Patient-Focused Drug Development: Incorporating Clinical Outcome Assessments Into Endpoints For Regulatory Decision-Making

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

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within ___ days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to https://www.regulations.gov. Submit written comments to the Dockets Management Staff (HFA-305), Food and Drug Administration, 5630 Fishers Lane, Rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document, contact (CDER) Office of Communications, Division of Drug Information at druginfo@fda.hhs.gov, 855-543-3784, or 301-796-3400 or (CBER) Office of Communication, Outreach and Development, 800-835-4709 or 240-402-8010 or (CDRH) Patient Science and Engagement Program at CDRH-PRO@fda.hhs.gov, 301-796-6715.

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)

April 2023
Procedural

Patient-Focused Drug Development: Incorporating Clinical Outcome Assessments Into Endpoints For Regulatory Decision-Making

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
10903 New Hampshire Ave., Bldg. 71, Room 3128
Silver Spring, MD 20993-0002
Phone: 800-835-4709 or 240-402-8010
Email: ocod@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-device-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)

April 2023
Procedural

TABLE OF CONTENTS

Patient-Focused Drug Development: Incorporating Clinical Outcome Assessments Into Endpoints for Regulatory Decision-Making Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders[1]

This draft guidance, when finalized, will represent 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 staff responsible for this guidance as listed on the title page.

I. INTRODUCTION

A. Overview of the Series of FDA Guidance Documents on Patient-Focused Drug Development

This guidance (Guidance 4) is the fourth 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 and caregivers 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 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, as well as 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 the Food and Drug Administration Reauthorization Act (FDARA) of 2017), 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; and (B) patient preferences with respect to treatment of the 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 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)[5]

- Approaches to identifying what is most important to patients with respect to their experience as it relates to burden of disease/condition and burden of treatment (Guidance 2)[6]

- Approaches to selecting, modifying, developing, and validating clinical outcome assessments (COAs) to measure outcomes of importance to patients in clinical trials (Guidance 3)[7]

- Methods, standards, and technologies for collecting and analyzing COA data for regulatory decision-making, including selecting the COA-based endpoint and determining clinically meaningful change in that endpoint (Guidance 4; current guidance)

Please refer to Guidance 1, Guidance 2, and other FDA guidances[8] for additional information on collecting patient experience data. When final, the PFDD guidance series will replace the guidance for industry *Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims* (December 2009).

FDA encourages stakeholders to interact early with FDA and obtain feedback from the relevant FDA review division when considering the collection of patient experience data related to the burden of disease and the benefits, burdens, and harms of treatment.[9] FDA recommends that stakeholders engage with patients and other appropriate subject matter experts (e.g., clinical and disease experts, qualitative researchers, 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.

---

[5] See the FDA guidance for industry, FDA staff, and other stakeholders *Patient-Focused Drug Development: Collecting Comprehensive and Representative Input* (June 2020). 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] See FDA's guidance for industry, FDA staff, and other stakeholders *Patient-Focused Drug Development Methods to Identify What is Important to Patients* (February 2022).

[7] See the draft FDA guidance for industry, Food and Drug Administration staff, and other stakeholders *Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments* (June 2022). When final, this guidance will represent the FDA's current thinking on this topic.

[8] See FDA's 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) and FDA's guidance for industry, Food and Drug Administration staff, and other stakeholders *Principles for Selecting, Developing, Modifying, and Adapting Patient-Reported Outcome Instruments for Use in Medical Device Evaluation* (January 2022), or subsequent guidances in the PFDD series, when available.

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

B. Purpose and Scope of PFDD Guidance 4

This guidance is intended to help sponsors of clinical trials for medical product development, as defined in footnote 4. This guidance focuses on COA issues associated with clinical trial (study) endpoints, design, conduct, and analysis and will be of most relevance for those designing and conducting trials using COAs as well as analyzing and interpreting the trial data.[10] This guidance builds on Guidance 3 by focusing on endpoints constructed from fit-for-purpose[11, 12] COAs which are intended to reflect, directly or indirectly, how patients feel, function, or survive. Some COAs provide direct insight on how patients feel or function (e.g., a patient-reported outcome (PRO) instrument measuring pain intensity). Other COAs, however, may provide more indirect information to evaluate clinical benefit (e.g., clinician-reported outcome (ClinRO) instruments measuring extent or activity of disease such as psoriasis area and severity). In these situations, it is important to understand how the COA-based endpoint corresponds to changes relevant to patients (e.g., the type and extent of change that is meaningful to patients).

Section II of this guidance discusses considerations for COA-based endpoints to align the study design, endpoint, and analysis with the clinical study objective to improve study planning and the interpretation of analyses.

Section III of this guidance describes methods to aid in the interpretation of treatment effects on COA-based endpoints in terms of patients' views on the effect of a medical product. This information is important because statistical significance does not, by itself, indicate whether the detected effect corresponds to a clinically meaningful treatment effect.

[10] The considerations addressed in this guidance may be relevant to a variety of regulatory decisions that require an assessment of benefit or risk, 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 314; device approval decisions under the standards in sections 513(a)(2) and 515(d) and regulations in 21 CFR part 814; biological product approval decisions under the standards in section 351(a) of the Public Health Service Act and regulations in 21 CFR 601; 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.

[11] See the Agency's draft guidance for industry, Food and Drug Administration staff, and other stakeholders *Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments* (June 2022). When final, this guidance will represent the FDA's current thinking on this topic.

[12] A COA is considered fit-for-purpose when the level of validation is sufficient to support its context of use. Note that having a fit-for-purpose COA is necessary for a strong endpoint rationale, but it is not sufficient. For example, a COA that is considered fit-for-purpose for assessing symptom intensity might be used for an endpoint based on the average symptom intensity score across 7 days. However, if worst intensity were identified as the most relevant patient experience for improvement based on patient input and the product's mechanism of action, the rationale for using an endpoint of average symptom intensity would be very weak—despite being based on a fit-for-purpose COA.

Section IV of this guidance includes a list of additional considerations when developing an endpoint from a COA and formatting and submitting patient experience data from a clinical study supporting medical product regulatory decision-making.

Though the text and examples in this guidance focus mostly on treatment benefit (e.g., improvement in disease-related symptoms or impaired functions), COAs also can be used to assess treatment harms including symptomatic adverse events and other burdens to the patient associated with the medical product under study. While many of the recommendations in this guidance will apply to the evaluation of treatment benefit or risk, additional considerations may be needed when using COAs to inform treatment 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 the word *should* in Agency guidances means that something is suggested or recommended, but not required.

II. COA-BASED ENDPOINT CONSIDERATIONS

This section discusses considerations for selecting COA-based endpoints, including the development of a well-justified rationale for the endpoints and considerations for statistical analyses of COA-based endpoints in clinical trials.

A. Endpoint of Interest: What Are You Measuring in the Target Study Population?

PFDD Guidance 3 discusses the importance of a fit-for-purpose COA. PFDD Guidance 4 complements PFDD Guidance 3 by focusing on the rationale for the proposed use of COA scores to construct endpoints that will support inferences about the effect of a medical product on how patients feel or function. As with the rationale for interpreting COA scores as measures of the concept of interest, the rationale for the use of COA scores as the basis for an endpoint should be well-supported by evidence.

1. Selecting and Justifying Endpoints

Generally, endpoints that are based on COAs should (1) reflect an aspect of the patient's health that is meaningful; and (2) be capable of supporting an inference of treatment effect within the context of the planned clinical trial. For a given COA score, there may be multiple options for constructing a trial endpoint (e.g., mean score at 12 weeks or time to complete symptom resolution).

Sponsors should clearly describe the COA-based endpoint, including:

- Type of assessment(s) made (e.g., Patient-Reported Outcome (PRO) measures, Observer-Reported Outcome (ObsRO) measures, Clinician-Reported Outcome (ClinRO) measures,

Performance Outcome (PerfO) measures).

- The COA(s) used to measure the concept(s) of interest. Note that it is important for endpoints to be assessed using a COA that is fit-for-purpose. For details, see draft PFDD Guidance 3.

- Specific score(s) from the COA (e.g., specific subscale score, total score).

- If a multi-component endpoint, the algorithm used to combine scores from two or more components into a single endpoint.

- Rules for handling missing item responses or task results when computing COA scores, along with justification for the rules.

- Timing of the assessments used to construct the endpoint, the timeframe over which COA scores are combined to construct the endpoint, and a detailed description of how COA scores collected during the treatment period are combined into an endpoint (e.g., score at week 12, average daily scores for 7 days prior to week 12 study visit, maximum value of the daily 200 mobile sensor assessments for 7 days prior to the week 36 study visit.). Also, if the endpoint is defined in terms of change from baseline to some follow-up assessment, then the definition of "baseline" should be clear.

FDA recognizes that constructing and selecting trial endpoint(s) often involves weighing the strengths and limitations of different approaches. Early in the planning of a clinical trial, sponsors should provide to FDA a well-supported rationale for the selection of the endpoint(s) by explaining why each endpoint is informative for the trial context. The rationale for endpoint selection typically will address the following:

- Concept(s) of interest.

- Clinical trial objective or hypothesis corresponding to the endpoint, ensuring that the objective/hypothesis is specific (e.g., "To compare the patient-reported physical functioning between arms at 24 weeks" rather than "To compare the patient-reported outcomes of product X vs. Y").

- The role of the endpoint (e.g., primary, secondary, other).

- Intended indication related to the COA-based endpoint.

- Explanation for why the selected COA is fit-for-purpose in the planned trial.

- Support for the importance of the endpoint to patients and/or caregivers from literature

review and/or primary data collection.[13] In some cases, for endpoints based on a COA that measures a concept of interest that is indirectly related to some meaningful aspect of health for the patient (e.g., based on a neurological functioning test that is thought to be indicative of the patients' cognitive functioning), it might be sufficient to provide support for the adequacy of the endpoint for measuring this aspect of health. Furthermore, there are well-established relevant outcomes such as organ failure and death that do not require additional support. If a multi-component endpoint, justification for the components included and the algorithm for combining them into the endpoint.

- Strengths and limitations of the proposed endpoint.

An endpoint's use in another trial evaluating a different product may not be adequate support for the use of the same endpoint for a trial under consideration, because the context of use can vary in important ways from trial to trial and science and/or policy might have evolved since the endpoint was last used. When disease-specific FDA guidances exist, sponsors should consult these for recommendations for suitable endpoints.[14]

2. Considerations for Constructing a COA-Based Endpoint

This section provides guidance on using scores from one or more COAs to construct endpoints for specific circumstances as well as guidance regarding particular types of endpoints. This is not a comprehensive review of all possible types of endpoints but rather a discussion of frequently encountered challenges for COA-based endpoints.

a. Considerations for baseline administration of COAs relevant to COA-based endpoints

Prior to discussing the different approaches, several considerations about collecting COAs at baseline should be noted:

- Some diseases, conditions, or clinical trial designs may necessitate more than one baseline assessment or longer/shorter baseline periods.

- When multiple baseline measurements are taken, the protocol should define how the baseline value will be calculated from the multiple measurements.

- A screening visit that includes administration of the COA is often used to ensure that patients enrolled in the trial have a sufficient level of severity so that improvement could

[13] For example, Stone et al. (2021) conducted semi-structured interviews with patients who have chronic pain (as well as clinicians and clinical trialists) to elicit their understanding of and preferences for seven different endpoints that could be constructed based on intensive longitudinal assessments of pain intensity (e.g., average pain over a week, worst pain intensity over a week, time spent with low or no pain). Patients were asked to rank the different endpoints in the order of what they were "most hoping for as a result of treatment."

[14] Please see the FDA guidance web page https://www.fda.gov/regulatory-information/search-fda-guidance-documents

be observed. To avoid regression to the mean and other potential sources of bias,[15] the COA score obtained at screening should not be used as the patient's baseline value. Rather, a separate, later pre-randomization assessment should be used as the patient's baseline value.

- If the trial includes a run-in period during which the patient's score from the COA might be expected to change (e.g., medication washout, patient behavior modification), then this should be considered when planning the timing of assessments.

b. Endpoints based on COA scores at a fixed time point or a summary of COA scores over time

In most situations in which a COA produces ordinal or continuous (interval or ratio scale) scores, the best and recommended endpoint will be the COA score at a predefined assessment point or summarized over some predefined post-baseline assessment period, and the most straightforward analysis will be a comparison of randomized groups with respect to the follow-up score(s) after adjusting for the baseline value (e.g., with a linear model to compare average follow-up scores).

When the endpoint is based on COA scores at a predefined assessment point, sponsors should justify the use of, and time at which, an analysis at a fixed time point (e.g., 12 weeks) is to be performed. For example, an analysis at a fixed time point might be justified if the COA score is not highly variable over time and the chosen time point (e.g., end of study) would be useful for reflecting the durability of the treatment effect. Justification of the fixed time point should also take the recall period of the COA (where applicable) into consideration.

When considering an endpoint based on summarizing COA scores over some predefined post-baseline assessment period, different summaries may be appropriate depending on the research questions. Common types of summaries include the patient's mean score over a fixed time period, the maximum (or minimum) score during some period (e.g., worst pain recorded during a 7-day period). For some types of summaries, an alternative approach is to use repeated measures modeling of all observed COA scores and derive summary estimates from the model. Regardless of the approach taken, sponsors who wish to construct an endpoint based on summaries of patients' COA scores over time should consider the robustness of the summary (or model) and any modeling assumptions, handling of missing COA scores, statistical power, and interpretability.

c. Endpoints constructed by dichotomizing COA scores

COA scores are often ordinal or continuous (interval or ratio scales) in nature. When this is the case, defining the endpoint using the ordinal or continuous COA score, rather than making the endpoint dichotomous, uses all the information and therefore usually maximizes statistical power. In some cases, dichotomized endpoints (e.g., "responder" status) are well-established and can be reasonable choices when it is important to evaluate the effect of treatment on the

[15] Shaw PA, Johnson LL, Proschan MA, 2018, Intermediate Topics in Biostatistics, In JI Gallin, FP Ognibene, LL Johnson (Eds) *Principles and Practice of Clinical Research* (4th ed), London: Academic Press, pp. 384-409.

probability of achieving clearly defined and important health states. Examples of such health states might be complete patient-reported symptom resolution or investigator's global assessment of acne lesions as "clear" or "almost clear" (see the May 2018 guidance for industry *Acne Vulgaris: Establishing Effectiveness of Drugs Intended for Treatment*). If a sponsor wishes to use an endpoint based on dichotomization from either ordinal or continuous data, the sponsor should prespecify a single score threshold and provide evidence to justify the dichotomization in the endpoint rationale. For example, FDA recommends that the rationale include evidence that patients and/or their caregivers view health states above the threshold to be meaningfully different from health states below the threshold. This recommendation also applies to the use of ordinal or continuous COA data to define an event for a time-to-event endpoint. Of note, data used to derive a score threshold(s) should be different than that used to demonstrate effectiveness (e.g., data from registration trial(s)). In addition to prespecifying a single score threshold, sponsors should also conduct analyses to explore treatment effects over a range of thresholds.

Sometimes the motivation for dichotomizing an ordinal or continuous COA-based score is to make the endpoint more interpretable for patients, caregivers, and/or clinicians. This is typically possible without creating a dichotomized endpoint for the primary analysis of treatment effect. (See Section III, *Evaluating the Meaningfulness of Treatment Benefit*).

d. Endpoints constructed by computing change from baseline or percent change from baseline COA scores

As discussed in Section II.A.2.a, in comparative trials, the preferred method for adjusting for baseline status is to do so in the context of a statistical model. Using the COA score's change-from-baseline as an endpoint is another option, but it has some important considerations:

- COA scores that are ordinal are challenging to interpret in terms of change from baseline because the difference between two ordinal scores cannot be assumed to have the same meaning across scores (e.g., for an ordinal score with 5 levels—when interpreting level 3 relative to level 1 and level 5 relative to level 3—both differ by two levels but might not correspond to the same degree of change in the underlying health state). Put another way, there might not be a linear relationship between the ordinal values and the true level of symptom severity or functioning being measured.

- If it aids interpretation to express treatment effects in terms of change-from-baseline, this can be done in the context of most models used to compare treatment groups on follow-up scores adjusting for baseline. For example, an ANCOVA model could be used to derive the predicted follow-up score on treatment for patients with a given baseline score, and these two values could be used to compute a predicted change-from-baseline score.

- For situations in which it is not possible to conduct a randomized, controlled trial and a single arm trial is done instead (e.g., to evaluate some devices), a change-from-baseline endpoint might be the best available option.

A similar endpoint that could be considered is the percent change-from-baseline. An advantage of this approach might be easier interpretability, but in addition to the considerations presented

for change-from-baseline endpoints, several important challenges are worth noting about percent change-from-baseline:

- Interpretation can be complicated by the fact that percent change-from-baseline is asymmetric; that is, it treats the baseline and follow-up COA scores differently (Berry and Ayers 2006). For example, consider two patients who are randomized to receive a new medical product. The first patient's COA score improves from 5 to 10 (change = +5) and the second patient's score decreases from 10 to 5 (change = -5). In both cases, the absolute change is 5, but the percent change is very different: +100% and -50%. This has important implications, including the fact that the average change on the original scale (0) indicates no overall change, whereas the average percent change ([+100 – 50]/2 = +25%) suggests an overall improvement.

- Percent change-from-baseline is undefined if the baseline score on a COA is zero, and some kind of imputation is required to include the observation in the analysis.

- Compared to follow-up scores or change-from-baseline scores, percent change-from-baseline scores may have highly non-normal distributions that can be challenging to model.

If the reason for considering percent change-from-baseline is that the treatment effect is expected to be multiplicative rather than additive (e.g., treatment improves a patient's symptom severity by 20% of the patient's severity level without treatment), then a logarithmic or similar transformation could be applied to continuously distributed COA scores prior to comparing groups (Senn 2007).

e. Endpoint strategies when a disease affects multiple aspects of feeling and functioning

A disease might manifest in multiple ways, in which case it is important to consider how or whether a medical product affects different aspects of health. Some aspects of health might be relevant for almost all patients with a given condition (e.g., pain associated with migraine). Other affected aspects of health might differ between patients and within patients over time with certain conditions (e.g., lupus, sarcoidosis, primary mitochondrial diseases, schizophrenia, and many rare diseases). In these situations, it may be challenging to identify one specific aspect of the disease for evaluating treatment benefit. It may be necessary to consider several different aspects to adequately assess benefit. FDA recognizes that selection of the endpoint(s) in these situations is likely to involve weighing the strengths and limitations of various approaches. When possible, sponsors can evaluate multiple endpoints in earlier phase trials to inform the selection of endpoints for later trials.

This section reviews three general strategies for constructing endpoints when multiple aspects of health might be of interest: (1) separate endpoints for each aspect of health, (2) a multi-component endpoint, and (3) a personalized endpoint.

Construct Separate Endpoints for Each Aspect of Health

As described in the guidance for industry *Multiple Endpoints in Clinical Trials* (October 2022), if a separate endpoint will be constructed for each aspect of health, their role should be described, with the main options as follows:

- *One primary endpoint and multiple secondary endpoints*. This option might be useful when there is one core or cardinal manifestation of a disease (primary endpoint) that most patients can be expected to experience and that is regarded by patients and/or caregivers as important. Secondary endpoints can be created for aspects of health that might not be experienced by all patients and/or are viewed as relatively less critical, but still important, to patients and/or caregivers.

- *Multiple primary endpoints*. This option might be useful when an improvement in at least one aspect of health would be regarded as evidence of treatment benefit.

- *Co-primary endpoints*. This option may be appropriate when there are multiple aspects of health that are critically important to the disease being studied, such that a treatment benefit can only be concluded if the medical product has an effect on each of the designated endpoints.

By creating a separate endpoint for each relevant aspect of health, there is clarity about which aspect of health has or has not been affected by the medical product, because each endpoint corresponds to only one aspect of health. But there are several issues with this approach that also should be considered. First, for diseases with many possible manifestations, the approach may be challenging to use if it is not known ahead of time which aspects of health are most likely to improve as a result of using the medical product under study. Second, depending upon the roles of the multiple endpoints, multiplicity adjustments might be needed, necessitating a larger sample size to ensure sufficient statistical power. Finally, if patients differ from one another in their symptoms or functional impacts due to the disease, then the treatment effect estimated for any one endpoint will be diluted by the patients for whom the endpoint is not relevant (e.g., patients who never had a given symptom cannot improve with treatment). Consult the guidance for industry *Multiple Endpoints in Clinical Trials* (October 2022) for additional information on constructing and analyzing multiple endpoints in a single trial.

Construct a Multi-Component Endpoint

A multi-component endpoint is based on a within-patient combination of two or more components, each reflecting a different aspect of health. Constructing the endpoint for an individual patient requires observation of all the specified components for that patient. Then a single overall rating or status on the endpoint is determined according to a prespecified algorithm.

A COA-based multi-component endpoint may take many forms. The individual components could be (a) scores from different COAs, (b) scores from multiple subscales of a single COA, or (c) responses to individual items or tasks that make up a single COA.[16]

Some COA-based multi-component endpoints are constructed by combining the patient's scores—in their original metric or transformed (e.g., dichotomized)—from two or more components according to an algorithm. Some examples include:

- An overall symptom index score created by using a well-justified weighted combination of responses to separate items that each assess a different type of symptom.

- Patients' endpoint values ("improved" versus "not improved") are assigned based on a more complex algorithm, for example, an algorithm requiring some minimum change-from-baseline for one COA and some minimum change on at least two of four other COAs.

Other multi-component endpoints are constructed with the objective of demonstrating the absence of all symptoms. Examples include:

- Achievement of complete resolution of all symptoms
- Total time without any symptoms during some predefined post-baseline period
- Time until complete resolution of all symptoms
- Time to sustained clinical recovery assessed over an appropriate duration

There are several advantages to using a multi-component endpoint, including:

- A multi-component endpoint has the potential to evaluate the entire range of important disease manifestations. Because patients may experience some aspects of a disease more than others—and some aspects, not at all—a multi-component endpoint lends itself to capturing a treatment effect more so than an endpoint that evaluates a narrower aspect of the disease.

- No multiplicity adjustment is needed to control the chance of erroneous conclusions (e.g., Type 1 error) for a multi-component endpoint compared to the use of multiple separate endpoints.

- The use of within-patient multi-component endpoints can be efficient if the treatment effects on the different components are generally concordant.

[16] Responses to individual items or tasks that make up a single COA could be treated as individual components of a multicomponent endpoint only when the COA is based on a composite indicator measurement model. In a composite measurement model, responses to the items or tasks are not assumed to be reflective of or caused by a single underlying aspect of health (as they would be for a reflective measurement model). Instead, each item or task addresses a separate health concept and, when combined, responses to all the items or tasks define the overall concept of interest. See the draft guidance for industry *Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments* (June 2022), Section IV.E.

These advantages should be weighed against important concerns and limitations with constructing certain types of multi-component endpoints, including:

- For endpoints that are based on complete resolution of all symptoms, it might be difficult to achieve complete resolution with a medical product in the context of a clinical trial. Furthermore, some patient populations might not require complete resolution of all symptoms to feel they have benefitted from treatment. Other endpoints may be advisable to assess treatment-related improvements in individual symptom intensity or frequency.

- For multi-component endpoints that sum or average over scores from multiple components, a clinically meaningful improvement in one COA becomes increasingly diluted as more COAs are included in the construction of the endpoint. For example, if a patient's only manifestation of a disease is symptom *A*, then the patient might appear to show little improvement if the multi-component endpoint averages the status on symptom *A* with symptoms *B, C, D,* and *E*. Therefore, sponsors considering this type of multi-component endpoint should balance the ability to observe improvements in any of several aspects of health with the chance that improvements in one aspect will be diluted by aspects that were never a problem for the patient. Sponsors might also consider the use of a personalized endpoint in such situations (see *Construct a Personalized Endpoint* below).

- All multi-component endpoints are based on some implicit or explicit weighting scheme. This includes multi-component endpoints that imply that all components have reasonably similar clinical importance, such as when taking the average across multiple COAs or assigning the status of "improved" to a patient who shows improvement in scores for any 1 of 5 COAs. Sponsors should be explicit about how each component is weighted in constructing the endpoint and provide justification for the weights.

- When a treatment effect is found using a multi-component endpoint, it may be helpful to examine the treatment effect for individual components. For more detail about when and how to examine individual components, see the guidance for industry *Multiple Endpoints in Clinical Trials* (October 2022).

- There are several challenges for endpoints that rely on categorizing meaningful changes in one or more COAs.

  - Endpoint values are strongly dependent on the thresholds selected for meaningful improvement and/or worsening and choosing such thresholds can be challenging. Thresholds for each COA should be predefined and justified. Sponsors should also conduct sensitivity analyses that explore treatment effects over a range of thresholds.

  - There is the potential for bias when those completing or administering the COA are aware of the thresholds for being considered a meaningful improvement (or worsening). It is important when possible that clinicians (for ClinRO measures), caregivers (for ObsRO measures), and/or any research staff (for PerfO measures)

involved in assessment are not made aware of the threshold definitions and are masked[17] to treatment assignment.

- Endpoints that assign values of worsened = -1, no change = 0, and improved = +1 assumes that the patients view the degree of improvement and deterioration in a concept of interest as symmetric, which may not be the case.

Construct a Personalized Endpoint

Personalized endpoints are sometimes proposed to reflect what is important to each individual patient enrolled in a clinical trial, especially for diseases with variable clinical manifestations that impact patients differently. Several examples include the following:

- The "most bothersome symptom" approach in which patients identify at baseline the one disease-related symptom that is most bothersome to them. The patient's status on that symptom post-randomization then becomes the outcome to be analyzed (Duke Margolis Center for Health Policy 2017). A similar approach is based on patients identifying at baseline the symptom that is "most severe" for them (which may or may not be the symptom that is most bothersome for them).

- Goal Attainment Scaling (GAS; Krasny-Pacini et al. 2016) in which each patient identifies a prespecified number of personal goals (e.g., being able to work in the garden) at baseline. At one or more post-randomization assessments, the patient records their status with respect to each goal using a standardized response scale and the responses are summarized across the patient's goals. Whereas the "most bothersome" and "most severe" symptom approaches are based on assessments of symptoms, GAS usually is based on assessment of functioning.

Personalized endpoints have several advantages, including:

- They are very patient focused in their attempt to reflect how each patient feels or functions in terms of what is most important to them at baseline.

- Because each patient's endpoint value is based only on what was identified as an issue for them at baseline, there is no dilution of treatment effect due to mixing affected and unaffected patients (i.e., when treating each aspect of health as its own endpoint) or mixing affected and unaffected aspects of health within a patient (i.e., when constructing some multi-component endpoints).

- Depending upon the context of use, a personalized endpoint could be considered along with another endpoint to inform decisions about the effect of a medical product. For example, the FDA guidance for industry *Migraine: Developing Drugs for Acute*

[17] Keeping study group assignment hidden from those involved in a study or trial 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.

Treatment (February 2018) describes using two co-primary endpoints: (1) having no headache pain at 2 hours after dosing; and (2) a demonstrated improvement on the patient's most bothersome migraine-related symptom at 2 hours after dosing. (Note that this approach is specific to the context of use and might not be appropriate in other contexts of use.)

These advantages should be weighed against several concerns, including:

- For personalized endpoints that rely on patients choosing a single "most bothersome" or "most severe" symptom, it might be difficult for patients to select a single symptom.

- Changes might occur over the duration of a clinical trial in what patients regard as their "most bothersome" symptom, "most severe" symptom, or their most important personal goals.

- It is possible that patients might choose symptoms or areas of functioning (for GAS) at baseline that are not targeted by the product being evaluated or that might not be realistic to achieve for patients in the target population.

- The outcomes chosen by patients might not reflect new or worsening symptoms and/or functional limitation(s) that occur during the trial duration. For this reason, the same set of outcome assessments should be assessed for all patients regardless of their own personalized endpoint.

- The processes for eliciting personalized endpoints have the potential for inconsistency. Therefore, the process to construct a personalized endpoint should be standardized and the criteria for selecting the outcome assessments should be consistent across sites and patients.

- As with multi-component endpoints, it is challenging to describe the specific effect of the treatment on a personalized endpoint. For this reason, it is important to measure all relevant symptoms and areas of functioning in addition to those identified as most important to the individual patients. This will make it possible to conduct prespecified treatment comparisons for individual symptoms and types of functioning.

3. Clinical Trial Duration and Timing of Assessments for COA-Based Endpoints

Generally, COA data should be collected over the duration of the clinical trial, as indicated for other measures of effectiveness or safety in the clinical trial protocol.

The timing of assessments plays a vital role in gaining reliable and meaningful information on the concept(s) of interest reflected in the COA-based endpoint and should be selected carefully and be scientifically justified. Clinical trials using COAs should include a schedule of COA administration as part of the overall study assessment schedule in the protocol. The COA schedule should consider the natural course of the disease or condition (i.e., acute, chronic, or episodic), the research questions to be addressed, the trial duration, patient burden, the disease

stage of the target patient population, the expected time frame when the investigational product is likely to affect the COA-based endpoint, and timing of collection of COAs if temporary study interruptions or discontinuation of study interventions are anticipated to occur.

In general, COA assessment frequencies or the rules governing when the COA is measured should be the same for all treatment arms (see event-triggered data collection below). In many instances, such as when a COA is planned to be frequently measured (e.g., event-triggered data collection) or when the COA is complex and potentially burdensome, sponsors might consider seeking input from members of the patient community to ensure that the planned length of the trial and timing of COA assessments is feasible and as convenient as possible for the patients and/or caregivers. This input may help to reduce missed assessments and study dropout. Sponsors can further reduce patient burden by including only those assessments that are well justified within the context of the study objectives. See Section IV.A.7 (*Minimizing Participant Burden*) for more discussions.

Other important considerations for determining the most appropriate timing of assessments for COA-based endpoints include, but are not limited to, the following:

- *Event-triggered data collection:* In some studies, COA administration may be triggered to occur during or following events such as urination or an asthma exacerbation. For this type of data collection, consider the windows for data collection around an event and whether it would be appropriate to prompt to ensure that all events were collected (i.e., at the end of the diary day). For example, for a trial evaluating a treatment for a disorder that results in difficulty or excessive frequency of urination, a participant could be asked to record each urination episode and complete a short assessment immediately following the event (e.g., pain or burning during urination, post-micturition dribble). Then, at the end of the diary day, the patient could be shown a list of reported urination episodes and asked if they had any other urination episodes that needed to be reported and assessed.

- *Anticipated rate of change in the underlying concept of interest to be measured:* The timing of assessments should align with the anticipated nature and rate of change in the underlying concept of interest to be measured. For example, if the concept of interest to be measured is expected to change rapidly over the course of the study period, then assessments should be placed closer together. If the concept of interest is expected to change slowly, then assessments can be placed further apart.

- *Ability to assess time-to-event endpoints*: If the trial endpoint is based on time to achieve an outcome of interest (e.g., time to complete symptom resolution), the frequency of assessment should be sufficient to assess clinically meaningful differences in the time to the outcome of interest. If assessments are made too infrequently, important differences between trial arms may not be detected.

- It will typically be of interest to understand treatment effects regardless of adherence to treatment, such that the protocol should include plans to continue to follow patients and administer the COA after discontinuation of treatment.

B. Estimation and Missing Data

The statistical analysis considerations for COA-based endpoints are similar to the statistical considerations for any other endpoint used in medical product development. This section briefly discusses several considerations that commonly arise when estimating COA-based estimands,[18] including missing data.

1. Analysis at a Fixed Time Point

For evaluating a treatment effect on COA scores at a fixed time point, the statistical power of the treatment group comparison is generally better when the comparison is statistically adjusted for patients' baseline scores[19] on the COA (see the draft guidance for industry *Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products* (May 2021)).[20] This recommendation also applies when the endpoint is the change in COA score from baseline to a predefined time point.

If a COA-based endpoint is collected repeatedly, data from intermediate time points (i.e., measurements taken prior to the fixed time point) can still be included in a longitudinal (e.g., mixed-effects or generalized estimating equations) model in which a treatment contrast is made for a prespecified fixed time point.

2. Analyzing Ordinal Data

Sometimes COA scores are used to construct an endpoint that results in an ordinal metric. Several analytic options exist for ordinally scaled endpoints. The choice of analytic approach might depend on the type of ordinal endpoint. For COA-based endpoints, there are generally two situations that generate an ordinal scale:

- *An ordinal endpoint based on a COA measuring a single aspect of health*. For example, a group comparison at a fixed time point might be made using a single item COA measuring the intensity of musculoskeletal pain might have response options of *none*, *mild*, *moderate*, and *severe*, which are scored as 0, 1, 2, and 3. The steps between successive levels might not reflect equal increments in pain, and so it might be challenging in some cases to interpret an estimate of treatment effect in terms of mean differences (e.g., as generated by an ANCOVA). On the other hand, an approach that tries to simplify the endpoint for analytic purposes by dichotomizing (e.g., [0 or 1] vs [2 or 3]) risks ignoring important information about patients' relative standing on the

[18] An *estimand* is defined as a precise description of the treatment effect reflecting the clinical question posed by the trial objective. It summarizes at a population-level what the outcomes would be in the same patients under different treatment conditions being compared (see the ICH guidance for industry *E9(R1) Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials* (May 2021) (ICH E9(R1)).

[19] Patient or clinician global impressions of severity, when used as anchor variables (see Section III), should be assessed at baseline. Note that patient or clinician global impressions of change used as anchor variables are not administered at baseline. Also, the concept of baseline or baseline symptoms may be complicated in certain study designs such as prophylaxis trials. Finally, some endpoints defined using event-triggered assessments might not be possible to assess at baseline.

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

concept of interest. An ordinal modeling approach (e.g., cumulative logistic regression; Agresti, 2013; Harrell, 2015) has different assumptions than a general linear model and may incorporate more information in the endpoint than the dichotomization approach. The key point when choosing an analytic approach is that the results are interpretable and address the appropriate clinical question. Regardless of the approach taken, sponsors should explore the potential impact of violation of assumptions.

- *A multi-component endpoint constructed by assigning ordinal values based on scores reflecting multiple aspects of health*. This type of multi-component ordinal endpoint might mix distinct aspects of a disease, such as symptom levels, hospitalization, and death. The ordinal values are assigned by an algorithm to reflect increasingly severe disease states. While the same analytic approaches could be considered for this type of ordinal endpoint, greater caution is required in interpreting the findings. There could be a situation where ordinal multi-component endpoints that mix distinct aspects of a disease in which treatments are beneficial in terms of one aspect of health (e.g., severity of symptoms) but are harmful in terms of another aspect (e.g., mortality). It is possible in these situations that estimates of treatment effect from common analytic methods such as ANCOVA and cumulative logistic regression may show overall treatment benefit but could obscure harmful effects. Sponsors should consult FDA when developing analytic plans for such ordinal, multi-component endpoints.

3. Missing Data

Missing data are problematic because they may lead to reduced power and potential bias in the estimated treatment effect when missingness is related to treatment effectiveness or to adverse events from the treatment. Two types of missingness can occur for COA-based endpoints: (1) missing responses to items or tasks that make up a COA; and (2) missing an entire COA at a given time point.[21]

Every effort should be made to avoid missing COA data. This begins with collecting only those COAs necessary to assess the endpoint (e.g., for efficacy, safety, tolerability) and designing a data collection plan that is least burdensome and as easy as possible for patients and/or caregivers. This includes counseling patients on the importance of completing the COA and providing reminders when the patient needs to complete the COA. When a person does not complete a COA at a given time point, the site should be notified so that research staff can contact the appropriate person (patient, caregiver, study, or site staff) to obtain the needed assessment. It is important to collect reasons for missing data to inform suitable sensitivity analyses of the study endpoints considering different approaches to account for the missing data. The ability of the COA-based endpoint to address the clinical question of interest will depend on the amount of and reasons for missing data and how plausible the missing data assumptions are for the study.

[21] Missing data should be distinguished from intercurrent events (e.g., death). Within the estimand framework, intercurrent events are things that happen after randomization that might affect the ability to observe or the interpretation of an endpoint. Potential intercurrent events and methods to handle intercurrent events should be addressed in the statistical analysis plan. For additional details, please see ICH E9(R1).

Missing item-level COA data should be handled based on the scoring algorithm for the instrument. In cases where patient-level COA data are missing for the entire domain(s) or the entire measurement(s), sponsors should propose statistical methods that properly account for missing data with respect to a particular estimand.

Methods to handle the missing data for a COA-based endpoint should be aligned with the estimand of interest and addressed in the statistical analysis plan.

III. EVALUATING THE MEANINGFULNESS OF TREATMENT BENEFIT[22]

In regulatory decision-making, FDA evaluates how well results of a COA-based endpoint correspond to a treatment benefit that is meaningful to patients. For endpoints based on COAs intended to reflect how patients feel or function (see Section I.B), sponsors should provide supporting evidence to justify the meaningfulness of an observed treatment benefit. Section III discusses what supporting evidence is recommended, how it could be collected, and how it can be applied to help interpret the trial results. FDA strongly recommends that sponsors seek FDA input as early as possible regarding the evaluation of meaningful treatment benefit.

A. Factors Affecting the Interpretability of COA Scores

To determine whether a medical product has a positive, meaningful effect on how a patient feels or functions (i.e., a treatment benefit[23]), FDA recommends that sponsors measure how a patient's status on a COA-based endpoint corresponds to the way they feel and/or function in their daily life. For example, if a treatment is shown to reduce scores on a performance outcome measure by an average of 2 points on a 15-point scale, it would be helpful to know whether a 2-point difference corresponds to something that patients would notice as important in their daily lives. Or, if a treatment is expected to increase a patient's score on a measure of functioning from 12 to 18, it would be helpful to know what kinds of things the patient could do (or do more easily) corresponding to a score of 18 versus 12. Knowing how COA scores relate to patients' experiences is central to interpreting the meaningfulness of a COA-based endpoint result(s). This is true whether the endpoint is based on scores generated from a single COA or multiple COAs (as in a multi-component endpoint).

Some COAs might produce scores that are easier to interpret than other COAs in terms of patients' experiences. How easily one can interpret a COA score depends on at least two factors:

[22] Most of the methods described in this section for interpreting trial results can apply to treatment impacts other than those described as "benefit." These could include treatment tolerability or harm in terms of how the patients feel, function, or survive. However, for brevity this section will refer only to treatment benefit.

[23] Treatment benefit is also demonstrated by a favorable effect on how patients survive, but this is not relevant for the discussion of COA-based endpoints.

718
719 *1. How Closely Does the Measured Concept of Interest Correspond to the Patients'*
720 *Experiences?*
721
722 Some COAs measure a concept of interest that is a directly interpretable reflection of the
723 patients' health-related experiences, such as a PRO measure of current pain intensity. For such
724 measures, it may be relatively easy to infer how different scores on the measure correspond to
725 different experiences the patients might have. Other COAs might measure a concept of interest
726 that is more indirectly related to the patient's health-related experiences, such as an ObsRO
727 measure of the patient's pain behavior (which is indirectly related to the patient's actual pain) or
728 a PerfO measure of leg strength (which is indirectly related to activities that require lower limb
729 function such as walking or climbing stairs).[24] For these types of measures, it may be more
730 challenging to infer how different scores on the measure correspond to different experiences the
731 patients might have; this means that additional empirical support is needed to translate scores on
732 the measures to corresponding patient experiences in their daily lives.
733
734 *2. How Simple or Familiar is the COA's Metric?*
735
736 In addition to how closely the concept of interest corresponds to the patient's direct experience,
737 the metric that is used to express the COA scores can also be more or less easy to interpret.
738 Some COAs produce scores that are easier to interpret on their own because they use a metric
739 that is relatively simple and/or familiar. For example, a daily diary that records the number of
740 times per night that a patient woke up to urinate would generate a directly interpretable metric
741 (i.e., number of times per night). Another example might be a simple ordinal rating of pain
742 severity (e.g., none, mild, moderate, severe) that generates a score that most patients have little
743 trouble interpreting in terms of noticeable gradations between patients' experiences. Cognitive
744 interview data might confirm that patients are comfortable evaluating their symptom severity
745 with this scale and that patients view each category as corresponding to a meaningfully distinct
746 experience. In this case, the scores themselves are directly interpretable in terms of patients'
747 experiences, and therefore, additional supporting evidence may not be necessary for
748 interpretation.
749
750 Other COAs produce scores that are more difficult to interpret on their own because they use a
751 metric that is unfamiliar and/or abstract, such as a COA measure that produces transformed
752 scores (e.g., linear transformation of a 0-4 raw score scale to a 0-100 score scale). There might
753 be very good reasons to generate a score on such a metric, but it increases the complexity of
754 describing the endpoint in labeling. In this case, FDA recommends additional evidence to justify
755 how scores relate to meaningful patient experiences.
756

[24] Indirect measures of patients' experiences could be recommended for many reasons, including the patients being incapable of self-reporting (e.g., too young, suffering from cognitive impairments) or a concern that heterogeneity in environments will create undesirable noise in self-reports of functioning (which may suggest the use of a PerfO measure).

B. Approaches for Collecting Evidence to Support Interpretability of COA-Based Endpoints

Sponsors should first review any existing evidence in support of the interpretability of the COA scores used to construct the endpoints. If the body of evidence supporting the interpretability of COA scores (e.g., from existing literature) is not sufficient, FDA recommends conducting empirical studies to support interpretability of COA scores prior to conducting a registration trial. When feasible, it is advantageous to use multiple methods to inform interpretations of scores. It is expected that empirical approaches will generate a range of plausible estimates reflecting the inherent uncertainty in interpreting scores. Based on such empirical studies, sponsors should prespecify the range of estimates that will be used to interpret the treatment effect(s) in a registration trial. The following sections describe two general approaches for conducting empirical studies to support the interpretability of COA scores–interpreting in terms of meaningful score differences (III.B.1) and in terms of meaningful score regions (III.B.2).

1. Interpreting in Terms of Meaningful Score Differences

This first approach identifies what size difference between any two COA scores would be viewed as meaningful for patients. This will be referred to as the *meaningful score difference* (*MSD*). Often, *MSD* is determined based on what patients would regard as a clinically meaningful within-patient change (i.e., improvement or deterioration from the patient's perspective), but other approaches might also be appropriate (e.g., those based on the patient's perception of the differences between hypothetical vignettes representing different degrees of symptom severity or functioning). Note that patients differ in their views of what might count as *MSD*, but for purposes of evaluating the results of clinical trials, a range of *MSD* should be selected that reflects most patients.

Regardless of the approach used to determine the *MSD*, the *MSD* can be used in at least two ways: (1) to evaluate the expected treatment effect for the average patient in some target population; or (2) to use as a threshold in descriptive analyses that identify individual patients who might have changed by a meaningful amount. Both of these applications will be discussed (see III.C) following a review of approaches for selecting a value or range of values for *MSD*.

Key assumptions should be identified and evaluated before *MSD* can be used to interpret the meaningfulness of a treatment effect in a clinical trial. Two common assumptions that should be evaluated are the following:

- The value of *MSD* is the same regardless of the baseline COA score (Crosby et al. 2003). For example, if *MSD* is specified as 4 points, then score differences of 5-1, 10-6, and 15-11 should all be regarded as meaningful differences by patients. If this assumption is not true, it is possible to use different values for *MSD* depending on the patient's baseline status.

- The value of *MSD* is the same for improvement and deterioration (Crosby et al. 2003). If this assumption is not true, then it is possible to use different values for *MSD* depending on the direction of change.

Sponsors can consider the use of anchor-based methods for identifying *MSD*. An anchor is some external variable, not derived from the COA whose scores require interpretation, for which meaningful differences are directly interpretable or already known.[25] Meaningful differences on the anchor can then be mapped onto differences in terms of the COA scores. For example, a patients' categorizations of their change in symptom severity (much better, a little better, no change, a little worse, much worse) could be used to find the range of changes in a multi-item COA that correspond to patients endorsing their change in symptom severity as "much better." (Considerations for the use of anchors are discussed in the next two sections.) Distribution-based methods (e.g., effect sizes, certain proportions of the standard deviation and/or standard error of measurement) do not directly consider the patient voice, and as such, are insufficient to serve as the sole basis for identifying an *MSD*. Distribution-based methods can provide helpful information about measurement variability. FDA is open to discussion about other well-justified methods developed for determining thresholds for *MSD* (e.g., Idio Scale Judgment; Cook et al. 2017).

a. Choice of anchor variables

FDA recommends that sponsors use multiple anchor measures to inform decisions about a plausible range of *MSD* values. Several factors should be considered when choosing anchor measures and, in the case of multiple anchor variables, when deciding how much weight to give an anchor when specifying *MSD* values:

- Ideally, the concept assessed by an anchor variable should match or be inclusive of the concept of interest being assessed by the COA-based endpoint. For example, a sponsor might propose a single item assessing the patient's global impression of severity for a symptom to use as an anchor variable to help interpret scores on a multi-item patient-reported outcome measure of severity for the same symptom. Sometimes it may not be possible to find an anchor that is a direct reflection of the patients' experiences related to the concept of interest measured by the COA-based endpoint. In such cases, sponsors can consider using multiple, less directly related anchors to aid in the interpretation of a meaningful difference in scores.

- An anchor should be plainly understood by respondents in the context of use. FDA recommends testing the proposed anchor item(s), including their response categories, in cognitive interviews.

- An anchor should have a well-justified definition for meaningful change or for meaningful increments. For example, consider the case of a single-item ordinal anchor to

[25] While it might be similar to the COA, an anchor variable would typically not be useful as the basis for the trial endpoint because it may be less sensitive than the COA and/or address a concept of interest that is broader or more specific than the concept of interest measured by the COA.

measure patients' perceptions of their symptom severity (e.g., with response options of none, mild, moderate, severe). Such an anchor might be used, for example, to help interpret scores from a multi-item COA intended to measure a symptom's severity. Qualitative data collected as part of cognitive interviews with patients could help to establish whether patients believe that the anchor variable's response options adequately represent meaningfully different experiences in their daily lives.

- Differences in COA scores should be related to differences documented by one or more anchors.[26] The stronger the relationship, the more confidence in translating differences in the anchor to differences in COA scores.

- Selected anchors should be assessed at comparable time points to the target COA. Sponsors should also ensure that, where applicable, the recall period of the anchor measure is consistent with the period covered by the COA-based endpoint.

- Sometimes sponsors wish to use a Global Impression of Change as an anchor, for example, a Patient Global Impression of Change (PGIC), in which patients report the direction and extent of change they have undergone between baseline and a follow-up time point using an ordinal categorical response scale. There should be evidence that the Global Impression of Change reflects the patient's/observer's/clinician's perception of the change they experienced (in the case of the patient) or observed (in the case of an observer or clinician). The usefulness of the Global Impression of Change as an anchor is reduced when there is excessive recall error and/or present state bias (i.e., the impression of change is influenced by the patient's status at follow-up more than the patient's actual change).

- Sometimes sponsors wish to use a Global Impression of Severity as an anchor, for example Patient Global Impression of Severity (PGIS), in which patients/observers/clinicians report the current or recent status of the severity or observation of symptoms or degree of functioning using a single ordinal response scale. Note that PGIS can be used to support either an *MSD* approach (by relating changes in the PGIS to changes in COA scores) or, as will be discussed in Section III.B.2, a meaningful score regions (MSRs) approach (by relating COA scores to their most likely PGIS response category).

In some situations, an acceptable anchor variable will not exist. When a suitable anchor cannot be found, sponsors can consider other methods to inform the choice of *MSD*, such as Idio Scale Judgment (Cook et al. 2017).

[26] Note that "differences in COA scores" is used here as a general term that includes differences that occur over time within a patient, i.e., changes in COA scores.

b. Analyses of anchors to inform choice of meaningful score difference

There are several options for relating differences in COA scores to anchor measures to arrive at *MSDs* (Coon and Cook 2018). Regardless of the analytic approach used, the following principles apply:

- Examine the distribution of the anchor scores or changes in anchor scores to ensure there is adequate variability for purposes of analysis. When changes in anchor scores are of interest, changes in the anchor scores should also be examined by baseline anchor score.

- Clearly describe the relationship between the COA score differences and the anchor (e.g., PGIC) or change in the anchor score (e.g., PGIS).

- Represent the distribution of COA difference scores corresponding to each response level of the anchor (e.g., PGIC) or each level of change in the anchor (e.g., PGIS). This presentation helps to inform a reasonable range of *MSD* estimates based on the heterogeneity among the patients studied.

- For ordinally-scaled anchors measured at two time points (e.g., PGIS), sponsors should first determine, based on evidence, what size changes in the anchor are regarded as meaningful (e.g., 1-category, 2-category). For each level of potentially meaningful change in the anchor (e.g., 1-category), sponsors should examine the distribution of COA difference scores separately by baseline anchor response. See Table 1 for an example table shell that could be used to determine for patients who experienced a 1-category improvement in the PGIS whether the COA change scores are distributed differently depending upon the patient's baseline PGIS category.

In Table 1, the lowest PGIS category of "None" is not shown because it is impossible for a patient with no severity to experience improvement in their PGIS.

Table 1. Sample Table Shell To Display the Distribution of COA Change-From-Baseline Scores for Patients With a 1-Category Improvement in Patient Global Impression of Severity.

PGIS at Baseline	N (%)	**Change in COA Score from Baseline to End of Study**				
		10^{th} Percentile	25^{th} Percentile	50^{th} Percentile	75^{th} Percentile	90^{th} Percentile
Mild						
Moderate						
Severe						
Very Severe						

- To select a range of thresholds to define *MSD*, sponsors should consider the following:[27]

  - Any choice of threshold *MSD* that attempts to distinguish between meaningful and non-meaningful differences will not correspond to some patients' experiences. That is, a difference below *MSD*, as measured, could be experienced as meaningful by some patients or a difference above *MSD*, as measured, could be experienced as not meaningful by some patients. Sponsors should consider and seek FDA input on how best to balance these two types of errors in the context of use. Note that this issue applies to any method used to derive thresholds, including anchor-based methods.

  - Generally, a wider range of thresholds should be selected when there is greater uncertainty about what patients would regard as an impactful difference. (Note that subsequent use of a wider range of thresholds to interpret a treatment effect will translate into correspondingly greater uncertainty about whether an obtained treatment effect is considered meaningful to patients.) A wider range of thresholds should be considered when any of the following are true:

    - There is a lower association between the COA difference scores and the anchor values, resulting in substantial overlap in the distributions of COA difference scores corresponding to different levels of the anchor scores (or differences between anchor scores). The greater the overlap, the less certainty there is that a given difference in COA score corresponds to a noticeable difference as indicated by the anchor. (See Coon and Cook 2018 for analytic approaches to examining overlap in distributions.)

    - Analyses of multiple anchor variables have generated different estimates of *MSD*. Note that in considering the range of *MSD*, threshold estimates from some anchors can be weighted more heavily than those estimates from other anchors based on the quality of the anchor (see III.B.1.a).

    - Analyses of the same anchor variable across multiple studies have generated different estimates of *MSD*.

    - There are several important prespecified patient subgroups, and analyses of the same anchor variable might generate different findings for different patient subtypes.

2. Interpreting in Terms of Meaningful Score Regions

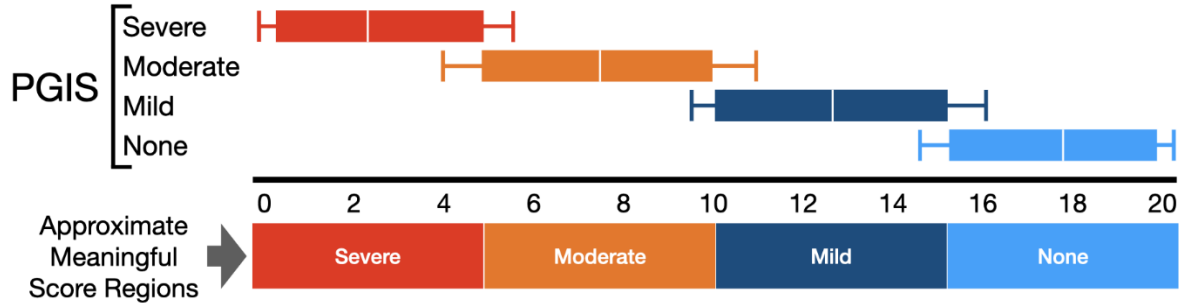
Another approach for interpreting the meaningfulness of treatment effect is to specify the meaning of individual COA scores so that it is easier to judge whether two or more scores (e.g., treatment group means at a prespecified time point) correspond to distinct health-related

[27] For a discussion of different methods for determining a threshold of meaningful score differences, see Coon and Cook 2018.

experiences of patients. For example, consider a measure of functioning that can generate scores from 0 to 20. Based on a study conducted with an independent sample of patients using the PGIS as an anchor, a figure can be constructed (Figure 1) to illustrate how different scores correspond to patients' global judgments of their functional impairment (none, mild, moderate, or severe). Assuming that the criteria for a strong anchor have been met (see III.B.1.a), the distributions of COA scores by PGIS response category could be examined to inform an approximate division of the COA score range into meaningful score regions (*MSRs*), as shown at the bottom of Figure 1. (Note that the figure shows an example in which the *MSRs* have equal widths; in other cases, the widths might differ.) In a later section (III.C), it is shown how *MSRs* could be used to help interpret a treatment effect on a COA-based endpoint.

In Figure 1, Box-and-whisker plots display the 25th (left edge of box), 50th (white line inside the box), and 75th (right edge of box) percentiles of the COA score distributions corresponding to each PGIS level. Whiskers indicate scores ± 1.5 interquartile range. *Approximate meaningful score regions* denote groups of scores that are thought to be similar to one another and different from other groups of scores in terms of the patient's experience of the symptom(s) measured by the COA.

Figure 1. Example of Approach for Interpreting COA Scores in Terms of Meaningful Score Regions Corresponding to Patient Global Impression of Severity (PGIS).



Different approaches to translate COA scores into their corresponding patient experiences may be appropriate if the approach is well justified within the context of use. Such approaches might include the following:

- Bookmarking or similar methods in which patients, caregivers, and/or clinicians make judgments to sort patient experiences into a small number of ordinal categories (e.g., none, mild, moderate, or severe) (Cook et al. 2019). By determining the COA scores corresponding to those patient experiences, it is possible to identify the COA score ranges or zones that correspond to the different ordinal levels.

- For COAs containing multiple items that are all thought to reflect the same underlying concept of interest, such as lower limb mobility, another way to facilitate interpretation of COA scores is to use one or more illustrative items from the COA measure to help

identify *MSRs*.[28] Essentially, this approach uses one or more of the COA's own items to serve as a kind of internal anchor variable.[29] If the illustrative item's response categories are easy to interpret in terms of patients' experiences, then this can be done by showing the predicted illustrative item responses for two or more COA scores. This allows a comparison of COA scores in terms of different ways the patient might feel or function as described by the illustrative item. For example, imagine a multi-item PRO measure of lower limb mobility with scores that range from 0 (poor mobility) to 100 (excellent mobility). Assume that the sponsor predefined *MSRs* based on data collected prior to the clinical trial by examining the relationship between scores on the PRO measure and responses to an individual item from the same measure that asks about difficulty walking up a flight of stairs. In this case, the response options for the individual item serve as approximate *MSRs* to guide interpretation of the expected scores in each treatment group. Suppose the mean scores for the randomized groups at a predefined follow-up time were 40 and 60. The *MSR* corresponding best to a score of 40 is "much difficulty" walking up a flight of stairs, compared to "little difficulty" for people whose score is 60. Items selected to serve as illustrative items should have item responses that are easily interpretable and are strongly associated with the COA score.

- For measures developed using Item Response Theory (IRT) (Chang and Reeve 2005), the meaning of different scores can be enhanced by using IRT item parameters to locate different items onto the measure's metric. For example, if a sponsor were using an IRT-based measure whose items assessed the level of assistance a patient needs to do different activities, the sponsor could show the activities that patients would be predicted to do "with no assistance" for different scores.

3. Additional Considerations for Justifying Meaningful Differences or Meaningful Score Regions

- FDA recommends that sponsors seek FDA input early regarding plans for determining *MSDs* (III.B.1) or *MSRs* (III.B.2). Ideally sponsors should evaluate and provide estimates of meaningful differences or scores prior to the start of the registration trial(s).

- When justifying a meaningful difference using transformed data, the sponsor should provide the threshold on the transformed and raw scales to aid in interpretation. For multi-item measures using a transformed scale, it is critical that the threshold *MSD* be at least equal to or greater than a one-category change for at least one item on the raw (untransformed) scale.

- For situations in which it is not feasible to obtain information to inform meaningful differences or scores before a registration trial (e.g., rare disease trials), sponsors can consider using exit interviews or surveys (refer to PFDD Guidance 2). Patients or their caregivers could be asked questions such as whether the patient experienced a change in

[28] This approach is known more generally as *content-based interpretation* (section 11.1.4 in Cappelleri et al. 2014).
[29] It is "internal" in the sense that the item is part of the COA and is used along with other items to generate a score for the COA.

their symptoms from baseline, whether the change was an improvement or worsening, and whether they believe the change in symptoms was meaningful (e.g., they can now walk around their house without assistance). The interviews should be conducted after the patients complete the main portion of the study to avoid any potential compromise to trial integrity. Note that this approach is susceptible to greater bias than other approaches and generally should only be used in trials in which patients and/or caregivers are unaware of their study group assignment. Sponsors who are considering conducting exit interviews or surveys should submit a study protocol and interview guide to FDA for review as early as possible, ideally prior to beginning the registration trial.[30]

- If sponsors wish to use data cited in the literature to propose *MSDs* or *MSRs*, sponsors should explain why it is reasonable to generalize the *MSDs* and *MSRs* from the literature to aid in interpreting the results of their registration trial. It is important to evaluate the comparability of context between the literature and the registration trial under consideration in terms of relevant factors such as disease, patient population, background standard of care, location, calendar time,[31] COA version,[32] endpoints, and length of follow-up.

C. Applying Information About Meaningful Score Differences or Meaningful Score Regions to Clinical Trial Data

Information about meaningful differences or scores can be used to help interpret the meaningfulness of treatment effects within a clinical trial. Determining whether a medical product produces an effect that is meaningful to patients involves careful consideration of multiple sources of information. This could include findings from multiple endpoints (e.g., primary and secondary endpoints), multiple anchors that inform a range of *MSDs* or *MSRs*, prespecified sensitivity analyses to supplement the main trial analysis of the COA-based endpoint, analyses to examine heterogeneity of treatment effect, and graphical and/or exploratory analyses to examine analytic assumptions or illustrate findings in alternative ways. Stakeholders should consider the strength of evidence to support decision making and the general considerations described in this section when creating justifications to support the interpretation of clinical trial data. In the broader picture of marketing authorization decisions, there are many factors to weigh simultaneously when making a decision about meaningfulness.

Sponsors should prespecify the method(s) used to interpret COA-based treatment effects and to convey the uncertainty around guides for score interpretation (e.g., estimates of *MSD or MSRs*) through describing a range of likely values, confidence intervals, or other representations of the uncertainty. The specific method of applying *MSDs* or *MSRs* will depend on the type of COA-based endpoint and the approaches taken to analyze the trial outcomes. The considerations and

[30] For a review of emerging qualitative methods for informing estimates of meaningful differences, see Staunton et al. 2019.

[31] Consider any changes relevant to the estimation of *MSDs* and *MSRs* that might have occurred since the time the study or studies in the literature were conducted.

[32] Note that a COA refers to any instructions, administration materials, content, formatting, and scoring rules associated with a COA.

examples in this section are meant to provide general suggestions for how to approach the interpretation of COA-based treatment effects.

Note that the roles of *MSD* or *MSRs* differ depending upon the type of endpoint. For endpoints based on continuous COA scores, the *MSD* or *MSRs* help to interpret the treatment effect. For this application, the sponsor can prespecify a range of *MSD* or *MSRs* that will be used to aid interpretation. For endpoints based on categorizing COA scores (e.g., a "responder" endpoint), the *MSD* or *MSRs* define the endpoint. In that case, the sponsor should prespecify a single threshold (for *MSD*) or set of thresholds (for *MSRs*) that will be used to define the endpoints.

1. Interpreting the Meaningfulness of Continuous COA-Based Endpoints

Different approaches can be used for interpreting treatment effects in terms of continuous COA-based endpoints depending upon whether *MSDs* or *MSRs* are used to aid in interpretation.

a. Meaningful score difference approach

An important consideration when applying *MSDs* to interpret a continuous COA-based endpoint is whether the estimates of *MSD* are relatively the same regardless of the patients' baseline COA scores. Sponsors who plan to interpret trial results in terms of *MSDs* should have already collected or cited evidence to evaluate this possibility.[33]

- *If there is evidence that MSD is relatively consistent over all baseline scores*: In this case, the difference between study arms may be compared to the value(s) of *MSD* to understand the meaningfulness of the treatment effect. For example, in a hypothetical clinical trial comparing a new product A to a current product B, scores (0-20) on a PRO measure of functioning were analyzed using an ANCOVA with baseline PRO functioning scores as the covariate. The primary prespecified group comparison was conducted at 12 weeks post-randomization. Figure 2 displays the treatment effect and 95% confidence interval.[34] Based on three different anchor-based analyses conducted using an independent sample of patients, the sponsor prespecified a range of *MSD* for the PRO functioning measure of 3 to 5 points. (The sponsor also conducted analyses to show that the value of *MSD* did not vary substantially by baseline COA score.) Because the x-axis reflects possible differences between scores on the PRO functioning measure, one can graph both the expected difference in scores between products A and B (i.e., the average treatment effect) and the range of *MSDs* thought to correspond to meaningfully different patient experiences. Figure 2 shows that values of the treatment effect that are

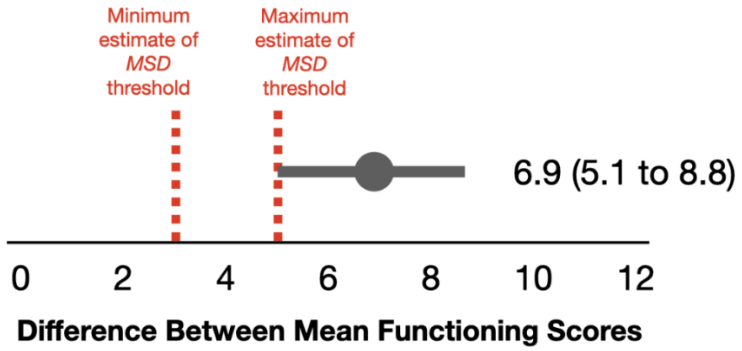
[33] Caution is needed when evaluating the potential baseline dependency of the *MSD*, because simple stratification on the baseline COA scores may lead to an erroneous finding of baseline dependency. There are other approaches that can be used (see Terluin B, Roos EM, Terwee CB, Thorlund JB, and LH Ingelsrud, 2021, Assessing Baseline Dependency of Anchor-Based Minimal Important Change (MIC): Don't Stratify on the Baseline Score! Qual Life Res, 30(10):2773-2782, doi:10.1007/s11136-021-02886-2).

[34] This treatment effect can be interpreted as a conditional treatment effect—that is, the treatment effect is assumed to be approximately constant across subgroups defined by the baseline PRO score in the ANCOVA model. In other words, this treatment effect is the difference in PRO score we would expect for the average patient. See FDA's draft guidance for industry *Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products* (May 2021). When final, this guidance will represent the FDA's current thinking on this topic.

consistent with the observed data (reflected by the 95% confidence interval) are above the maximum estimate of the threshold for *MSD*. This strongly suggests that the average treatment effect corresponds to a difference in experience that most patients would consider meaningful. In contrast, Figure 3 displays a scenario that does not clearly correspond to a meaningful overall difference due to treatment using the predefined MSDs, although a small portion of patients might experience a treatment effect that they regard as meaningful.

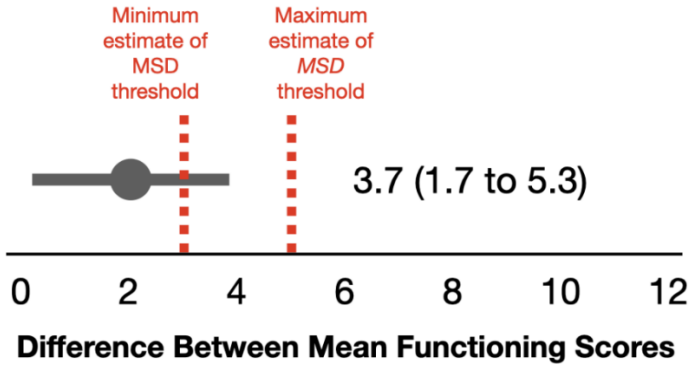
In Figure 2, dotted red lines indicate the minimum and maximum estimates of meaningful difference thresholds (D thresholds) obtained from anchor-based studies conducted independently of the registration trial. Differences greater than a threshold estimate are considered noticeably different by patients.

Figure 2. Estimated Difference in Adjusted Means (With 95% Confidence Interval) Between Products A and B on Functioning Measure Scores at Follow-Up Time Point Relative to Thresholds for Meaningful Score Differences



In Figure 3, dotted red lines indicate the minimum and maximum estimates of meaningful difference thresholds (D thresholds) obtained from anchor-based studies conducted independently of the registration trial. Differences greater than a threshold estimate are considered noticeably different by patients.

Figure 3. Estimated Difference in Adjusted Means (With 95% Confidence Interval) Between Products A and B on Functioning Measure Scores at Follow-up Time Point Relative to Thresholds for Meaningful Differences



- *If there is evidence that MSD varies substantially depending upon the patients' baseline scores*: This might occur if, for example, estimates of *MSD* ranged from 2 to 8 with larger values of *MSD* found for patients whose baseline COA scores reflected lower severity. In this case, if the treatment effect is larger than the largest estimate of *MSD* (e.g., *MSD* = 8 for patients who are least severe at baseline), this suggests that the treatment effect corresponds to a meaningful difference in patients' experiences. If the treatment effect is smaller than the largest estimate of *MSD*, it means that the treatment effect might be meaningful for only some or even none of the patients depending upon their baseline COA scores. To explore this, the sponsor could compare the treatment effect estimate to the estimates of *MSD* corresponding to each level of baseline COA score to better understand the meaningfulness of the treatment effect in patients across the range of baseline severity.

In addition to directly interpreting the estimate of treatment effect as described above, other analyses and displays may aid interpretation. If within-patient changes from baseline in the COA-based endpoint can be meaningfully estimated and interpreted from the trial data, sponsors can also plot the empirical probability density function (ePDF) or empirical cumulative distribution function (eCDF) of changes from baseline for each trial arm. The graphs should be annotated with a range of *MSD* values and the proportion of patients in each trial arm whose change-from-baseline exceeds one or more values of *MSD*. At times, other descriptive statistics by trial arm, such as the median and other quantiles of the change-from-baseline distributions, can provide additional relevant information.

These and other supplementary analyses should be interpreted in the context of the estimates of treatment effect overall and, if applicable, by prespecified patient subgroups. A judgement about the overall meaningfulness of the treatment effect could be made based on all the different analyses described in the example, along with data from complementary endpoints, any other clinical trials, and other factors that define the context of use.

b. Meaningful score regions approach

Figure 1 (presented earlier) illustrated how a study conducted with an independent sample of patients using the PGIS as an anchor informed a decision about approximate *MSRs*. These regions corresponded to patients' experiences of their health state as *none*, *mild*, *moderate*, or *severe*.

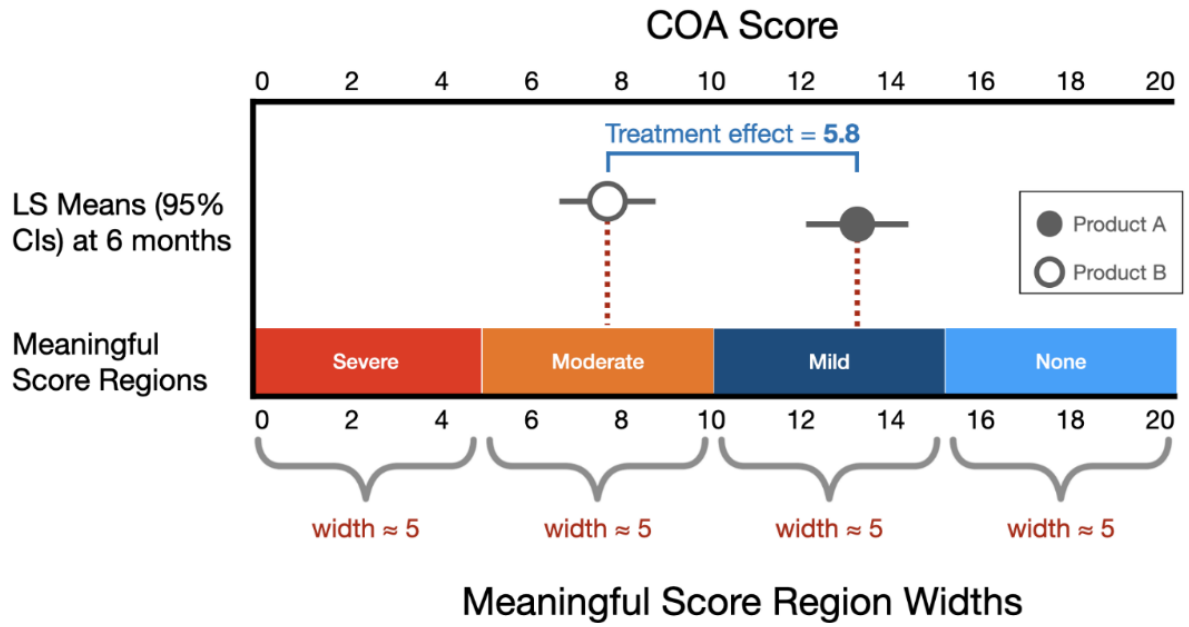
When examining the treatment effect in terms of *MSRs*, sponsors should predefine whether a difference of 1, 2, or more regions is required for patients to view the treatment effect as meaningful. The discussion that follows uses a 1 region difference, which would need to be supported by patient and/or caregiver input and which might not apply to other COAs and contexts of use.

An important consideration when applying the *MSRs* approach to interpret a continuous COA-based endpoint is whether the widths of the *MSRs* are relatively similar. For example, the widths of the regions in Figure 1 are all approximately 5 points. The following are general considerations regarding the width of the *MSRs* and the size of the treatment effect:

- *If there is evidence that the widths of the MSRs are relatively similar*: In this case, if the treatment effect is larger than the width of each of the *MSRs*, this suggests the treatment effect could be considered meaningful (i.e., because no matter where along the score range the treatment effect occurs, the average treatment effect will always correspond to a difference in score regions). This is illustrated in which the overall treatment effect is shown in terms of the adjusted means at the predefined follow-up time generated from an ANCOVA. However, if the average treatment effect is smaller than the common width of the *MSRs*, then additional analyses may be necessary to understand the nature of the treatment effect, such as exploring predicted COA scores at follow-up for each study arm over a range of baseline COA scores. This analysis may help identify which, if any, COA values at baseline are associated with a treatment effect that crosses two or more *MSRs*.

In Figure 4, dotted red lines are drawn to illustrate how adjusted means are mapped onto meaningful score regions derived using PGIS data.

Figure 4. Least Squares (LS) Means Scores (With 95% Confidence Interval) on Functioning Measure Scores at Follow-up Time Point for Products A and B Relative to Meaningful Regions of Scores Based on Patient Global Impression of Severity.



- *If there is evidence that the widths of the score regions are relatively different*: In this case, if the treatment effect is larger than the width of the widest score region, this suggests that the treatment effect reflects a meaningful difference to patients and/or caregivers. If the treatment effect is smaller than the widest score region, then the meaningfulness of the treatment effect may be different for different patients, depending upon their baseline status. This possibility could be explored as described in the prior bullet, by examining predicted COA scores at follow-up for each study arm over a range of baseline COA scores.

In addition to directly interpreting summaries of COA scores by treatment group, sponsors may also plot the ePDF or eCDF of COA scores separately by treatment group, annotating the graph with a guide for *MSRs* (e.g., as shown in the X axis of Figure 4). Such graphs might help to assess whether, for example, a small average treatment difference is driven by a small location shift in the entire curve or by a bigger shift in a small part of the curve. Sponsors may also compute, separately by treatment group, the proportion of patients with scores at follow-up that are greater (or less) than a specific score corresponding to the border between two *MSRs* (e.g., in the example used for Figure 4, scores less than 10 would reflect moderate to severe problems with functioning).

2. Interpreting the Meaningfulness of Ordinal and Dichotomous COA-Based Endpoints

When a COA-based endpoint is on an ordinal scale, interpreting effects in terms of meaningfulness to patients will depend upon the COA. Some measures produce an ordinal score consisting of a small number of categories that may have already been shown through cognitive interviews to be well understood and to reflect meaningfully distinct experiences of the patients (e.g., pain intensity rating of none, mild, moderate, severe). For these types of ordinal scales, no additional work may be needed to interpret the meaningfulness of the score, though additional analyses might need to be done to understand the nature of the treatment effect. In contrast, some measures might produce an ordinal score with many levels (e.g., 0 – 7) that may have been shown through cognitive interviews to be less interpretable in terms of patients' experiences. Additional work is recommended using the *MSRs* approach to understand which score ranges correspond to distinct experiences of patients.

Some endpoints are based on defining a state or status with respect to a COA score (see II.A.2.b). The status could be defined based on an *MSD* approach by classifying patients' changes from baseline (e.g., as "observed improvement," "observed worsening," "no change"). The endpoint could also be defined using a *MSRs* approach (e.g., patients scoring below some thresholds are classified as "symptoms resolved" and those scoring at or above the threshold are classified as "symptomatic"). For these situations, the sponsor should prespecify the threshold (in the case of *MSD*) or set of thresholds (in the case of *MSRs*) that will be used to define the endpoint.

IV. ADDITIONAL CONSIDERATIONS

A. Other Study Design Considerations

1. Masking[35]

Patients', clinicians', and/or caregivers' knowledge of treatment assignment (e.g., in single arm trials, open label trials, open-label treatment extension periods) is likely to influence how they report information on a PRO, ClinRO, or ObsRO measure, or how they engage with PerfO tasks (e.g., amount of encouragement given to patients when measuring walking distance), which will bias estimates of treatment effect. The protocol should specify to what extent masking will be maintained among the investigators, evaluators/raters, and reporters (e.g., clinicians, patients, caregivers).

2. Practice Effects

A practice effect (sometimes also called a learning effect) is any change that results from practice or repetition of particular tasks or activities including repeated exposure to an

[35] See footnote 20.

instrument. A simple example is taking a math test, which measures math ability. After completing the same test three times, a person's speed (and accuracy in answering) likely will improve because they recognize the questions and have 'learned' the test. While potentially an issue for any COA, practice effects may be of particular concern in studies utilizing PerfOs with within-subject designs in which repeated measurements are taken over time, (i.e., over the course of the study period; American Psychological Association 2018; Shadish et al. 2002).

Practice effects may be problematic for studies conducted to support a medical product regulatory application. If severe enough, practice effects could lead to improvements in the score of the assessment that might change the effective range of an assessment (e.g., if it creates a ceiling effect), potentially limiting the size of the observed treatment effect, which might impact the study's statistical power. Aside from this possibility, in a randomized, double-masked[36] trial, practice effects are unlikely to bias the difference of the outcomes between arms. For randomized trials that are not masked, differences might arise between trial arms in practice effects (e.g., due to differences in patient motivation or in how research staff interact with patients) and could impact group differences in the endpoint in a way that is not due to the treatment effect. For non-randomized trials, especially trials using external controls whose COA assessment schedule differs from treated patients, an apparent difference (or lack of difference) between trial arms may be due to practice effects and not due to any difference in the medical products.

Currently, approaches exist for attenuating, but not eliminating, practice effects (Jones 2015). In addition, no consensus on best practices for attenuating practice effects has yet been reached. Some general strategies for mitigating practice effects are summarized below. These strategies can be used in isolation but may be more effective when used in combination.

- **Consider available evidence on practice effects when identifying an instrument:** Some instruments may be more robust to practice effects than others. When selecting an instrument, one may wish to consider available evidence of the candidate instrument's robustness (or vulnerability) to practice effects. Such evidence can be obtained through, for example, a review of the literature and/or consulting the instrument's user manual or developer. If no evidence exists for a candidate measure, sponsors can conduct their own empirical study of potential practice effects.

- **Increase length of time (spacing) between assessments:** In general—and all else being equal—the magnitude of practice effects is expected to decrease as time between assessments increases (Shadish et al. 2002). Decisions regarding the length of time (spacing) to place between assessments should take into consideration how rapidly (or slowly) change in the underlying construct is expected to occur.

- **Increase the length and number of assessments for the run-in period:** In general, the magnitude of practice effects is largest at the beginning of a study and gradually levels off or decreases as the number of assessments increases. Having a long run-in period allows large practice effects to occur for the first few assessments until its magnitude

[36] Also referred to as "double-blind."

does not significantly increase such that the baseline and post-baseline scores are minimally affected by practice effects. Note that this strategy would not reduce any ceiling or floor effects caused by practice.

- **Use alternative forms (sometimes also referred to as parallel forms or equivalent forms):** Alternative forms are different versions of an instrument "that are considered interchangeable, in that they measure the same constructs in the same ways, are built to the same content and statistical specifications, and are administered under the same conditions using the same directions" (American Educational Research Association 2014).

3. Use of Assistive Devices

If a patient starts to use an assistive device after beginning the clinical trial, the interpretation of COA-based endpoints can be affected. Use of assistive devices may particularly impact PerfO assessment of mobility and can impact other types of COAs (e.g., use of a walker may impact both PerfO and PRO measures assessing physical functioning). For diseases where patients' underlying disease status is expected to change during the trial, with corresponding changes in the use and the type of assistive device, sponsors should consider the following:

- Some COAs address the use of assistive devices in the instructions or administration manual, detailing how the conduct of the assessment and scoring should occur when a patient is using an assistive device. If this is the case, sponsors should follow the directions for administering and scoring the chosen COA.

- When the COA does not explicitly address how to incorporate assistive devices into the assessment, then the sponsor should consider one of the following two strategies:[37]

  - If the use of the assistive device could be influenced by treatment and altering the need for the assistive device is one of the primary goals of treatment, then incorporate the information on the use of assistive device into the COA-based endpoint construction, as the use of an assistive device may reflect either an improvement or a deterioration in the patient's disease status.

  - If the use of the assistive device could be influenced by treatment and altering the need for the assistive device is not a primary goal of treatment, construct a supportive endpoint based on whether an assistive device is used.

- Case report forms for data collection should include information on whether an assistive device (and what type) was used during the test.

[37] These strategies are based on the estimand framework—namely the ways to address intercurrent events (i.e., things that happen after randomization that might affect the ability to observe or the interpretation of an endpoint). For additional details, see ICH E9(R1).

4. Considerations When Using a Nonrandomized Design, External Controls, or Nonconcurrent Control

Whenever possible, COA-based endpoints should be assessed in the context of randomized, controlled clinical trial designs. Sponsors considering COA-based endpoints in nonrandomized, external control, or nonconcurrent control (randomized groups but at different calendar times) trial designs should be aware of the significant potential for bias in estimating treatment effects:

- Depending on the study, the inability to effectively mask treatment assignment could cause group differences due to expectations of outcome held by patients, caregivers, clinicians, or research staff. To mitigate this risk, sponsors using these designs may consider assessing concepts of interest that require less subjective judgments (e.g., ability to do certain activities instead of perceived difficulty in doing activities). Though there might still be effects of patient expectation, sponsors could also use PerfO measures for which the patient's performance is rated by study personnel who are masked to treatment assignment or rated automatically by some device or computer.

- There might be differences in the measures used to assess the concept(s) of interest, method of COA administration, and/or the COA assessment frequency/schedule that could lead to differences between the groups that is unrelated to the effect of treatment. It is important to establish comparability of the COAs across the groups, to use well-defined and reliable COA-based endpoints in conjunction with standardized rater training and instructions for administration within each comparator arm and across comparator arms. Every effort should be made to ensure comparability in the assessment methods and timing of COA administration, together with the use of standardized data collection methods (e.g., standardized modes of administration).

- There might be preexisting differences between the groups that affect the estimate of treatment effect. (This potential source of bias is not unique to COA-based endpoints.)

These considerations apply to clinical trials as well as natural history studies,[38] disease registries, baseline-controlled trials, and trials with a more complicated sequential on-off-on (medical product-control-medical product) designs. Considerations for the various types of control groups are discussed at length in the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) guidance for industry *E10 Choice of Control Group and Related Issues in Clinical Trials* (May 2001).[39]

[38] See the draft guidance for industry *Rare Diseases: Natural History Studies for Drug Development* (March 2019), *Rare Diseases: Common Issues in Drug Development* (January 2019) and final guidance for industry *Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices* (August 2017).

[39] Available at the FDA guidance web page.

5. Analysis of Treatment Effects for Subgroups Based on Post-Baseline Events

If subgroups of a trial population are defined based on post-baseline events (e.g., patients who are alive and on treatment), interpretation of direct comparisons between treatment arms are likely to be misleading. By no longer reflecting the randomization intended to support a strong inference, the treatment arms will likely no longer be comparable due to differences in patient characteristics based on post-baseline events.

6. Computerized Adaptive Testing

One option for collecting scores from patients in clinical trials is to use computerized adaptive testing (CAT). This involves the use of an algorithm to iteratively select and administer items from a bank of items based on previous responses of the person being assessed. With each item that is answered, an updated estimate of the person's status on the concept of interest (e.g., symptom severity) is generated. That updated estimate is used by the CAT algorithm to select items that best match the current estimated severity and provide the most information for further estimation. The general goal of CAT is to provide individualized testing on a large scale by automatically selecting the most appropriate items for a person. However, generally the item selection is based on the likelihood that an item will be helpful in improving the estimate of the person's score, not on the relevance of the item content. (Note that special CATs can be constructed to ensure that items reflecting particular content are administered.) Thus, FDA recommends special considerations to assess whether CAT is appropriate for a given concept of interest and context of use.

Because a CAT is based on IRT modeling, sponsors who wish to use CAT should demonstrate that (1) the underlying IRT parameters are statistically sound and come from the population of interest; (2) the assumptions of the IRT model and CAT are tenable; and (3) the adaptive and scoring algorithms were correctly implemented.

Sponsors should consider the concept of interest and if the specific items have sufficient content coverage when using CAT. Hybrid CAT, where a small number of static items (i.e., those seen by all respondents) are administered along with the administration of items using the CAT algorithm, may be useful when CAT administration of items serves to supplement the static short form. When thoughtfully implemented, CAT or hybrid CAT may present advantages over static administrations, such as short forms.

In general, sponsors should consider whether administering items from an item bank via CAT will be more advantageous than administering a short form consisting of the same set of items to every patient in the trial. In some cases, CAT administration can bring statistical efficiency and help lower patient burden. It allows for tighter control of score reliability, while often reducing the number of items administered. However, depending on the concept of interest being measured and the range of severity in the target population, CAT may or may not provide a significant advantage over a short form in terms of precision, number of items recommended, and/or ceiling/floor effects. Research has shown that in some cases, CAT only provides benefits to measurement precision on the very high and low levels of severity when the sample is

representative of the full spectrum of severity (Choi et al. 2010; Rothrock et al. 2019; Amtmann et al. 2018). When weighing CAT versus short form in clinical study settings, sponsors should consider the make-up of their target population throughout the study, including at baseline, peak effect, and end of study. For specific populations with a limited range of severity, a short form can be created from the same item bank to target precise measurement over the range of severity expected in the study.

Sponsors should carefully consider the potential benefits and drawbacks to employing CAT in a clinical study. Discussion and alignment with the appropriate review division are strongly encouraged.

7. Minimizing Participant Burden

To demonstrate respect for the patients and/or caregivers who participate and maximize the quality and completeness of information collected in a clinical trial, sponsors should consider ways to minimize the burden of participation and increase the convenience and value of participation to patients and/or caregivers. Early engagement with patient communities (see PFDD Guidance 1) and the involvement of patient representatives in the development of a clinical trial can improve the patient-centeredness of trial procedures and assessments. With respect to COA-based endpoints, patient communities can provide input on the relevance, type, length, and frequency of COAs. Pilot testing of procedures for recruitment and assessment can also help minimize patient burden. A failure to evaluate and address potential issues with burden or fatigue can result in a trial with greater missing data, poorer quality data (e.g., when overly burdened participants quickly respond and select the first response to every item rather than carefully reading and considering their answer), and/or more dropout.

B. Formatting and Submission Considerations

Regardless of how patient experience data is collected in a given study, patient experience data collected and submitted to FDA to support a regulatory medical product application are subject to statutory and regulatory submission requirements that apply to the study data and submission type. Guidance documents that address data formatting and submission include, but are not limited to, the following:

- ICH guidance for industry *M8 Electronic Common Technical Document (eCTD) v4.0 DRAFT Implementation Guide v2.0; and eCTD Implementation Package DRAFT Specification for Submission Formats v2.0* (April 2015)

- Code of Federal Regulations, (CFR) Title 21, Chapter 1 (21 CFR Chapter 1)—with particular attention given to Parts 11, 21, 312.57, 312.62(b) and (c), and 812.140

- FDA draft guidance for industry *Use of Electronic Records and Electronic Signatures in Clinical Investigations Under 21 CFR Part 11–Questions and Answers* (June 2017)[40]

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

- FDA guidance for industry *Computerized Systems Used in Clinical Investigations* (May 2007)

- FDA guidance for industry *Electronic Source Data in Clinical Investigations* (September 2013)

- FDA guidance for industry *Providing Regulatory Submissions in Electronic Format—Standardized Study Data* (June 2021)

- FDA guidance for industry *Providing Regulatory Submissions in Electronic Format—Submissions Under Section 745A(a) of the Federal Food, Drug, and Cosmetic Act* (December 2014)

- FDA guidance for industry *Providing Regulatory Submissions in Electronic Format—Certain Human Pharmaceutical Product Applications and Related Submissions Using the eCTD Specifications* (February 2020)

- FDA Study Data Standards Resources[41] which includes links to FDA technical specifications documents and guidances for CDER, CBER, and CDRH including the *Study Data Technical Conformance Guide* and the *eCTD Technical Conformance Guide*. This resource and its documents are frequently updated.

Electronic devices used to administer COAs in studies conducted to support a regulatory medical product application can present special development, testing, and deployment considerations common to digital health technologies. For example, usability studies may be needed to assess study participants' ability to enter timely and accurate data. The following FDA guidances have more information about these considerations:

- FDA draft guidance for industry, FDA staff, and other stakeholders *Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments* (June 2022).[42]

- FDA draft guidance for industry and FDA staff *Contents of a Complete Submission for Threshold Analyses and Human Factors Submissions to Drug and Biologic Applications* (September 2018)[43]

- FDA draft guidance for industry *Comparative Analyses and Related Comparative Use Human Factors Studies for a Drug-Device Combination Product Submitted in an ANDA* (January 2017)[44]

[41] Available at https://www.fda.gov/industry/fda-resources-data-standards/study-data-standards-resources.
[42] When final, this guidance will represent the FDA's current thinking on this topic.
[43] Ibid.
[44] Ibid.

- FDA draft guidance for industry, investigators, and other stakeholders *Digital Health Technologies for Remote Data Acquisition in Clinical Investigations* (December 2021)[45]

- FDA guidance for industry and FDA staff *Applying Human Factors and Usability Engineering to Medical Devices* (February 2016)

- FDA draft guidance for industry and FDA staff *Human Factors Studies and Related Clinical Study Considerations in Combination Product Design and Development* (February 2016)[46]

- FDA guidances with digital health content[47]

Sponsors may also consult SPIRIT (Calvert et al. 2018, 2021) and CONSORT (Calvert et al. 2013), consensus documents that include an extensive, detailed discussion of PRO information that can be included in trial protocols and manuscripts to improve the completeness and clarity of reporting. Much of the discussion in SPIRIT and the CONSORT PRO extension is applicable to other types of COAs as well.

[45] Ibid.
[46] Ibid
[47] Available at https://www.fda.gov/medical-devices/digital-health-center-excellence/guidances-digital-health-content.

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.

Agresti, A, 2013, Categorical Data Analysis (3rd Ed), Hoboken, NJ: Wiley-Interscience.

American Educational Research Association, American Psychological Association, and The National Council on Measurement in Education, 2014, The Standards for Educational and Psychological Testing, Washington (DC): American Educational Research Association Publications.

American Psychological Association, 2018, APA Dictionary of Psychology, accessed July 16, 2019, https://dictionary.apa.org.

Amtmann, D, A Bamer, J Kim, F Bocell, H Chung, R Park, R Salem, and BJ Hafner, 2018, A Comparison of Computerized Adaptive Testing and Fixed-Length Short Forms for the Prosthetic Limb Users Survey of Mobility (PLUS-M™), Prosthet Orthot Int, 42(5):476-482.

Berry, DA and GD Ayers, 2006, Symmetrized Percent Change for Treatment Comparisons, The American Statistician, 60(1):27-31.

Calvert, M, J Blazeby, DG Altman, DA Revicki, D Moher, MD Brundage, and CONSORT PRO Group, 2013, Reporting of Patient-Reported Outcomes in Randomized Trials: the CONSORT PRO Extension, JAMA, 309(8):814-822.

Calvert, M, M King, R Mercieca-Bebber, O Aiyegbusi, D Kyte, A Slade, A Chan, E Basch, J Bell, A Bennett, V Bhatnagar, J Blazeby, A Bottomley, J Brown, M Brundage, L Campbell, JC Cappelleri, H Draper, AC Dueck, C Ells, L Frank, RM Golub, I Griebsch, K Haywood, A Hunn, B King-Kallimanis, L Martin, S Mitchell, T Morel, L Nelson, J Norquist, D O'Connor, M Palmer, D Patrick, G Price, A Regnault, A Retzer, D Revicki, J Scott, R Stephens, G Turner, A Valakas, G Velikova, M von Hildebrand, A Walker, and L Wenzel, 2021, SPIRIT-PRO Extension Explanation and Elaboration: Guidelines for Inclusion of Patient-Reported Outcomes in Protocols of Clinical Trials, BMJ Open 11(6):1-35.

Calvert, M, D Kyte, R Mercieca-Bebber, A Slade, A Chan, MT King, the SPIRIT-PRO Group, A Hunn, A Bottomley, A Regnault, A Chan, C Ells, D O'Connor, D Revicki, D Patrick, D Altman, E Basch, G Velikova, G Price, H Draper, J Blazeby, J Scott, J Coast, J Norquist, J Brown, K Haywood, LL Johnson, L Campbell, L Frank, M von Hildebrand, M Brundage, M Palmer, P Kluetz, R Stephens, RM Golub, S Mitchell, and T Groves, 2018, Guidelines for Inclusion of Patient-Reported Outcomes in Clinical Trial Protocols: The SPIRIT-PRO Extension, JAMA, 319(5):483-494.

Cappelleri, JC, KH Zou, AG Bushmakin, JMJ Alvir, D Alemayehu, and T Symonds, 2014, Patient-Reported Outcomes: Measurement, Implementation, and Interpretation, Boca Raton: CRC Press.

Chang, CH and BB Reeve, 2005, Item Response Theory and its Application to Patient-Reported Outcomes Measurement, Eval Health Prof, 28(3):264-282.

Choi, SW, SP Reise, PA Pilkonis, RD Hays, and D Cella, 2010, Efficiency of Static and Computer Adaptive Short Forms Compared to Full-Length Measures of Depressive Symptoms, Qual Life Res, 19(1):125-136.

Coon, CD and KF Cook, 2018, Moving From Significance to Real-World Meaning: Methods for Interpreting Change In Clinical Outcome Assessment Scores, Qual Life Res, 27(1):33-40.

Cook, KF, D Cella, and BB Reeve, 2019, PRO-Bookmarking to Estimate Clinical Thresholds for Patient-Reported Symptoms and Function, Medical Care, 57(5 Suppl 1):S13-S17.

Cook, KF, MA Kallen, CD Coon, D Victorson, and DM Miller, 2017, Idio Scale Judgment: Evaluation of a New Method for Estimating Responder Thresholds, Qual Life Res, 26(11):2961-2971.

Crosby, RD, RL Kolotkin, and GR Williams, 2003, Defining Clinically Meaningful Change in Health-Related Quality of Life, J Clin Epidemiol, 56(5):395-407.

Duke Margolis Center for Health Policy, 2017, Developing Personalized Clinical Outcome Assessments, accessed August 15, 2021, https://healthpolicy.duke.edu/sites/default/files/2020-03/discussion_guide_4_5_17.pdf.

FDA Draft Guidance for Industry, Benefit-Risk Assessment for New Drug and Biological Products, September 2021.

FDA Draft Guidance for Industry, Comparative Analyses and Related Comparative Use Human Factors Studies for a Drug-Device Combination Product Submitted in an ANDA, January 2017.

FDA Draft Guidance for Industry and FDA Staff, Contents of a Complete Submission for Threshold Analyses and Human Factors Submissions to Drug and Biologic Applications, September 2018.

FDA Draft Guidance for Industry, Investigators, and Other Stakeholders, Digital Health Technologies for Remote Data Acquisition in Clinical Investigations, December 2021.

FDA Draft Guidance for Industry and FDA Staff, Human Factors Studies and Related Clinical Study Considerations in Combination Product Design and Development, February 2016.

FDA Guidance for Industry, Multiple Endpoints in Clinical Trials, October 2022.

FDA Draft Guidance for Industry, Rare Diseases: Common Issues in Drug Development, January 2019.

FDA Draft Guidance for Industry, Rare Diseases: Natural History Studies for Drug Development, March 2019.

FDA Draft Guidance for Industry, Use of Electronic Records and Electronic Signatures in Clinical Investigations Under 21 CFR Part 11—Questions and Answers, June 2017.

FDA Guidance for Industry and Food and Drug Administration Staff, Applying Human Factors and Usability Engineering to Medical Devices, February 2016.

FDA Guidance for Industry, Computerized Systems Used in Clinical Investigations, May 2007.

FDA Guidance for Industry, Electronic Source Data in Clinical Investigations, September 2013.

FDA Guidance for Industry, Food and Drug Administration Staff, and other Stakeholders, Patient-Focused Drug Development: Collecting Comprehensive and Representative Input, June 2020.

FDA Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders, Patient-Focused Drug Development: Methods to Identify What Is Important to Patients, February 2022.

FDA Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders, Patient Preference Information—Voluntary Submission, Review in Premarket Approval Applications, Humanitarian Device Exemption Applications, Humanitarian Device Exemption Applications and De Novo Requests, and Inclusion in Decision Summaries and Device Labeling, August 2016.

FDA Guidance for Industry, Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims, December 2009.

FDA Guidance for Industry, Food and Drug Administration Staff and Other Stakeholders, Principles for Selecting, Developing, Modifying and Adapting Patient-Reported Outcome Instruments for Use in Medical Device Evaluation, January 2022.

FDA Guidance for Industry, Providing Regulatory Submissions in Electronic Format—Certain Human Pharmaceutical Product Applications and Related Submissions Using the eCTD Specifications, February 2020.

FDA Guidance for Industry, Providing Regulatory Submissions in Electronic Format—Standardized Study Data, June 2021.

FDA Guidance for Industry, Providing Regulatory Submissions in Electronic Format—Submissions Under Section 745A(a) of the Federal Food, Drug, and Cosmetic Act, December 2014.

FDA Guidance for Industry and Food and Drug Administration Staff, Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices, August 2017.

Gawlicki, MC, SM McKown, MJ Talbert, and BA Brandt, 2014, Application of Bother in Patient Reported Instruments Across Cultures, Health Qual Life Outcomes, 12(18):1-7.

Goldberg, TE, PD Harvey, KA Wesnes, PJ Snyder, and LS Schneider, 2015, Practice Effects Due to Serial Cognitive Assessment: Implications for Preclinical Alzheimer's Disease Randomized Controlled Trials, Alzheimer's Dement (Amst) 1(1):103-111.

Guyatt, GH, GR Norman, EF Juniper, and LE Griffith, 2002, A Critical Look at Transition Ratings, J Clin Epidemiol, 55(9):900-908.

Harrell, F, 2015, Regression Modeling Strategies (2nd Ed.), New York: Springer.

ICH Guidance for Industry, E9(R1) Statistical Principles for Clinical Trials: Addendum on Estimands and Sensitivity Analysis in Clinical Trials, May 2021.

ICH Guidance for Industry, E10 Choice of Control Group and Related Issues in Clinical Trials, May 2001.

ICH Guidance for Industry, M8 Electronic Common Technical Document (eCTD) v4.0 DRAFT Implementation Guide v2.0; and eCTD Implementation Package DRAFT Specification for Submission Formats v2.0, April 2015.

Jones RN, 2015, Practice and Retest Effects in Longitudinal Studies of Cognitive Functioning, Alzheimer's Dement (Amst), 1(1):101-102.

Krasny-Pacini, A, J Evans, M Moore Sohlberg, and M Chevignard, 2016, Proposed Criteria for Appraising Goal Attainment Scales Used as Outcome Measures in Rehabilitation Research, Arch Phys Med Rehab, 97(1):157-170.

Lachin JM, 1999, Worst-Rank Score Analysis With Informatively Missing Observations in Clinical Trials, Control Clin Trials, 20(5):408-422.

Norman, GR, P Stratford, and G Regehr, 1997, Methodological Problems in the Retrospective Computation of Responsiveness to Change: The Lesson of Cronbach, J Clin Epidemiol, 50(8):869-879.

Rothrock, NE, AJ Kaat, MS Vrahas, RV O'Toole, SK Buono, S Morrison, and RC Gershon, 2019, Validation of PROMIS Physical Function Instruments in Patients With an Orthopaedic Trauma to a Lower Extremity, J Orthop Trauma, 33(8):377-383.

Senn, S, 2007, Statistical Issues in Drug Development (2nd Edition), Chichester (England): John Wiley & Sons.

Shadish, WR, TD Cook, and DT Campbell DT, 2002, Experimental and Quasi-Experimental Designs for Generalized Causal Inference, Accessed August 2, 2021, https://www.alnap.org/system/files/content/resource/files/main/147.pdf.

Shiffman, S, AA Stone, and MR Hufford, 2008, Ecological Momentary Assessment, Annu Rev Clin Psychol, 4(1):1-32.

Song, MK, and SE Ward, 2015, Assessment Effects in Educational and Psychosocial Intervention Trials: An Important But Often-Overlooked Problem, Res Nurs Health, 38(3):241-247.

Staunton, H, T Willgoss, L Nelsen, C Burbridge, K Sully, D Rofail, and R Arbuckle, 2019, An Overview of Using Qualitative Techniques to Explore and Define Estimates of Clinically Important Change on Clinical Outcome Assessments, J Patient Rep Outcomes, 3(1):16.

Stone, AA, JE Broderick JE, Goldman RE, Junghaenel DU, A Bolton, M May, and S Schneider, 2021, I. Indices of Pain Intensity Derived from Ecological Momentary Assessments: Rationale and Stakeholder Preferences, J Pain, 22(4):359-370.

Terluin B, Roos EM, Terwee CB, Thorlund JB, Ingelsrud LH, 2021, Assessing Baseline Dependency of Anchor-Based Minimal Important Change (MIC): Don't Stratify on the Baseline Score! Qual Life Res, 30(10):2773-2782, doi:10.1007/s11136-021-02886-2.

Uryniak, T, ISF Chan, VV Fedorov, Q Jiang, L Oppenheimer, SM Snappin, CH Teng, and J Zhang, 2011, Responder Analyses—A PhRMA Position Paper, Statistics in Biopharmaceutical Research, 3(3):476-487.

Wilson, IB, and PD Cleary, 1995, Linking Clinical Variables With Health-Related Quality of Life, JAMA, 273(1):59-65.

Vickers AJ, 2001, The Use of Percentage Change From Baseline as an Outcome in a Controlled Trial is Statistically Inefficient: A Simulation Study, BMC Med Res Methodol, 1(1):6-6.