interpretation*. Chapman & Hall/CRC Biostatistics series. CRC Press, Taylor & Francis Group; 2014.

Cheung GW, Rensvold RB. Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Struct Equ Modeling*. 2002;9(2):233-255. doi:10.1207/S15328007sem0902_5

Critical Path Institute ePRO Consortium. Best Practices for Electronic Implementation of Response Scales for Patient-Reported Outcomes. 2018a. https://c-path.org/wp-content/uploads/BestPractices2_Response_Scales.pdf

Critical Path Institute ePRO Consortium. Best Practices for Maximizing Electronic Data Capture Options during the Development of New Patient-Reported Outcome Measures. 2018b. https://c-path.org/wp-content/uploads/BestPractices_Maximizing_Data_Capture.pdf

de Vet HCW, Terwee CB, Mokkink LB, Knol DL. *Measurement in Medicine: A Practical Guide*. 1st ed. Practical Guides to Biostatistics and Epidemiology. Cambridge University Press; 2011.

Dworkin RH, Turk DC, Farrar JT, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. *Pain*. Jan 2005;113(1-2):9-19. doi:10.1016/j.pain.2004.09.012

Edgar CJ, Bush EN, Adams HR, et al. Recommendations on the Selection, Development, and Modification of Performance Outcome Assessments: A Good Practices Report of an ISPOR Task Force. *Value Health*. Jul 2023;26(7):959-967. doi:10.1016/j.jval.2023.05.003

Edwards MC, Slagle A, Rubright JD, Wirth RJ. Fit for purpose and modern validity theory in clinical outcomes assessment. *Qual Life Res*. 2017;27(7):1711-1720. doi:10.1007/s11136-017-1644-z

Embretson SE, Reise SP. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates; 2000.

Eremenco S, Coons SJ, Paty J, et al. PRO Data Collection in Clinical Trials Using Mixed Modes: Report of the ISPOR PRO Mixed Modes Good Research Practices Task Force. *Value Health*. Jul 2014;17(5):501-516. doi:10.1016/j.jval.2014.06.005

Eremenco S, Pease S, Mann S, Berry P. Patient-Reported Outcome (PRO) Consortium translation process: consensus development of updated best practices. *Journal of Patient-Reported Outcomes*. 2018;2(1):12-12. doi:10.1186/s41687-018-0037-6

Fayers PM, Machin D. *Quality of Life*. 3rd ed. Wiley Blackwell; 2016.

Hawker GA, Mian S, Kendzerska T, French M. Measures of adult pain: Visual Analog Scale for Pain (VAS Pain), Numeric Rating Scale for Pain (NRS Pain), McGill Pain Questionnaire (MPQ), Short-Form McGill Pain Questionnaire (SF-MPQ), Chronic Pain Grade Scale (CPGS), Short Form-36 Bodily Pain Scale (SF-36 BPS), and Measure of Intermittent and Constant Osteoarthritis Pain (ICOAP). *Arthritis Care & Research (Hoboken)*. Nov 2011;63 Suppl 11:S240-52. doi:10.1002/acr.20543

Holland PW, Wainer H. *Differential Item Functioning*. Lawrence Erlbaum Associates; 1993.

Houts CR, Bush EN, Edwards MC, Wirth RJ. Using validity theory and psychometrics to evaluate and support expanded uses of existing scales. *Qual Life Res*. Oct 2022;31(10):2969-2975. doi:10.1007/s11136-022-03162-7

Kane MT. Validating the Interpretations and Uses of Test Scores. *J Educ Meas*. 2013;50(1):1-73. doi:10.1111/jedm.12000

Leitgöb H, Seddig D, Asparouhov T, et al. Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. *Social Science Research*. 2022;110. doi:10.1016/j.ssresearch.2022.102805

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*. Jun 2013;16(4):461-479. doi:10.1016/j.jval.2013.04.004

McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods*. Mar 1996;1(1):30-46. doi:Doi 10.1037/1082-989x.1.1.30

McKown S, Acquadro C, Anfray C, et al. Good practices for the translation, cultural adaptation, and linguistic validation of clinician-reported outcome, observer-reported outcome, and performance outcome measures. *J Patient Rep Outcomes*. Nov 2020;4(1):89. doi:10.1186/s41687-020-00248-z

Millsap RE. *Statistical Approaches to Measurement Invariance*. Routledge; 2012.

Mokkink LB, Eekhout I, Boers M, van der Vleuten CPM, de Vet HCW. Studies on Reliability and Measurement Error of Measurements in Medicine - From Design to Statistics Explained for Medical Researchers. *Patient Relat Outcome Meas*. 2023;14:193-212. doi:10.2147/PROM.S398886

Mowlem FD, Elash CA, Dumais KM, et al. Best Practices for the Electronic Implementation and Migration of Patient-Reported Outcome Measures. Value Health. Jan 2024;27(1):79-94. doi:10.1016/j.jval.2023.10.007

O'Donohoe P, Reasner DS, Kovacs SM, et al. Updated Recommendations on Evidence Needed to Support Measurement Comparability Among Modes of Data Collection for Patient-Reported Outcome Measures: A Good Practices Report of an ISPOR Task Force. *Value Health*. 2023;26(5):623-633. doi:10.1016/j.jval.2023.01.001

Papadopoulos EJ, Patrick DL, Tassinari MS, et al. Clinical Outcome Assessments for Clinical Trials in Children. *Pediatric Drug Development: Concepts and Applications*. 2nd ed. John Wiley & Sons Ltd; 2013:545:chap 42.

Papadopoulos EJ, Bush EN, Eremenco S, Coons SJ. Why Reinvent the Wheel? Use or Modification of Existing Clinical Outcome Assessment Tools in Medical Product Development. *Value Health*. Oct 2019;23(2):151-153. doi:10.1016/j.jval.2019.09.2745

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*. Dec 2011a;14(8):967-977. doi:10.1016/j.jval.2011.06.014

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*. Dec 2011b;14(8):978-988. doi:10.1016/j.jval.2011.06.013

Petrillo J, Cano SJ, McLeod LD, Coon CD. Using Classical Test Theory, Item Response Theory, and Rasch Measurement Theory to Evaluate Patient-Reported Outcome Measures: A Comparison of Worked Examples. *Value Health*. Jan 2015;18(1):25-34. doi:10.1016/j.jval.2014.10.005

Powers III JH, Patrick DL, Walton MK, et al. Clinician-Reported Outcome Assessments of Treatment Benefit: Report of the ISPOR Clinical Outcome Assessment Emerging Good Practices ask Force. *Value Health*. 2017;20(1):2-14.

Qin SS, Nelson L, McLeod L, Eremenco S, Coons SJ. Assessing test-retest reliability of patient-reported outcome measures using intraclass correlation coefficients: recommendations for selecting and documenting the analytical formula. *Quality of Life Research*. Apr 2019;28(4):1029-1033. doi:10.1007/s11136-018-2076-0

Romero H, DeBonis D, O'Donohoe P, et al. Recommendations for the Electronic Migration and Implementation of Clinician-Reported Outcome Assessments in Clinical Trials. *Value Health*. Jul 2022;25(7):1090-1098. doi:10.1016/j.jval.2022.02.012

Rothman MP, Burke LRMPH, Erickson PP, Leidy NKP, Patrick DLPM, Petrie CDP. Use of Existing Patient-Reported Outcome (PRO) Instruments and Their Modification: The ISPOR Good Research Practices for Evaluating and Documenting Content Validity for the Use of Existing Instruments and Their Modification PRO Task Force Report. *Value Health*. 2009;12(8):1075-1083. doi:10.1111/j.1524-4733.2009.00603.x

Streiner DL, Norman GR, Cairney J. *Health Measurement Scales: A practical guide to their development and use*. 5th ed. Oxford University Press; 2015.

Teresi JA, Wang C, Kleinman M, Jones RN, Weiss DJ. Differential Item Functioning Analyses of the Patient-Reported Outcomes Measurement Information System (PROMIS®)

Measures: Methods, Challenges, Advances, and Future Directions. *Psychometrika*. Sep 2021;86(3):674-711. doi:10.1007/s11336-021-09775-0

U.S. Food and Drug Administration (FDA). *Applying Human Factors and Usability Engineering to Medical Devices: Guidance for Industry and FDA Staff*. 2016. FDA-2011-D-0469. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/applying-human-factors-and-usability-engineering-medical-devices

U.S. Food and Drug Administration (FDA). Clinical Outcome Assessment (COA) Qualification Program. https://www.fda.gov/drugs/drug-development-tool-ddt-qualification-programs/clinical-outcome-assessment-coa-qualification-program

U.S. Food and Drug Administration (FDA). *Digital Health Technologies for Remote Data Acquisition in Clinical Investigations: Guidance for Industry, Investigators, and Other Stakeholders*. 2023. FDA-2021-D-1128. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/digital-health-technologies-remote-data-acquisition-clinical-investigations

U.S. Food and Drug Administration (FDA). Medical Device Development Tools (MDDT). https://www.fda.gov/medical-devices/medical-device-development-tools-mddt?utm_medium=email&utm_source=govdelivery

U.S. Food and Drug Administration (FDA). *Multiple Endpoints in Clinical Trials: Guidance for Industry*. 2022. FDA-2016-D-4460. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials-guidance-industry

U.S. Food and Drug Administration (FDA). *Patient-Focused Drug Development: Collecting Comprehensive and Representative Input: Guidance for Industry, FDA Staff, and Other Stakeholders.* 2020. FDA-2018-D-1893. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-focused-drug-development-collecting-comprehensive-and-representative-input

U.S. Food and Drug Administration (FDA). *Patient-Focused Drug Development: Incorporating Clinical Outcome Assessments Into Endpoints for Regulatory Decision-Making: Draft Guidance for Industry, FDA Staff, and Other Stakeholders*. 2023. FDA-2023-D-0026. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-focused-drug-development-incorporating-clinical-outcome-assessments-endpoints-regulatory

U.S. Food and Drug Administration (FDA). *Patient-Focused Drug Development: Methods to Identify What Is Important to Patients: Guidance for Industry, FDA Staff, and Other Stakeholders*. 2022. FDA-2019-D-4247. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-focused-drug-development-methods-identify-what-important-patients

U.S. Food and Drug Administration (FDA). *Patient Preference Information - Voluntary Submission, Review in Premarket Approval Applications, Humanitarian Device Exemption Applications, and De Novo Requests, and Inclusion in Decision Summaries and Device Labeling: Guidance for Industry, FDA Staff, and Other Stakeholders*. 2016. FDA-2015-D-1580. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-preference-information-voluntary-submission-review-premarket-approval-applications

U.S. Food and Drug Administration (FDA). *Principles for Selecting, Developing, Modifying, and Adapting Patient-Reported Outcome Instruments for Use in Medical Device Evaluation: Guidance for Industry, FDA Staff, and Other Stakeholders*. 2022. FDA-2020-D-1564. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/principles-selecting-developing-modifying-and-adapting-patient-reported-outcome-instruments-use

Wainer H. *Computerized Adaptive Testing: A Primer*. 2nd ed. Lawrence Erlbaum Associates; 2000.

Walton MK, Powers JH, Hobart J, et al. Clinical Outcome Assessments: Conceptual Foundation-Report of the ISPOR Clinical Outcomes Assessment - Emerging Good Practices for Outcomes Research Task Force. *Value Health*. Sep 2015;18(6):741-752. doi:10.1016/j.jval.2015.08.006

Weinfurt KP. Constructing arguments for the interpretation and use of patient-reported outcome measures in research: an application of modern validity theory. *Quality of Life Research*. Jun 2021;30(6):1715-1722. doi:10.1007/s11136-021-02776-7

Weinfurt KP. Constructing and evaluating a validity argument for a performance outcome measure for clinical trials: An example using the Multi-luminance Mobility Test. Clin Trials. Apr 2022;19(2):184-193. doi:10.1177/17407745211073609

Weinfurt KP. A Commentary on the ISPOR Task Force's Report on Developing, Selecting, and Modifying Performance Outcome Assessments. *Value Health*. 2023;27(7):957-958. doi:10.1016/j.jval.2023.04.011

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*. Mar-Apr 2005;8(2):94-104. doi:10.1111/j.1524-4733.2005.04054.x

Willis GB. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. SAGE Publications; 2005.

Willis GB. *Analysis of the Cognitive Interview in Questionnaire Design*. Oxford University Press; 2015.

APPENDIX A: PATIENT-REPORTED OUTCOME MEASURES

I. INTRODUCTION

A PRO is a type of COA that is a self-assessment of a patient's health condition directly from the patient without interpretation of the patient's response by others. A PRO is likely to be the best COA type to assess a concept of interest when the concept of interest is any of the following and the patient is able to reliably self-report:

- A feeling or experience known only to the patient, such as pain, itch, shortness of breath as no one else has direct access to feelings except for the patient
- Any type of functioning or activity that is part of the patients' day-to-day life
- Degree of impact on day-to-day life associated with one or more symptoms

Note that a PRO is intended to be completed by the patient themselves and cannot be completed by a proxy reporter, i.e., someone reporting on behalf of the patient (see Appendix B ObsRO and Appendix C ClinRO for further discussion). This does not exclude a third party from aiding (e.g., disability accommodations specified by study protocol) a patient so the patient can make their own report, e.g., an aide marking down the responses of a patient with fine motor impairments. The aide would not be considered a proxy reporter because the report is coming from the patient without any interpretation by the aide. However, the use of an aide or an administrator to facilitate PRO completion by the patient may potentially impact patient responses (e.g., Bowling 2005; Braekman et al. 2020) and requires careful consideration and a standardized process for implementation.

II. HYPOTHETICAL EXAMPLE

Imagine that a sponsor wishes to assess the effects of a new medical product on MAHs associated with Disease W. Following the Roadmap described in section III, the sponsor conducts research that includes qualitative studies of patients with Disease W (using methods described in PFDD Guidances 1 and 2) to understand all the important aspects of health that are impacted by Disease W, i.e., the MAHs. Then, with input from patients and clinical experts, the sponsor selects from among the different MAHs three MAHs that have the potential to be affected by the new medical product within the time frame of a clinical trial—Function A, Symptom A intensity, and Symptom B intensity.

Table A displays the COA-based endpoint approach the sponsor proposes for these three MAHs. The sponsor proposes that the first MAH, Function A, can be measured directly as a concept of interest using a 5-item PRO known as the Function A Assessment Scale, which uses a 7-day recall period. The sponsor plans to interpret the Total Score from this measure as reflecting the concept of interest, Function A. The sponsor includes an evidence-based rationale to support this interpretation (see section IV).

Table A. Example Description of a COA-Based Endpoint Approach using Patient-Reported Outcomes

		Clinical Outcome Assessment			
Meaningful Aspect of Health	**Concept of Interest for Measurement**	*Type*	*Name*	*Score*	**COA-based Endpoint**
Function A	Function A	PRO	Function A Assessment Scale	Function A Total Score	Function A Total Score at 12 weeks post-randomization
Symptom A Intensity	Worst Symptom A Intensity in past 24 hours	PRO	Disease W Daily Diary	Symptom A Score*	Patient's worst Symptom A Score over 7 days assessed at 12 weeks post-randomization
Symptom B Intensity	Worst Symptom B Intensity in past 24 hours	PRO	Disease W Daily Diary	Symptom B Score*	Patient's worst Symptom B Score over 7 days assessed at 12 weeks post-randomization

*Higher scores reflect greater symptom intensity.

Note that one component of the rationale is that the method of scoring responses to the Function A Assessment Scale is appropriate for assessing Function A (see section IV.E). The scoring assumes a particular measurement model for combining responses across the multiple items—in this case, a reflective indicator model. To help communicate this to reviewers, the sponsor illustrates the measurement model in Figure A.

Figure A. Measurement Model for the Function Assessment Scale

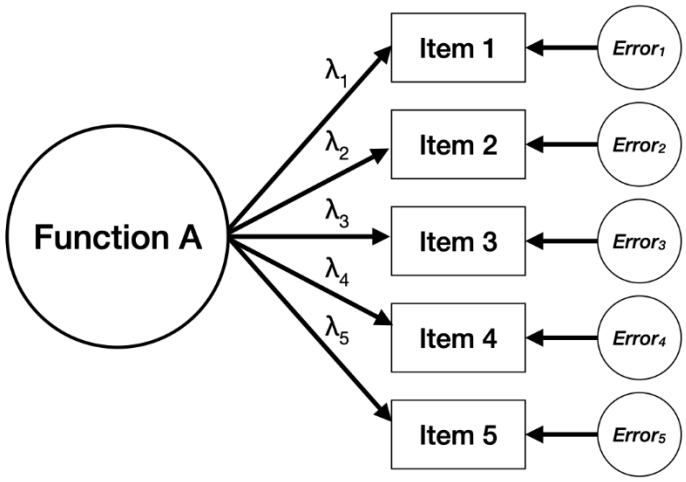


Table A also shows that two other MAHs will be targeted in the proposed clinical trial—Symptom A intensity and Symptom B intensity. For both, the sponsor has decided to assess the worst intensity during a 24-hour period (the concept of interest) using a multi-item PRO known as the Disease W Daily Diary. The daily diary includes single items measuring the worst

intensity over the past 24 hours for both Symptom A and B. The ordinal responses *None*, *Mild*, *Moderate*, and *Severe* are assigned scores of 0, 1, 2, or 3. For each score (Symptom A and Symptom B), the sponsor includes an evidence-based rationale (see section IV) to support the proposed interpretation. Note that because the scoring is based on single-item measures, a measurement model similar to the multi-item Function A Assessment Scale earlier may not be needed. Table A also shows that an endpoint will be constructed for each symptom by taking the maximum score obtained over 7 daily assessments of the Disease W Daily Diary PRO (with a prespecified approach to handling missing data), taken at 12 weeks post-randomization (see Guidance 4 for a discussion of COA-based endpoints).

Note that this example presents the end result of a process described by the Roadmap (see section III) and which should be based on regular communication with FDA.

APPENDIX B: OBSERVER-REPORTED OUTCOME MEASURES

I. INTRODUCTION

An ObsRO is a type of COA that assesses observable signs, events, or behaviors related to a patient's health condition and is reported by someone other than the patient or a health professional (e.g., parent, caregiver, or someone who cares for the patient the most or spends significant time with the patient during the relevant observation window in daily life). An ObsRO does not rely on medical judgment or interpretation[45] and can be particularly useful for patients who cannot report for themselves.

Example ObsROs for Use in Clinical Trials
• Rating scales completed by a caregiver, such as: – Acute Otitis Media Severity of Symptoms scale, a measure used to assess signs and behaviors related to acute otitis media in infants • Counts of events recorded by a caregiver (e.g., observer-completed log of seizure episodes)

Observation versus proxy report

Observable signs, events, or behaviors do not require the reporter to interpret or infer on behalf of the patient. A proxy report is when someone other than the patient reports on behalf of the patient's experiences as if they are the patient. Concepts that are only known by the patient (e.g., symptoms) should be measured by a PRO using information that contemporaneously comes directly from the patient. If patients are unable to self-report, a different type of COA (e.g., an ObsRO) should be used in the study.

For example, itch intensity is something known only to the patient and should be assessed using a PRO. FDA strongly discourages a proxy report, whereby someone other than the patient tries to infer the intensity of the patient's itching, as the proxy report would be unreliable and generally inappropriate. However, someone other than the patient (i.e., a caregiver or other observer) could report on the patient's observable behaviors (i.e., scratching behaviors) using an ObsRO. FDA acknowledges there are instances when it is impossible to collect valid and reliable self-report data from the patient. In these instances, it is recommended an ObsRO be used rather than a proxy report.

[45] A measure that relies on medical judgment or interpretation is a ClinRO.

Examples of ObsRO Versus Proxy-Reported Item Stem Phrasing
ObsRO items
• "In the last hour, how often did you see your child holding their stomach or abdomen?"
• "How frequently did the patient do household chores (e.g., laundry, washing dishes) in the past week?"
• "Based on what you observed, how often did your child scratch themselves from the time your child woke up today until now?"
Proxy-reported outcome items
• "How severe was your child's pain from the time your child woke up until right now?"
• "Please rate your child's tiredness over the past 24 hours."
• "My child felt out of breath because of their asthma."
• "My child felt sad when they had pain."

ObsRO Selection and Implementation Considerations

- Consider predefining the minimum amount of regular contact between the observer and the patient that is required for an individual to be selected as an observer.
- When implementing an ObsRO in a clinical study, to the extent feasible, the same observer should complete the assessments throughout the trial to minimize unwanted variability due to different reporters. If another caregiver (e.g., a schoolteacher) will provide reports, the other caregiver should be sufficiently trained to provide responses to the ObsRO. In all cases, the reporter associated with an observation should be recorded.

II. HYPOTHETICAL EXAMPLE

Imagine that a sponsor wishes to assess the effects of a new medical product on MAHs associated with Disease X in pediatric patients ages 2 to 5. Following the Roadmap described in section III, the sponsor conducts research that includes qualitative studies of patients and their caregivers with Disease X (using methods described in PFDD Guidances 1 and 2) to understand all the important aspects of health that are impacted by Disease X, i.e., the MAHs. Then, with input from patients, caregivers, and clinical experts, the sponsor selects from among the different MAHs one that has the potential to be affected by the new medical product within the time frame of a clinical trial—Symptom A frequency. Based on results from earlier qualitative studies, it is determined that children in the target population are unable to consistently and accurately self-report Symptom A frequency. However, Symptom A causes various behaviors that can be observed by a parent or caregiver. Additional research (e.g., pilot study) indicates that measuring the frequency of observable Symptom A-related behaviors provides information about the child's Symptom A frequency (i.e., the MAH). The multi-item ObsRO, the Disease X Parent Inventory, includes the relevant observable Symptom A-related behaviors and is selected for use in the clinical trial. The sponsor plans to interpret the Symptom A Frequency Subscale score from this ObsRO as reflecting the concept of interest, the frequency of observable Symptom A-related behaviors. The sponsor includes an evidence-based rationale to support this interpretation

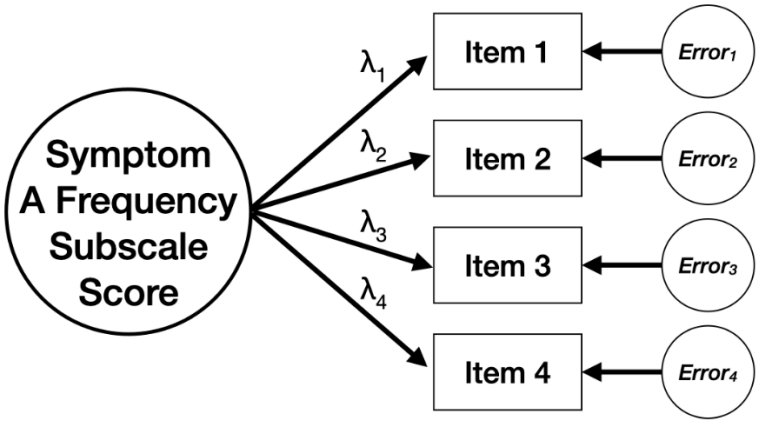
(see section IV). Table B displays the COA-based endpoint approach the sponsor proposes for the MAH in this trial.

Table B. Example Description of a COA-Based Endpoint Approach using an Observer-Reported Outcome

		Clinical Outcome Assessment			
Meaningful Aspect of Health	**Concept of Interest for Measurement**	*Type*	*Name*	*Score*	**COA-based Endpoint**
Symptom A Frequency	Frequency of Observable Symptom A-related Behaviors	ObsRO	Disease X Parent Inventory	Symptom A Frequency Subscale Score	Symptom A Frequency Subscale Score at 12 weeks post-randomization

One component of the rationale is that the method of scoring responses to the 4-item Symptom A Frequency Subscale of the Disease X Parent Inventory is appropriate for assessing the frequency of observable Symptom A-related behaviors (see section IV.E). The scoring assumes a particular measurement model for combining responses across the multiple items—in this case, a reflective indicator model. To help communicate this to reviewers, the sponsor illustrates the measurement model in Figure B.

Figure B. Measurement Model for the Disease X Parent Inventory, Symptom A Frequency Subscale



Note that this example presents the end result of a process described by the Roadmap (see section III) and which should be based on regular communication with FDA.

APPENDIX C: CLINICIAN-REPORTED OUTCOME MEASURES

I. INTRODUCTION

ClinROs are typically used when clinical judgment is needed to assess some aspect of a patient's health. ClinROs can include reports of clinical signs or events, ratings of a sign, and clinician's global assessments of the patient's current status or of the change the patient undergoes (Powers et al. 2017).

Examples: ClinROs
• Psoriasis Area and Severity Index, a measure used to assess the severity and extent of a patient's psoriasis • Clinician global assessment of psoriasis severity, such as through a single-item verbal rating scale

ClinRO Selection and Implementation Considerations

Below are key considerations and recommendations for selecting and implementing a ClinRO in a clinical study:

- When developing a new ClinRO, conduct cognitive interviews with clinicians to confirm the clinical relevance of the concepts, as well as to ensure that the measure (i.e., instructions and items) is understandable and that the response options reasonably capture the clinician assessment of the patient for each item.
- Include a user manual with clear instructions and directions for standardized administration.
- Implement standardized process for training, qualification, and certification of raters (where applicable) to help ensure that rating assessments are based on consistent criteria for the ratings to minimize unwanted variability. Ongoing refresher trainings should also be conducted as needed during the clinical trial.
- Scales should be developed and tested as they will be used in the registration trial (e.g., it is inappropriate to assume the measurement properties for a dermatology scale used to assess a patient's condition by photographs will be the same when the scale is used during an in-person (non-photographic) assessment).
- Implement a standardized case report form for data collection.
- Evaluate intra- and inter-rater reliability prior to using a proposed ClinRO in a registration trial.
- If visual aids (e.g., photo guides) are used, ensure that they cover a wide variety of patients (e.g., demographics, severity of disease/condition), and environmental characteristics and pilot test them with clinician raters to ensure they are well understood.
- For ClinROs used for primary endpoints, use an assessor who is masked from study group assignment and study visit, if feasible and appropriate in the context of use; in some cases, a centralized independent blinded review and an adjudication process in the event of rating discrepancies may be necessary to ensure consistent assessment.

- To the extent feasible, the same clinician should conduct the assessments for the same patients throughout the trial to minimize unwanted variability due to different reporters.
- Evaluate whether alternative methods for the ClinRO assessments (e.g., phone contact, virtual visit, alternative location for assessment) could be consistently implemented.

An overview of other issues relevant to the development and use of ClinROs is found in Powers et al. (2017).

II. HYPOTHETICAL EXAMPLE

Imagine that a sponsor wishes to assess the effects of a new medical product on MAHs associated with a dermatologic Disease Y. Following the Roadmap described in section III, the sponsor conducts research that includes qualitative studies of patients with Disease Y (using methods described in PFDD Guidances 1 and 2) to understand all the important aspects of health that are impacted by Disease Y, i.e., the MAHs. In the case of Disease Y, the MAHs include symptoms such as pain and itching and functional impacts in terms of daily activities of living. With input from patients and clinical experts, the sponsor chooses to use PROs to assess symptom severity and functioning more directly from patients. However, the concept of interest of severity of clinical signs of Disease Y, such as the severity of skin lesions (Powers et al. 2017), would provide additional information on the treatment benefit of the new medical product. The sponsor decides to use a 4-item ClinRO known as the Disease Y Severity Index to assess severity of clinical signs of Disease Y. The sponsor plans to interpret the total score from this measure as reflecting the concept of interest—severity of clinical signs of Disease Y. The sponsor includes an evidence-based rationale to support this interpretation (see section IV). An important part of this rationale that will need to be supported with evidence is that scores from the measure correspond with the patients' symptom severity (see section IV.H). Table C displays the ClinRO portion of the COA-based endpoint approach the sponsor proposes for the MAHs.

Table C. Example Description of a COA-Based Endpoint Approach using a Clinician-Reported Outcome

		Clinical Outcome Assessment			
Meaningful Aspect of Health	**Concept of Interest for Measurement**	*Type*	*Name*	*Score*	**COA-based Endpoint**
Symptom severity of Disease Y	Severity of Clinical Signs of Disease Y	ClinRO	Disease Y Severity Index	Total score	Total score at Week 16 post-randomization

Note that this example presents the end result of a process described by the Roadmap (see section III) and which should be based on regular communication with FDA.

APPENDIX D: PERFORMANCE OUTCOME MEASURES

I. INTRODUCTION

A PerfO is a type of COA that is used to generate patient experience data through standardized task(s) performed by a patient. A PerfO is administered and evaluated by an appropriately trained individual or independently completed. PerfOs are commonly used to assess patient physical or cognitive functioning, or perceptual/sensory functioning, through standardized tasks completed by the patient. The patient's performance on these tasks is then quantified and reported using predefined procedures.

A PerfO can be considered for use when patient functioning is the concept(s) of interest (e.g., mobility, memory, visual acuity) and the patient is able to follow the instructions to perform the required task(s).

Because PerfOs are based on patients' actual performance on a set of standardized tasks, they may be advantageous for the following reasons:

- When appropriately designed, PerfOs may reduce the influence of culture and language variability on outcome assessment in multinational and multilanguage trials.
- By having patients perform standardized tasks in a controlled, standardized environment, PerfOs may be less influenced by variability between and within patients in the types and settings of daily activities performed by the patients in their natural environment (e.g., driving a car versus taking public transportation, living in rural area versus living in big cities; Edgar et al. 2023).
- By assessing real-time functioning, PerfOs are not vulnerable to errors of recall that can occur for some PROs, ObsROs, and ClinROs that use a recall period (e.g., during the past 7 days).
- PerfOs may be less vulnerable to external changes in the patient's environment, such as seasonal impacts on daily routines (e.g., a patient might take fewer and shorter walks during a cold winter or a hot summer).
- Results of PerfOs may be communicated in units that are familiar and readily interpretable such as meters (e.g., distance walked in 6 minutes), seconds (e.g., time to climb a flight of stairs), and frequency counts (e.g., number of words recalled).

PerfO Selection and Implementation Considerations

Although using a PerfO can be beneficial in a clinical trial, the following are examples of unique challenges and recommendations:

- ***Potentially less direct relationship to a meaningful aspect of health.*** Each task usually assesses a specific function in a standardized environment. Therefore, the patient's performance on the standardized task(s) may provide only limited information about the patient's overall functioning outside of the assessment setting.
- ***Effects are potentially difficult to translate.*** Some PerfOs, and changes in their scores, are difficult to translate into noticeable, pragmatic, and meaningful changes in how patients function or feel in their daily lives. In many cases a supplementary PRO, ObsRO, and/or ClinRO should also be used to evaluate the meaningfulness.

- ***Potential interference of functions or abilities that are not part of the concept of interest*** (see section IV.F)**.** Some PerfO tasks require multiple functions to complete. For example, fine motor skills might be important in providing a response to a neuropsychological measure of memory functioning (Edgar et al. 2023), and so someone with fine motor impairment might receive a score that does not reflect the person's true memory functioning. Care should be taken to ensure that functions other than the concept of interest do not unduly influence scores on the PerfO. Pilot testing a PerfO prior to selecting it for use in a study is recommended. If the patient's cognitive ability may interfere with the performance of the tasks, sponsors should consider whether the selected PerfO is fit-for-purpose.
- ***Potential for patient fatigue or burden.*** Because a PerfO involves assessing how well and/or how quickly a patient performs a task, it is important to consider how patient fatigue or burden may impact their performance. This is especially the case when PerfOs are time- or effort-sensitive. Some causes of fatigue (e.g., traveling in busy traffic to the clinic visit) may be more challenging to prevent than others. When developing the clinical trial protocol, sponsors should consider the cumulative burden on the patient and the placement of the PerfO within the clinic visit day. For example, in a trial for a disease in which fatigue is a primary concern for patients, consider the impact of preceding the PerfO with other assessments that may impact fatigue.
- ***Standardization.*** If a specific published administrator's manual is selected for the test, it is important to conduct the test in accordance with the selected manual.
- ***Inaccessible equipment for task administration.*** Required equipment or assessment setup may not be available or feasible for certain clinical trial sites (e.g., a flight of stairs, air-conditioned rooms) or the materials may not be consistent across sites. Special attention should be paid to maintaining standardization of PerfOs, especially in multisite and multinational clinical trials, to ensure that the assessment results are reliable, valid, interpretable, and can be pooled for analysis.
- ***Practice effects.*** There are some instances in which patients improve their performance after repeated exposure to the same tasks, even though their underlying disease state has not changed. Steps should be implemented in trials to minimize the practice effect so that it does not influence the assessment results, including increasing the time in between PerfO assessments and allowing all patients to practice the task prior to randomization. Sponsors should consider potential practice effects associated with the selected performance-based tasks. The study protocol should include plans and/or procedures that will be put in place to minimize the influence of practice effects on the interpretation of the PerfO-based endpoint results.
- ***Standardized case report forms, assistive devices, and documentation.*** The use of a standardized case report form is recommended, which should include information on whether an assistive device was used during the test. The use of assistive devices should be standardized, and the type of device, if used, should be recorded. If the test was not completed, sponsors should collect the reason for not completing the test. These pieces of information should be part of the analysis data sets and may play a role in analysis and interpretation of the data.

Several of these potential challenges including practice effects, burden, and inaccessible equipment may differ among trial participants in ways that are not related to disease, making

score interpretation even more challenging. Data collection and careful context considerations in trial design and analyses are important to mitigating these trial interpretation challenges.

II. HYPOTHETICAL EXAMPLE

Imagine that a sponsor wishes to assess the effects of a new medical product on MAHs associated with Disease Z. Following the Roadmap described in section III, the sponsor conducts research that includes qualitative studies of patients with Disease Z (using methods described in PFDD Guidances 1 and 2) to understand all the important aspects of health that are impacted by Disease Z, i.e., the MAHs. Then, with input from patients and clinical experts, the sponsor selects from among the different MAHs one that has the potential to be affected by the new medical product within the time frame of a clinical trial—lower limb-related functioning. Because of the heterogeneity among patients in their activities and environments, the sponsor decides not to assess lower limb-related functioning directly as the concept of interest using a PRO-based primary endpoint. Instead, the sponsor, with input from clinical experts, identifies three subfunctions that are important to lower limb-related functioning: functional walking capacity, leg muscle strength, and walking speed. The sponsor proposes to measure each using a different PerfO. The sponsor includes an evidence-based rationale to support the proposed interpretation of the scores from each PerfO (see section IV). Table D displays the COA-based endpoint approach the sponsor proposes for this MAH.

Table D. Example Description of a COA-Based Endpoint Approach using a Reported Performance Outcome[46]

		Clinical Outcome Assessment			
Meaningful Aspect of Health	**Concept of Interest for Measurement**	*Type*	*Name*	*Score*	**COA-based Endpoint**
Lower-limb Related Functioning	Functional Walking Capacity	PerfO	6 Minute Walk Test	Distance walked (meters)	Distance walked at 6 months post-randomization
	Mobility and Balance	PerfO	Timed Up and Go	Time to complete (seconds)	Time at 6 months post-randomization
	Walking Speed	PerfO	Timed 25-foot Walk	Time to Complete (seconds)	Time at 6 months post-randomization

Note that this example presents the end result of a process described by the Roadmap (see section III) and which should be based on regular communication with FDA.

[46] Adapted from Figure 1in Edgar et al. (2023). This example is meant to illustrate how a sponsor might communicate their COA-based endpoint approach. It is not meant as an endorsement of the COAs or endpoints by FDA.

APPENDIX E: EXAMPLE TABLE FORMAT TO SUMMARIZE RATIONALE AND SUPPORT FOR A COA

Each rationale can be tailored to the proposed interpretation of COA scores. Each component should be accompanied by supporting evidence and justification (see section IV for a discussion on possible sources of evidence to evaluate each component).

As sponsors work to develop a comprehensive rationale table, keep this important question in mind: "How is this measure supposed to work and what evidence do we have that the measure works as intended?"

It is helpful to precede the rationale with a concise statement of the following:

- **Intended context of use**
- **Description and justification for the MAH:** Sponsors should report or cite data collected from patients and/or caregivers on the nature and importance of the MAH (see sections III.A and B)
- **Concept of interest:** If not identical to the MAH, then sponsors should explain and justify how measurement of the concept of interest helps us to understand how patients feel or function (i.e., the MAH). For example, in cases where the concept of interest is one of several concepts of interests that make up the MAH, the sponsor could explain how measurement of that concept of interest will be combined with other measured concepts of interest to support an inference about the MAH.

A table may be considered to summarize the rationale and its support (see Table E). Detailed support for each row of the summary table should be clearly described and included after the table. Sponsors can include in the table what part (e.g., section) of the application includes the supporting evidence.

Table E. Example Table Format to Summarize Rationale and Support for a COA to Measure a Concept of Interest in a Specific Context of Use

	Component[a]	Support[b]
A	**The concept of interest, [FILL IN], should be assessed by a [PRO/ObsRO/ClinRO/PerfO], because . . .**	
B	**[NAME OF MEASURE] includes all the important parts of [CONCEPT OF INTEREST].**	
C	**[NAME OF MEASURE] is administered as intended by the measure developer.**	
D	**[RESPONDENTS PROVIDING INFORMATION] understand the instructions and [ITEMS or TASKS] as intended by the measure developer.**	
E	**The method of scoring responses to the [NAME OF MEAURE] is appropriate for assessing [CONCEPT OF INTEREST].**	
E.1	Results of the assessment of measure model assumptions show the [ITEM or TASK] responses are consistent with the statistical assumptions of the proposed measurement model.	
E.2	Empirical assessment of missing data rule shows the method for handling missing [ITEM or TASK] responses in scoring is appropriate for assessing [CONCEPT OF INTEREST]	

E.3	*Other?*	
F	**Scores from the [NAME OF MEASURE] are not overly influenced by processes/concepts that are not part of [CONCEPT OF INTEREST].** *[Select and comment on appropriate rows for the type of COA]*	
F.1	[ITEM OR TASK] interpretation or relevance does not differ substantially according to respondents' [SEX, AGE, EDUCATION LEVEL, CULTRUAL/LINGUISTIC BACKGROUND, etc].	
F.2	Recollection errors do not overly influence assessment of [CONCEPT OF INTEREST].	
F.3	Respondent fatigue or burden does not overly influence assessment of [CONCEPT OF INTEREST].	
F.4	*Other?*	
G	**Scores from the [NAME OF MEASURE] are not overly influenced by measurement error.**	
G.1	Test-retest reliability coefficient indicates scores from the [NAME OF MEASURE] are not overly influenced by variation over time within clinically stable patients.	
G.2	*Other?*	
H	**Scores from the [NAME OF MEASURE] correspond to [MEANINGFUL ASPECT OF HEALTH] related to [CONCEPT OF INTEREST].**	
H.1	Correlation coefficients for the relationship between scores on the [NAME OF MEASURE] and [OTHER COA] are as hypothesized *a priori*.	
H.2	Empirical comparisons of scores from the [NAME OF MEASURE] for patient groups known to differ with respect to the [MEANINGFUL ASPECT OF HEALTH] show relationships hypothesized *a priori*.	
H.3	*Other?*	
Other?	*Other components needed to justify interpreting scores as measures of the concept of interest?*	

[a] Language shown is suggestive only. Sponsors should describe components using language that is clearest for their situation.

[b] Summary of supporting evidence with reference or link to more details about the evidence.