U.S. FOOD & DRUG ADMINISTRATION

PATIENT-FOCUSED DRUG DEVELOPMENT
GUIDANCE PUBLIC WORKSHOP

Incorporating Clinical Outcome Assessments into Endpoints for Regulatory Decision-Making

Workshop Date: December 6, 2019

Discussion Document for Patient-Focused Drug Development Public Workshop on Guidance 4:

INCORPORATING CLINICAL OUTCOME ASSESSMENTS INTO ENDPOINTS FOR REGULATORY DECISION-MAKING

TABLE OF CONTENTS

I. INTRODUCTION 1
A. Guidance Series 1
B. Document Summary 2
II. ESTIMAND FRAMEWORK OVERVIEW 3
A. COA Research Objective: Foundation for Your Work 4
B. Target Study Population: In Whom Are You Going to Do the Research and Which Subject Records Are in the Analysis? 5
C. Endpoint of Interest: What Are You Testing or Measuring in the Target Study Population? 6
1. Endpoint Definition(s) 6
2. Pooling Different Tools and/or Different Concepts to Construct the Endpoint 6
a. Correlation of Subcomponents and Effect on Power and Type 1 Error 6
b. Multidomain Responder Index 7
c. Personalized Endpoints 8
d. Pooling Scores Across Reporters 8
e. Pooling Across Delivery Modes (Same Tool, Same Reporter) 8
3. Timing of Assessments 9
4. Defining Improvement and Worsening 10
5. Clinical Trial Duration and COA-Based Endpoints 10
D. Intercurrent Events: What Can Affect Your Measurement's Interpretation? 11
1. Use of Assistive Devices, Concomitant Medications, and Other Therapies 11
2. Impact of Disease/Condition Progression, Treatment, and Potential Intercurrent Events 12
3. Practice Effects 12
4. Participant Burden 14
5. Mode of Administration 15
6. Missing Data and Event-Driven COA Reporting 15
7. Missing Scale-Level Data 15
E. Population-Level Summary: What Is the Final Way All Data Are Summarized and Analyzed? 16
1. Landmark Analysis 16
2. Analyzing Ordinal Data 16
3. Time-to-Event Analysis 16
4. Responder Analyses and Percent Change From Baseline 17
III. MEANINGFUL WITHIN-PATIENT CHANGE 18
A. Anchor-Based Methods to Establish Meaningful Within-Patient Change 19
B. Using Empirical Cumulative Distribution Function and Probability Density Function Curves to Supplement Anchor-Based Methods 20
C. Other Methods 22
1. Potentially Useful Emerging Methods 22
2. Distribution-Based Methods 22

3. Receiver Operator Characteristic Curve Analysis 22
D. Applying Within-Patient Change to Clinical Trial Data 22
IV. ADDITIONAL CONSIDERATIONS 25
A. Other Study Design Considerations 26
B. Formatting and Submission Considerations 27
APPENDIX 1: CASE STUDY OF ESTIMAND FRAMEWORK 29
A. Example Research Objective 29
1. Define COA Scientific Research Question A Priori 30
2. Define Target Study Population Based on the Research Question A Priori 30
3. Define Endpoint of Interest Based on the Research Question A Priori 31
4. Address Intercurrent Events in Alignment with the Research Question 32
5. Define Population-Level Summary Based on Research Question A Priori 33
6. Prespecify Statistical Analysis Plan 34
B. Summary of Decisions Made in This Case Study 35
APPENDIX 2: EXAMPLE FROM GENE THERAPY 36
APPENDIX 3: REFERENCES 42
APPENDIX 4: GLOSSARY 44

TABLE OF FIGURES

Figure 1: Attributes of an Estimand Placed in Context 4
Figure 2: Example of Empirical Cumulative Distribution Function Curves of Change in COA Score from Baseline to Primary Time Point by Change in PGIS Score 21
Figure 3: Example of Density Function Curves of Change in COA Score from Baseline to Primary Time Point by Change in PGIS Score 21
Figure 4: An eCDF Curve by Treatment Arm Showing Consistent Separation Between Two Treatment Arms 23
Figure 5: An eCDF Curve Where Treatment Effect Is Not in Range Considered Clinically Meaningful by Patients 24
Figure 6: MLMT Scores in Phase 3 Trial 40

TABLE OF TABLES

Table 1: Considerations of Defining a COA Target Study Population 5
Table 2: Considerations When Defining a COA-Based Endpoint 31
Table 3: Considerations When Addressing Intercurrent Events 33
Table 4: Considerations When Defining a COA Population-Level Summary 33
Table 5: Summary of Estimand Decisions Made 35
Table 6: MLMT Illuminance Level, Score Code, and Real-World Examples 37

I. INTRODUCTION

A. Guidance Series

The Food and Drug Administration (FDA) recognizes the need to obtain meaningful ***patient experience data***[1] to understand patients' experience with their disease and its treatment. This can help inform development of ***endpoint*** measures to assess ***clinical outcomes*** of importance to ***patients*** and ***caregivers*** in medical product development. To ensure a ***patient-focused*** approach to medical product[2] development and regulation, FDA is developing guidance on methods to identify what matters most to patients for measurement in clinical trials; specifically, how to design and implement studies to capture the patient's voice in a robust manner. FDA created this Discussion Document to facilitate discussions at the December 6, 2019, public meeting that will inform FDA's development of a ***patient-focused drug development*** (PFDD)[3] guidance on incorporating ***clinical outcome assessments*** (COAs) into endpoints for regulatory decision-making.

This public workshop will inform the development of the fourth in a series of four methodological PFDD guidance documents[4] that FDA is developing to describe in a stepwise manner how stakeholders (patients, researchers, medical product developers, and others) can collect and submit patient experience data and other relevant information from patients and caregivers for medical product development and regulatory decision-making. The topics that each guidance document will address are:

- Methods to collect patient experience data that are accurate and representative of the intended patient population (Guidance 1)[5]

[1] The Glossary defines many of the terms used in this Discussion Document. Words or phrases found in the Glossary appear in bold italics at first mention.

[2] A drug, biological product, or medical device.

[3] See https://www.fda.gov/Drugs/NewsEvents/ucm607276.htm.

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

[5] See draft guidance for industry, FDA staff, and other stakeholders *Patient-Focused Drug Development: Collecting Comprehensive and Representative Input* (June 2018). When final, this guidance will represent FDA's current thinking on this topic. For the most recent version of a guidance, check the FDA guidance web page at https://www.fda.gov/RegulatoryInformation/Guidances/default.htm.

- Approaches to identify what is most important to patients with respect to their experience as it relates to ***disease burden*** and ***treatment burden*** (Guidance 2) [6]
- Approaches to identify and develop methods to measure impacts in clinical trials (Guidance 3)
- Methods, standards, and technologies to collect and analyze COA data for regulatory decision-making (Guidance 4)

All documents in the series encourage stakeholders to obtain feedback from FDA during the study and trial development period when considering collection of patient experience data. FDA encourages engagement of broader disciplines during clinical development (e.g., qualitative researchers, survey methodologists, statisticians, psychometricians, ***patient preference*** researchers, data managers) when designing and implementing studies because the logistics in some cases can be daunting for a seemingly simple piece of patient data to address a simple research objective.

B. Document Summary

The purpose of this Discussion Document is to help stakeholders understand what FDA considers when a COA in a ***clinical study*** will be used to eventually support medical product regulatory decision-making.

The document first lays out a framework that aims to align the clinical study objective with the study design, endpoint, and analysis to improve study planning and the interpretation of analyses. Several examples are provided to help illustrate the framework.

The document then describes methods to aid in the interpretation of study results to evaluate what constitutes a meaningful within-patient change (i.e., improvement and deterioration from the patients' perspective) in the concepts assessed by COAs. This information is important because statistical significance can be achieved for small differences between comparator groups, but this finding does not indicate whether individual patients have experienced meaningful ***clinical benefit***.

A list of considerations when developing an endpoint from a COA is included in Section IV of this Discussion Document.

[6] See draft guidance for industry, FDA staff, and other stakeholders *Patient-Focused Drug Development: Methods to Identify What Is Important to Patients* https://www.fda.gov/media/131230/download. When final, this guidance will represent FDA's current thinking on this topic.

II. ESTIMAND FRAMEWORK OVERVIEW

Section Summary

An ***estimand*** is a quantity used to define a treatment effect in a clinical study. The estimand framework aims to align the clinical study objective with the study design, endpoint, and analysis to improve study planning and the interpretation of analyses. The attributes of an estimand include specifically defining:

- Who is the target population for the study?
- What is the endpoint (e.g., what variables will be used including which time points)?
- How will events precluding observation or affecting interpretation be accounted for in the analyses, e.g., dropouts, use of rescue medication, not following prescribed regimen?
- What is the population level summary (e.g., comparing means, hazard ratios)?

Decisions for all the attributes, implicitly or explicitly, are currently present in every data analysis that is performed. The choices made strongly impact interpretation of the analysis, power, and data collected.

Technical Summary: Key Messages in This Section

To develop endpoints from COAs, a fundamental issue must first be addressed: What is the clinical question or research objective that the clinical study should be designed to answer? The estimand framework based on International Council on Harmonisation (ICH) E9(R1) aims to improve clinical studies by putting the focus on a set of attributes to ensure they align with the study research objectives. This section discusses four attributes:

Target Study Population

- Patients who are targeted by the scientific question; who will be included in the analysis
- A different population may be appropriate for each scientific question

Endpoint of Interest

- Outcome obtained for each patient that will be statistically analyzed to address the scientific question; this may include data from multiple variables
- ***Research protocols*** should define the ***concept***, COA instrument, ***score***/summary score, type of endpoint, and thresholds/estimates for clinical interpretation

Intercurrent Events

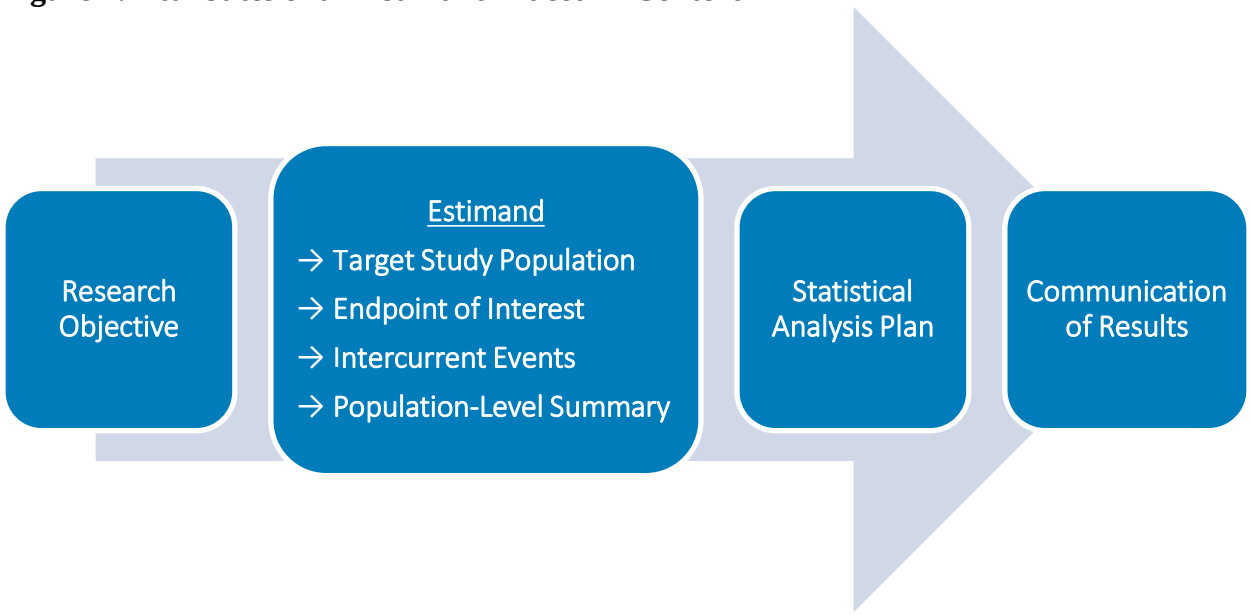
- Events that occur after randomization/treatment initiation/or study start that could preclude observation of the variable or affect its interpretation
    - An example of an intercurrent event: taking subsequent therapy beyond treatment discontinuation with an endpoint of physical function measured at a time point several weeks after treatment discontinuation
    - For nonrandomized trials or trials borrowing data, intercurrent events could occur at any time a subject is considered ‘on study’
- Protocols should specify intercurrent events and how they will be accounted for in analyses to address the scientific question of interest

Population-Level Summary

- Basis for comparison between treatment arms, treatment conditions, other groups, or otherwise summarizes information
    - Examples include a) mean physical function score at baseline for everyone in an ***observational research*** study and b) difference compared to control of a new medical product's median time to pain resolution

The attributes listed above should be clearly defined prior to developing a protocol and included in both the protocol and Statistical Analysis Plan (SAP). They will determine the data collected, procedures, and other sections of the protocol beyond statistical methods. The attributes also drive the SAP and communication of trial results, as highlighted in Figure 1.

Figure 1: Attributes of an Estimand Placed in Context



A. COA Research Objective: Foundation for Your Work

The essence of clinical research is to ask important questions and answer them with appropriate studies (ICH E8(R1)). The research objective should be clearly and explicitly stated. To develop the objective, both the natural history of the disease and the treatment goal for the intended product must be considered. For example, the choice of an endpoint will likely be very different between a product intended to treat an acute disease, where the ***symptoms*** of many patients will likely resolve within several weeks, versus a product intended to be used for patients living with a chronic disease. Even for a chronic disease, endpoint selection could vary depending on whether the disease is degenerative/progressive, relapsing and remitting, episodic, or relatively stable. Heterogeneity of symptoms or functional status of patients with the disease is also a crucial issue. As an example, relating to the intended goal of the treatment, a product intended to cure a disease is likely to have a different research objective from a product designed to decrease the symptom severity of a chronic disease.

B. Target Study Population: In Whom Are You Going to Do the Research and Which Subject Records Are in the Analysis?

The target study population used to address a COA research objective (the COA 'analysis population') may vary based on the COA-derived endpoint and scientific research question. There may be multiple COA analysis populations in a single trial. The COA analysis population(s) should be defined *a priori* in the protocol and SAP, with clear justification made for each COA analysis population. The choice of COA analysis population will affect the COA-related estimand and interpretation of patient experience.

Table 1 presents COA target study population examples. There may be other target study populations of interest depending on your research objective.

Table 1: Considerations of Defining a COA Target Study Population

Target Study Population (Examples)	

• ITT: All patients randomized according to assigned treatment arm, regardless of adherence
• Safety: All patients who received at least one dose of product, regardless of randomization
	• Analysis populations are often defined based on their availability of COA data
	  • All patients who are eligible for the COA
	  • Completed the COA at baseline
	  • Completed baseline and at least one postbaseline assessment
	  • COA data are available at any trial timepoint

Abbreviations: ITT = intent-to-treat; COA = clinical outcome assessment

For sponsors considering an effectiveness claim from a COA-derived endpoint in a randomized trial, the intent-to-treat (ITT) population generally should be used to preserve the benefits of randomization. Justification should be provided if treatment comparisons are made using a COA analysis population different from the ITT population. Any justification should incorporate discussion of trial blinding[7] procedures and their potential impact on data interpretation. If the COA objective is safety or tolerability, including patients who received at least one dose of the investigational product, regardless of randomization, may be more appropriate. Consider how interpretation of the COA-derived endpoint changes if all patients in a trial are not eligible for the COA. For example, ***generalizability*** of the results may be narrowed if some patients do not have access to a COA because it is unavailable in a language in which they are fluent. Additionally, depending on the trial protocol, eligibility to complete a COA may change over time.

Every effort should be made to have high completion rates throughout the study. At baseline, this is important otherwise all postbaseline assessments will be difficult to put into context without a reference variable. Because there is the potential for patients to have missing assessments, sponsors should clearly specify in the SAP how missing observations will be dealt with for clear

[7] Blinding is sometimes referred to as "masking"; for purposes of this document, we will use blinding.

interpretation. Removing subjects without a baseline measurement is common but depending on the research question it may not be the better option.

C. Endpoint of Interest: What Are You Testing or Measuring in the Target Study Population?

1. Endpoint Definition(s)

An endpoint is a precisely defined variable intended to reflect an outcome of interest that is statistically analyzed to address a specific research question. An endpoint definition typically specifies the type of assessments made, the timing of those assessments, the tools used, and possibly other details, as applicable, such as how multiple assessments within an individual are to be combined. ***Measurement properties*** remain crucial in the context of developing useful endpoints, as endpoints (as well as COAs) should be understood as imperfectly measuring concepts. Hence, assessment of an endpoint's ***reliability***, ***content validity, construct validity,*** as well as ***ability to detect change*** are important (refer to FDA PFDD G3 Public Workshop Discussion Document for details). Within the protocol, the specific COA concept(s) should be assessed by ***fit-for-purpose*** COA(s) and should be incorporated into a corresponding clinical trial objective or hypothesis and reflected in the endpoint definition and positioning in the testing hierarchy.

A ***multidomain*** COA may successfully support claims based on one or a subset of the ***domains*** measured if an analysis plan prespecifies (1) the domains that will be targeted for supporting endpoints and (2) the method of analysis that will adjust for the multiplicity of tests for the specific claim. The use of domain subsets to support clinical trial endpoints assumes the COA was adequately developed and validated to measure the subset of domains independently from the other domains. A complex, multidomain claim cannot be substantiated by instruments that do not adequately measure the individual components of the domain.

2. Pooling Different Tools and/or Different Concepts to Construct the Endpoint

a. Correlation of Subcomponents and Effect on Power and Type 1 Error

Since most diseases have more than one relevant clinical outcome, trials can be designed to examine the effect of a medical product on more than one endpoint (i.e., multiple endpoints). For example, a COA with multiple domains may be used in a clinical trial to assess the most relevant and meaningful clinical outcomes (i.e., each domain corresponds to one clinical outcome) to patients. In such a case, a multiple endpoint approach would be of clinical interest, specifically a ***multicomponent endpoint*** approach (refer to FDA draft guidance *Multiple Endpoints in Clinical Trials* (January 2017)).

Other analytical methods, such as global tests, could potentially be used to pool scores from different tools of a similar type, e.g., ***patient-reported outcomes*** (PROs). The use of these methods should be discussed with the FDA.

Some researchers have considered combining different scores from different measurement tools that evaluate different parts of the latent construct to create a new endpoint using item response theory and other methods to take in to account the different potential dimensions of the COAs. FDA is open to discussion of well-defined and reliable endpoints.

For a COA composed of multiple domains, with each domain measured using either an ordinal or a continuous ***response scale***, a within-patient combination (e.g., sum or average) of all the individual domain (i.e., component) scores to calculate a single overall rating creates a type of multicomponent endpoint. Other types of multicomponent endpoints may include a dichotomous (event) endpoint corresponding to an individual patient achieving prespecified criteria on each individual component. Careful considerations should be made regarding the choice of individual components and whether all components will have reasonably similar clinical importance or whether the algorithm combining the scores uses differential weighting and how those weights are determined. Since multicomponent endpoints are constructed as a single endpoint, no multiplicity adjustment is necessary to control Type I error. In addition, multicomponent endpoints may provide gains in efficiency if different components are not that highly correlated. Regardless of how a multicomponent endpoint is constructed, the choice of the endpoint needs to be clinically relevant and interpretable.

b. Multidomain Responder Index

For some rare diseases with heterogeneous patient populations and variable disease manifestations, it may be challenging to assess a single concept of interest across all patients. Stakeholders occasionally propose combining multiple measurement concepts (e.g., a variety of individual COA-based endpoints) into a single dichotomous (event) endpoint. While FDA regulations allow for flexibility and judgement in considering multicomponent endpoints, the selection of these endpoints faces similar challenges as those described under responder endpoints—responder analyses (refer to Section II.E.4). In addition, the choice of the individual components relies on the requirement that all components are of reasonably similar clinical importance (*Multiple Endpoints in Clinical Trials* (January 2017)). An example dichotomous (event) endpoint is the multidomain responder index (MDRI) approach, which thus far has not been demonstrated as a viable approach based on evidence submitted to FDA.

In general, the MDRI approach combines multiple individual domains or endpoints with a prespecified responder threshold for each endpoint. Various methods have been proposed to construct an MDRI endpoint. For example, each domain score is presented as +1 for improvement, 0 for no change, and -1 for decline, and an overall MDRI response for a patient is defined based on the individual scores (e.g., if any domain shows improvement). It is important to note that successful creation of an MDRI requires clearly defined and clinically relevant endpoints with appropriate responder thresholds (i.e., what constitutes a clinically meaningful within-patient change score) established *a priori* for those endpoints. In practice, these responder thresholds are often hard to establish, especially for rare diseases with small patient populations and limited natural history data. Additionally, defining an endpoint score as +1, 0, or -1 relies on the assumption that the degree of improvement and deterioration in a concept of interest is symmetric, which often is not a valid assumption. Another important consideration for the MDRI

approach is the amount of missing data for each domain, component, or individual endpoint of the MDRI. Large amounts of missing data will impede the interpretation of the endpoint results.

c. Personalized Endpoints

Similar concerns exist with personalized or individualized endpoints, which often are analyzed descriptively as exploratory endpoints. 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. The same set of outcome assessments should be assessed for all patients, regardless of their own personalized endpoint, to allow for an assessment of any new or worsening symptoms and/or functional limitation(s) during the trial duration. Certain outcome assessments may not be applicable to all trial patients. However, if an outcome is not assessed in a patient at a given time point, the reason for the assessment not being performed should be noted, included in the analysis data set, and used as part of the analysis.

d. Pooling Scores Across Reporters

To evaluate the treatment benefit of a medical product, sometimes it may be necessary to use different types of COAs to assess the same construct(s) in the same clinical trial (e.g., in a pediatric trial in which a PRO measure is used for older children who can reliably self-report and an ***observer-reported outcome*** (ObsRO) measure is used by caregivers to report ***signs*** and behaviors of younger children who are unable to reliably self-report). In general, scores generated by different types of COAs, (i.e., PROs, ObsROs, ***clinician-reported outcomes*** (ClinROs), and ***performance outcomes*** (PerfOs)), cannot be pooled to form a single clinical trial endpoint, even if they are developed to assess the same construct(s), in analyses submitted to FDA to support a medical product application. Because these different types of COAs are developed for different contexts of use (e.g., PRO measures to report direct experiences of symptoms by the patients themselves and ObsRO measures to report observable signs and behaviors of the patients by their caregivers), they are distinct outcome assessments, and it is therefore inappropriate to pool the resulting sets of scores.

Scores generated by different types of COAs should be analyzed separately with—where feasible—enough ***reporters*** included in each group to support any subsequent inferences to the ***target population***. Simply because each tool has the same score range (e.g., 0 to 10) does not mean data can be pooled.

e. Pooling Across Delivery Modes (Same Tool, Same Reporter)

Scores generated by the same tool administered ("delivered") via different modes (e.g., interactive voice response; interview; paper-based; electronic device) may be pooled under very specific and limited conditions. Although scores yielded by different modes are generally considered to be comparable when there is no difference between modes in terms of the wording of item stems and response options, item formats, the appearance and usage of graphics or other visuals, or order of the items (see FDA PFDD G3 Public Workshop Discussion Document for

further discussion), administering a tool using more than one mode or method per study can introduce noise (i.e., construct-irrelevant variance in COA score) that may not be completely random and may make it more difficult to discern treatment effects.

3. Timing of Assessments

Clinical trials using COAs should include a schedule of COA administration as part of the overall study assessment schedule in the protocol. The timing of assessments plays a vital role in gaining reliable and meaningful information on the concept(s) of interest and should be selected carefully and supported by adequate rationale for the choice of assessment time points. The COA schedule should correspond directly with the natural course of the disease or condition (i.e., acute, chronic, or episodic), research questions to be addressed, trial duration, disease stage of the target patient population, and current treatment of patients, and be administered within the expected time frame for observing changes in the concept(s) of interest. Other important considerations for determining the most appropriate timing of assessments for COA-based endpoints include, but are not limited to, the following:

- **Recall period:** A COA should not be administered more frequently than the ***recall period*** allows (refer to FDA PFDD G3 Public Workshop Discussion Document Section VI.B.7 for in-depth discussion of considerations regarding an instrument's recall period). For example, an instrument with a 1-week recall period should be administered no more frequently than 1 week (7 days) after the previous administration. If the recall period implies assessment at a specific time of day (e.g., in the morning, at night) or at a specific time relative to treatment (e.g., since last dose) or relative to some other event (e.g., since waking up today, since going to bed last night, since last bowel movement), assessments should be timed accordingly. This issue also arises in wording of COAs administered using ecological momentary assessment.
- **Anticipated rate of change in the underlying construct to be measured:** The timing of assessments should align with the anticipated rate of change in the underlying construct to be measured (but, as mentioned above, should be no more frequent than what the instrument's recall period allows). For example, if the construct to be measured is expected to change rapidly over the course of the study period, assessments should be placed closer together. If the construct is expected to change slowly, one might place assessments further apart. Note that rate and direction of change in the underlying construct is linked to the rate and direction of change in the underlying disease/condition to be treated (i.e., linked to the pace of improvement or deterioration/progression in the underlying disease), but the two may not move together in lock-step (i.e., they probably would move in the same direction but may move at different rates).
- **COA administration burden:** The length and frequency of COA administration should take into consideration patient burden which may result in patient fatigue and lead to an increase of missing data, as well as impact data quality.
- **COA administration schedule:** The schedule of COA administration should align with the administration of other prespecified endpoints (i.e., primary and secondary) and proposed SAP.

- **Collect COA data at baseline:** The COAs should be administered at baseline. If the trial includes a run-in period during which the effect on the COA might be expected to change (e.g., medication washout, patient behavior modification), this should be considered when considering the timing of assessments. Note that some diseases, conditions, or clinical trial designs may necessitate more than one baseline assessment and several COA administrations during treatment.
- **Align anchor administration time:** The timing of anchor scale administration should align with both the recall period and the administration of the corresponding COA (e.g., patient global impression of severity (PGIS) with PRO timing; clinician global impression of severity with ClinRO timing).
- **Use same COA administration order:** The order of COA administration should be standardized to help reduce measurement error.
- **Timing of treatment administration:** If treatment is administered repeatedly over the clinical trial period and change in the target construct(s) is to be assessed repeatedly over the trial period, it may be sensible to measure the construct at the same time relative to treatment administration throughout the trial—unless treatment considerations dictate otherwise.

4. Defining Improvement and Worsening

Clinically relevant within-patient thresholds for improvement and worsening should be predefined and justified. A few suitable supplementary analyses may be conducted to evaluate a range of thresholds when appropriate. See Section III of this Discussion Document for additional information.

Superiority versus noninferiority or equivalence testing of a COA-based endpoint must be predefined in the SAP. It is inappropriate to conclude "no worsening" when there is a nonsignificant test of superiority (e.g., $p > 0.05$). Trials with small sample sizes lead to wide confidence intervals of the treatment effect of the COA-based endpoint, which will likely not demonstrate superiority.

5. Clinical Trial Duration and COA-Based Endpoints

Generally, the duration a COA is collected should be the same duration as indicated for other measures of effectiveness or safety in the clinical trial protocol. It is important to consider whether the clinical trial's duration is of adequate length to assess a durable COA-based endpoint in the disease or condition being studied. Determination of the clinical trial duration should be driven by the disease course as well as treatment and endpoint objectives outlined in the clinical trial protocol.

D. Intercurrent Events: What Can Affect Your Measurement's Interpretation?

Intercurrent events are events that occur after randomization/treatment initiation/or trial start that either preclude observation of the variable (and potentially subsequently the endpoint) of interest or affect its interpretation (e.g., taking rescue medication). While missing data is a part of the definition, it is not the only definition.

1. Use of Assistive Devices, Concomitant Medications, and Other Therapies

It is important to consider what other activities may impact the COA score and endpoint value, such as use of assistive devices (e.g., walkers), concomitant medications including rescue therapies (e.g., bronchodilators or pain medication), and other therapies (e.g., physical therapy). For example:

- Use of assistive devices may particularly impact PerfO assessment of mobility and can impact other types of COAs
- If a specific published administrator's manual is selected for a performance-based test, it is important to conduct the test in accordance with the selected manual, including the use of standardized assistive devices, if allowed
- If study procedures are not aligned with the instrument's user manual, changes should be detailed in the study documents and training should occur specific to the changes
- ***Case report forms*** (CRFs) for data collection should include information on whether an assistive device (and what type) was used during the test

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, it would be informative to incorporate the information on the assistive device into the COA-based endpoint construction, as the change in assistive device may reflect either an improvement or a deterioration in the patient's disease status.

Two other examples of intercurrent events:

- If an item assesses difficulty buying groceries and wording does not account for use of a food delivery service, an intercurrent event could occur.
- If a patient in a trial breaks their leg in a car accident, that likely impacts the physical function PRO instrument's score.

Use of other supportive therapies that may impact the interpretation of the endpoint should be assessed consistently. Data should be collected and recorded in a standardized manner, and incorporated into the endpoint model and supplementary analyses. A discussion with study coordinators, statisticians, clinicians, and patients will result in a list of likely intercurrent events to include in study planning.

2. Impact of Disease/Condition Progression, Treatment, and Potential Intercurrent Events

In the planning stages of a clinical study, it is important to consider how both the disease/condition and treatment may impact a patient's ability to function cognitively and physically over the course of the study as the disease/condition progresses or as treatment ***side effects*** manifest, including ability to communicate, follow instructions (verbal and written), receive and understand information, and complete the assessment.[8] Missed or incomplete assessments due to disease progression or treatment side effects "may provide meaningful information on the effect of a treatment and hence may be incorporated into a variable [(or endpoint)], with appropriate summary measure, that describes a meaningful treatment effect" (ICH E9(R1)).

Since model-based estimates generally tend to be "very sensitive" to model misspecification, it is recommended that supplementary and sensitivity analyses be conducted to examine how much the results/findings change under various assumptions about the missing data mechanism (National Research Council, 2010). Principles and methods for sensitivity analyses are discussed further in ICH E9(R1) and Chapter 5 of the National Research Council's 2010 report on *The Prevention and Treatment of Missing Data in Clinical Trials* (National Research Council, 2010).

Changes in physical or cognitive function due to disease/condition progression and/or treatment effects are important outcomes to be measured and either incorporated into the study endpoint structure or reported as safety findings.

For some risk factors of cognitive or physical change unrelated to disease or treatment (such as advancing age), the chances of a patient's cognitive or physical function changing over the course of the study may increase with study duration. Use of appropriate inclusion and exclusion criteria may help mitigate some potential causes of cognitive and physical change. However, restrictive criteria can impact the ability to recruit and the generalizability of study results. Because changes in cognitive and physical function may still occur during the study, it is important to note sources of ***competing risks*** and other intercurrent events in the SAP and Study Report.

3. Practice Effects

A ***practice effect***[9] (sometimes also called a ***learning effect***) is any change that results from practice or repetition of completing particular tasks or activities including repeated exposure to an instrument. A simple example is taking a math test. After completing the same test three times

[8] When disease progression or treatment side effects result in missed or incomplete assessments, those missing COA data are considered to be *informatively missing* or *missing not at random* (MNAR). Missing observations (e.g., missing COA data) are considered to be informatively missing or MNAR "when there is some association between whether or not an observation is missing (or observed) and the status of the patient's underlying disease" (Lachin, 1999). Failing to incorporate both observed and unobserved (i.e., missing but potentially observable) COA data from the entire ITT population in analyses involving the COA-based endpoint will likely yield biased (erroneous; misleading) results.

[9] Note that ***practice effects*** may be referred to using different terminology in different disciplines.

your speed (and maybe accuracy in answering) likely will improve because you 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, Cook, & Campbell, 2002).

Practice effects may be problematic for studies conducted to support a medical product regulatory application. Practice effects, by definition, lead to improvement in the score of the assessment. This score improvement confounds score changes attributable to the clinical trial intervention. In randomized controlled trials, if practice effects are constant across trial arms, they will not bias the difference of the outcomes between arms. However, if practice effects interact with clinical trial intervention such that the magnitude and direction of practice effects differ by trial arm, the treatment effects may be deflated or inflated (Song & Ward, 2015). Deflation of treatment effects may result in delayed patient access to effective treatment options, and inflation of treatment effects may expose patients to risk due to wasted time and resources spent pursuing ineffective treatments. Whether the practice effects are constant or differ across clinical trial arms is generally unknown. Therefore, the best strategy is to minimize the potential for practice effects in clinical studies.

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 (Jones, 2015). Some general strategies for mitigating practice effects are summarized below. These strategies may 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 instruments' robustness (or vulnerability) to practice effects. Such evidence may be obtained through, for example, a thorough review of the literature.
- **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, Cook, & Campbell, 2002). Decisions regarding the length of time (spacing) to place between assessments should take into consideration both how rapidly (or slowly) change in the underlying construct is expected to occur and the recall period utilized by the instrument. Refer to Section II.C.3 of this Discussion Document for more detailed considerations regarding timing of assessments.
- **Increase the length of 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 does not significantly increase such that the baseline and postbaseline score are minimally affected by practice effects.
- **Use *alternate forms* (sometimes also referred to as *parallel forms* or *equivalent forms*):** Alternate 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" (Test Design and Development, 2014). Administering different forms comprised of distinct sets of items may make practice effects less likely to occur.

For the use of alternate forms to attenuate practice effects without introducing additional bias: (1) alternate forms must be truly psychometrically equivalent;[10] and (2) alternate forms must be administered in a random order that differs by study arm (i.e., a counterbalanced, randomized order) (Jones, 2015; Goldberg, Harvey, Wesnes, Snyder, & Schneider, 2015).

4. Participant Burden

The possibility of participant burden compromising the validity of the endpoint should be assessed. Burden may lead to missing data or inaccurate data (e.g. answering the first response to every item). When an endpoint is derived from multiple administrations of a COA, attention should be paid to whether study subject fatigue or patient burden might diminish the validity of COA scale scores. This, in turn, could compromise the validity of the endpoint itself, leading to biased estimates of treatment effects and inaccurate hypothesis tests. Study subject fatigue is less likely to occur if an endpoint is based on a small number of widely spaced administrations, and more likely to occur if an endpoint is based on a larger number of administrations over a limited period of time. The effort required for the subject to complete the COA also influences the probability that subject fatigue will compromise scale and endpoint validity.

For the sake of illustration, suppose subjects are expected to complete a 25-item PRO for seven consecutive days, with the endpoint being the average of the seven daily scores. Some study subjects may grow fatigued at needing to complete the PRO for seven consecutive days, and such fatigue could manifest itself in a variety of ways:

- Subjects stop completing the PRO at some point after the initial administration and/or choose not to respond to some items at a given administration.
- Subjects recall item responses they made the previous day and repeat prior item responses rather than carefully considering how to respond to each item.
- Subjects tend to give the same rote response to each item rather than carefully considering how to respond to each item.

The first type of fatigue response will increase endpoint missingness. While this is not, strictly speaking, an issue of reliability or validity, it clearly compromises the use of the endpoint for assessing its construct. The second and third types of fatigue responses compromise the validity

[10] For two different instruments to be considered *parallel*, they must have matching content (i.e., each instrument must measure the same symptom, function, or impact); estimated item parameters and corresponding standard errors must not significantly differ; estimated score reliability and corresponding standard errors must not significantly differ; and score means and standard deviations (surrogates for the distributions of the two sets of scores) in the target population must not significantly differ (Test Design and Development, 2014).

of the endpoint, as the validity of at least some of the PRO administrations per fatigued subject are compromised.

5. Mode of Administration

Changes or disruptions to standardized instrument administration procedures should be documented and may need to be included in the data analyses. Depending on the construct being measured, the assessment environment should provide the reporter with reasonable comfort and minimal distractions to avoid introducing construct-irrelevant variance into the resulting COA scores (American Educational Research Association; American Psychological Association; National Council on Measurement in Education, 2014).

COA data collection modes can include paper-based and/or electronic-based approaches. Types of COA administration can include self-administration, interviewer-administration (e.g. face-to-face, via telephone or electronic means), clinician-administration, and/or trained administrator-administration (FDA PFDD G3 Discussion Document). To help ensure the instrument's established psychometric measurement properties hold in the study at hand, the COA must be administered in accordance with standardized administration and ***scoring algorithm*** specified by the instrument developer (such as in the instrument's user manual or website). For modes of data collection that do not include a date and time stamp (e.g., paper diaries), it is difficult to ensure that patients enter data at the protocol-specified time.

6. Missing Data and Event-Driven COA Reporting

Programming errors can result in significant amounts of missing data which impedes interpretation of analysis results. For example, a COA may be designed to give patients the option to report additional events and event-related symptoms not reported during the day; however, a potential programming error could cause the additional questions to not be administered at the end of the day. Large amounts of missing data would be generated, resulting in underreporting of the event and the study endpoint itself being unreliable and uninterpretable.

7. Missing Scale-Level Data

Missing data should be distinguished from data that do not exist or data that are not considered meaningful due to an intercurrent event. The protocol and the SAP should address plans for how the statistical analyses will handle missing COA data when evaluating clinical benefit and when considering patient success or patient response.

In cases where patient-level COA data are missing for the entire domain(s) or the entire measurement(s), sponsors should clearly define missing data and propose statistical methods that properly account for such data with respect to a particular estimand. Methods to handle the missing data for a COA-based endpoint and any related supportive endpoints should be addressed in the protocol and the SAP. In addition, the supplementary and sensitivity analyses of the COA-based endpoints should be prospectively proposed in the protocol and the SAP. These

analyses investigate assumptions used in the statistical model for the main analytic approach, with the objective of verifying that inferences based on an estimand are robust to limitations in the data and deviations from the assumptions.

E. Population-Level Summary: What Is the Final Way All Data Are Summarized and Analyzed?

The population-level summary serves as the basis for comparison between treatment arms, treatment conditions, other groups, or otherwise summarizes information. Examples include a) mean physical function score at baseline for everyone in an observational research study and b) difference compared to control of a new medical product's median time to pain resolution.

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 analyzing COA-based endpoints.

1. Landmark Analysis

Sponsors should justify the use of and time in which a landmark analysis (an analysis at a fixed time point, e.g. 12 weeks) is to be performed. If a COA-based endpoint is collected repeatedly, information may be lost in conducting a landmark analysis. However, even when conducting a landmark analysis at a fixed time point, data from intermediate time points (i.e., measurements taken prior to the fixed time point) can still be included in the model. Interpretation of an analysis of overall COA score over time may be difficult in the presence of missing data. The interpretation of potential analyses when COA data collection is truncated due to death or other events should be carefully discussed within the research team.

2. Analyzing Ordinal Data

When an ordinal endpoint has a limited number (e.g., 3 to 7) of categories, you should describe, analyze, and interpret the study result on this endpoint using methods appropriate for ordinal variables, e.g., ordinal regression. For descriptive statistics, mean and standard deviation should not be used on an ordinal endpoint. Percentiles and bar graphs can be informative.

3. Time-to-Event Analysis

Defining and identifying an event is an issue for time-to-event analysis of COA-based endpoints and responder analyses of ordinal or continuous COA data. A clinically relevant threshold for deterioration, maintenance, or improvement must be predefined and justified. Relevant information on intercurrent events, censoring rules, and defining an event should be prespecified in the protocol. It is important to explicitly state how to handle intercurrent events. For example, estimates may differ if death is considered a deterioration event versus censored. Censoring rules in the presence of missing COA data should be prespecified in the SAP. Furthermore, analyses to

evaluate assumptions of the primary time-to-event analysis should be performed under differing censoring rules.

4. Responder Analyses and Percent Change From Baseline

As previously mentioned, COA data often are ordinal or continuous in nature. Sponsors should consider analyzing COA-based endpoints as continuous or ordinal variables rather than as a responder (i.e., dichotomized from either ordinal or continuous COA data) to avoid misclassification errors and potential loss of statistical power. There tends to be more precision in the evaluation of medical product effects on continuous variables (i.e., based on a comparison of means), especially when sample size is of concern. Alternative approaches for analysis (e.g., analyses based on ranks) should be included, if appropriate, in the SAP to account for occurrence of extreme outliers.

If a responder endpoint is deemed appropriate for a trial and the endpoint is proposed based on dichotomization from either ordinal or continuous data, it is prudent for the sponsor to prespecify a single responder threshold and provide evidence to justify that the proposed responder threshold constitutes a clinically meaningful within-patient change prior to the initiation of the trial. Proposed responder threshold(s) should be discussed with FDA prior to the initiation of the trial as it is crucial for sample size planning and to appropriately power the study.

In general, for COA-based endpoints FDA does not recommend a responder analysis endpoint or a percent change from baseline endpoint unless the targeted response is complete resolution of signs and symptoms. While percent change from baseline is popular in other contexts, the statistical measurement properties are poor. Strange occurrences arise, for example in randomized withdrawal studies we have seen subjects needing to reach a percent change from baseline threshold who end up needing significantly higher symptom burden to go back on treatment compared to symptom levels needed to enter the trial based on inclusion criteria. Extreme caution should be exercised, and all potential endpoint situations explored especially near the floor and ceiling of the COA or COA-based endpoint's values, before using percent change from baseline as the population-level summary.

The Appendix contains a case study that illustrates several of these concepts and guides us to using the estimand framework to better develop the SAP and ultimately more transparently communicate study results.

III. MEANINGFUL WITHIN-PATIENT CHANGE

Section Summary

To aid in the interpretation of study results, FDA is interested in what constitutes a meaningful within-patient change (i.e., improvement and deterioration from the patients' perspective) in the concepts assessed by COAs. Statistical significance can be achieved for small differences between comparator groups, but this finding does not indicate whether individual patients have experienced meaningful clinical benefit.

Technical Summary: Key Messages in This Section

- What constitutes, from a patient perspective, a meaningful within-patient change in the concepts evaluated by COAs.
- FDA recommends the use of anchor-based methods to establish meaningful within-patient changes, although there are other methods that can be used.
- Anchors selected for the trial should be plainly understood in context, easier to interpret than the clinical outcome itself, and sufficiently associated with the target COA and/or endpoint.
- Anchor-based methods should be supplemented by the use of empirical cumulative distribution function (eCDF) curves and probability density function (PDF) curves.

Interpretation of Within-Patient Meaningful Change

To holistically determine what is a meaningful change, both ***benefit*** and ***risk***, improvement and deterioration, may need to be accounted for. This document is not directly addressing this integration of benefit and risk, but the methods described can be used to help interpret benefit or risk. As such, special consideration should be given by the sponsor to assess how meaningful the observed differences are likely to be. To aid in the interpretation of the COA-based endpoint results, sponsors should propose an appropriate threshold(s) (e.g., a range of score change) that would constitute a clinically meaningful within-patient change in scores in the target patient population for FDA review.

In addition, if the selected threshold(s) are based on transformed scores (e.g., linear transformation of a 0-4 raw score scale to a 0-100 score scale), it is important to consider score interpretability of the meaningful change threshold(s) for both transformed scores and raw scores. Depending on the proposed score transformation, selected threshold(s) based on transformed scores may reflect less than one category change on the raw score scale, which is not useful for the evaluation and interpretation of clinically meaningful change.

Meaningful Within-Patient Change Versus Between-Group Difference

It is important to recognize that individual within-patient change is different from between-group difference. From a regulatory standpoint, FDA is more interested in what constitutes a meaningful within-patient change in scores from the patient perspective (i.e., individual patient level). The between-group difference is the difference in the score endpoint between two trial arms that is commonly used to evaluate treatment difference. Between-group differences do not address the individual within-patient change that is used to evaluate whether a meaningful score change is observed. A treatment effect is different from a meaningful within-patient change. The terms minimally clinically important difference (MCID) and minimum important difference (MID) do not define meaningful within-patient change if derived from group-level data and therefore should be avoided. Additionally, the minimum change may not be sufficient to serve as a basis for regulatory decisions.

A. Anchor-Based Methods to Establish Meaningful Within-Patient Change

Anchor-based methods utilize the associations between the concept of interest assessed by the target COA and the concept measured by separate measure(s), referred to as anchoring measure(s), often other COAs. FDA recommends the use of anchor-based methods supplemented with both empirical cumulative distribution function (eCDF) and PDF curves to establish a threshold(s), or a range of thresholds, that would constitute a meaningful within-patient change score of the target COA or the derived endpoint for the target patient population. The anchor measure(s) are used as external criteria to define patients who have or have not experienced a meaningful change in their condition, with the change in COA score evaluated in these sets of patients. Sponsors should provide evidence for what constitutes a meaningful change on the anchor scale by specifying and justifying the anchor response category that represents a clinically meaningful change to patients on the anchor scale, e.g., a 2-category decrease on a 5-category patient global impression of severity scale.

Considerations for Anchor Measures

- Selected anchors should be plainly understood in context, easier to interpret than the COA itself, and sufficiently associated with the target COA or COA endpoint
- Multiple anchors should be explored to provide an accumulation of evidence to help interpret a clinically meaningful within-patient score change (can also be a range) in the clinical outcome endpoint score
- Selected anchors should be assessed at comparable time points as the target COA but completed after the target COA

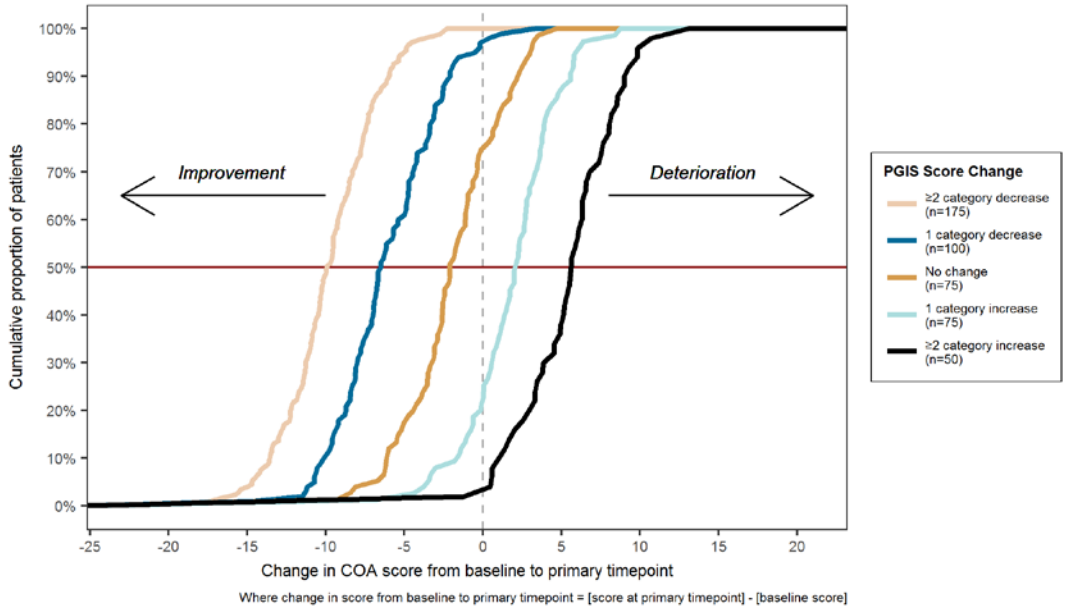
- The following anchors are sometimes recommended to generate appropriate threshold(s) that represent a meaningful within-patient change in the target patient population:
  - Static, current-state global impression of severity scale (e.g., PGIS)
  - Global impression of change scale (e.g., patient global impression of change or PGIC)
  - Well-established clinical outcomes (if relevant)
- A static, current state global impression of severity scale is recommended at minimum, when appropriate, since these scales are less likely to be subject to recall error than global impression of change scales; they also can be used to assess change from baseline.

B. Using Empirical Cumulative Distribution Function and Probability Density Function Curves to Supplement Anchor-Based Methods

The eCDF curves and PDF curves can be used to supplement anchor-based methods. The eCDF curves display a continuous view of the score change (both positive and negative) in the COA-based endpoint score from baseline to the proposed time point on the horizontal axis, with the vertical axis representing the cumulative proportion of patients experiencing up to that level of score change. An eCDF curve should be plotted for each distinct anchor category as defined and identified by the anchor measure(s) (e.g., much worse, worse, no change, improved, much improved).

As a reference, Figure 2 provides an example of eCDF curves. Note that the median change is indicated by the red line in this example. The number of PGIS category increases and decreases defines the example's curves. In some instances, not all two (or 1 or 0) category changes are the same. This should be considered when choosing an anchor summary and interpreting these figures and data.

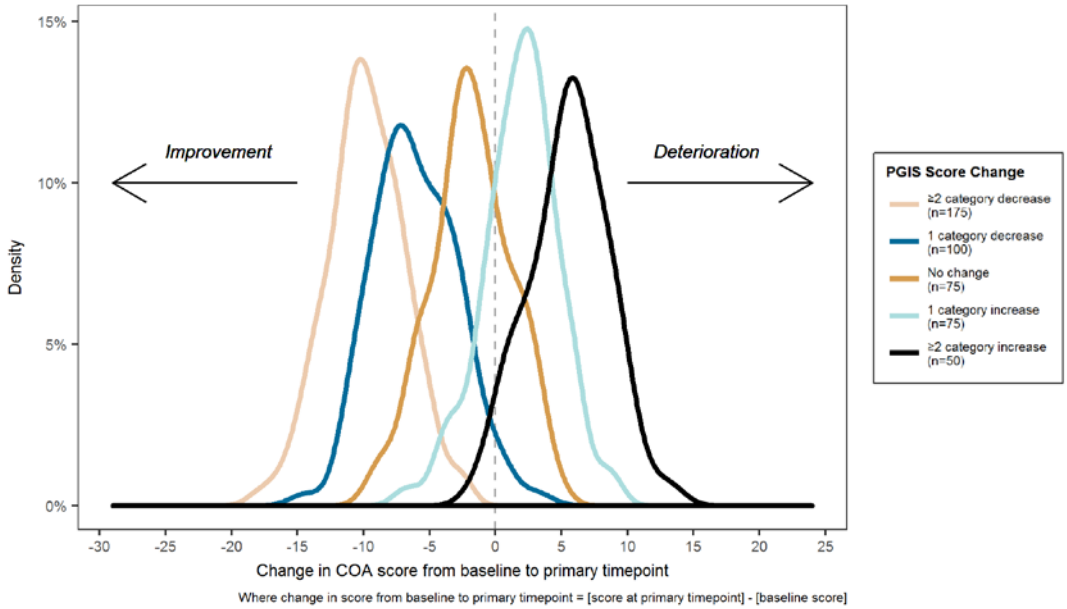
Figure 2: Example of Empirical Cumulative Distribution Function Curves of Change in COA Score from Baseline to Primary Time Point by Change in PGIS Score



Abbreviations: PGIS = patient global impression of severity; COA = clinical outcome assessment

The PDF curves are useful in aiding the interpretation of eCDF curves. Compared with eCDF curves, PDF curves may provide a more intuitive overview of the shape, dispersion, and skewness of the distribution of the change from baseline in the endpoint of interest across various anchor categories. Figure 3 provides an example of PDF curves.

Figure 3: Example of Density Function Curves of Change in COA Score from Baseline to Primary Time Point by Change in PGIS Score



Abbreviations: PGIS = patient global impression of severity; COA = clinical outcome assessment

C. Other Methods

1. Potentially Useful Emerging Methods

Other methods may be explored to complement the anchor-based methods or when anchor-based methods are not feasible (i.e., when no adequate anchor measure(s) are available). For example, mixed methods may be used to triangulate and interpret COA-based endpoint results. The ***qualitative research methods*** in the PFDD Guidance 1 and Guidance 2 documents are frequently used, including ***cognitive interviews***, exit interviews, or surveys to help inform the improvement threshold. In addition, patient preference studies, typically surveys or interviews, may be utilized to help interpret and support clinical trial results.

There are several methods emerging in the health sector as potential ways to derive and interpret clinically meaningful change (Duke Margolis meeting summary, 2017), including scale-judgement and bookmarking/standard-setting. These methods are relatively new in the regulatory setting.

2. Distribution-Based Methods

Distribution-based methods (e.g., effect sizes, certain proportions of the standard deviation and/or standard error of measurement) do not directly take into account the patient voice and as such cannot be the primary evidence for within-patient clinical meaningfulness. Distribution-based methods can provide information about measurement variability.

3. Receiver Operator Characteristic Curve Analysis

Unless there is significant knowledge about how a COA performs in a specific ***context of use***, FDA does not recommend using receiver operator characteristic (ROC) curve analysis as a primary method to determine the thresholds for within-patient meaningful change score. The ROC curve method is a model-based approach, such that different models may yield different threshold values. Additionally, the ROC curve method is partially a distributional-based approach, such that the distribution of the change scores of the two groups will determine the location of the threshold. The most sensitive threshold identified by ROC may not actually be the most clinically meaningful threshold to patients.

The ROC curve method is appropriate for evaluating the performance (e.g., sensitivity and specificity) of the proposed responder thresholds derived from the anchor-based methods.

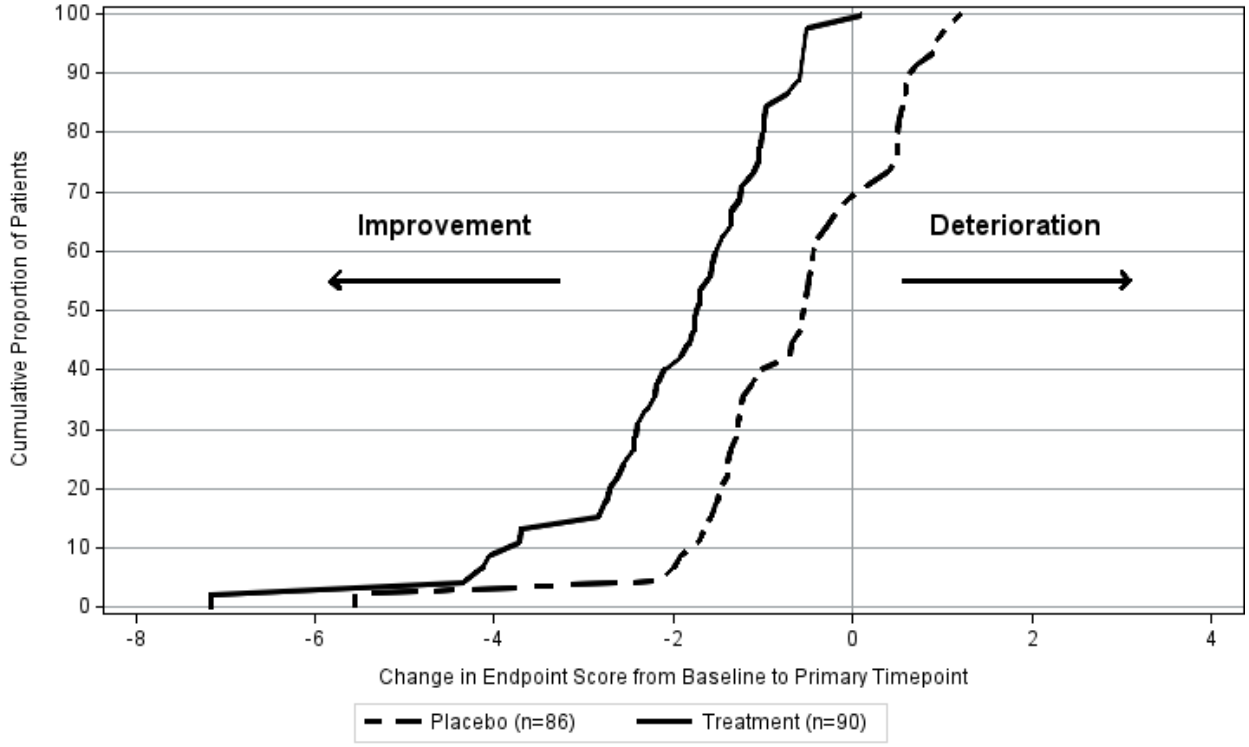
D. Applying Within-Patient Change to Clinical Trial Data

Clinical trials compare groups. To help evaluate what constitutes a meaningful within-patient change (i.e., improvement and deterioration from the patients' perspective), you should examine whether treatment arms show separation in the range of clinically meaningful within-patient change thresholds evaluated using methodologies described in other parts of this document.

When analyzing a COA-based endpoint as either a continuous or an ordinal variable, it is important to evaluate and justify the clinical relevance of any observed treatment effect. Sponsors should plan to evaluate the meaningfulness of within-patient changes to aid in the interpretation of the COA-based endpoint results by submitting a supportive graph (i.e., eCDF) of within-patient changes in scores from baseline with separate curves for each treatment arm. The graph will be used to assess whether the treatment effect occurs in the range that patients consider to be clinically meaningful.

Figure 4 provides an example of an eCDF curve by treatment arm, where there is consistent separation between the treatment arms. The treatment effect occurs in the range patients consider to be clinically meaningful.

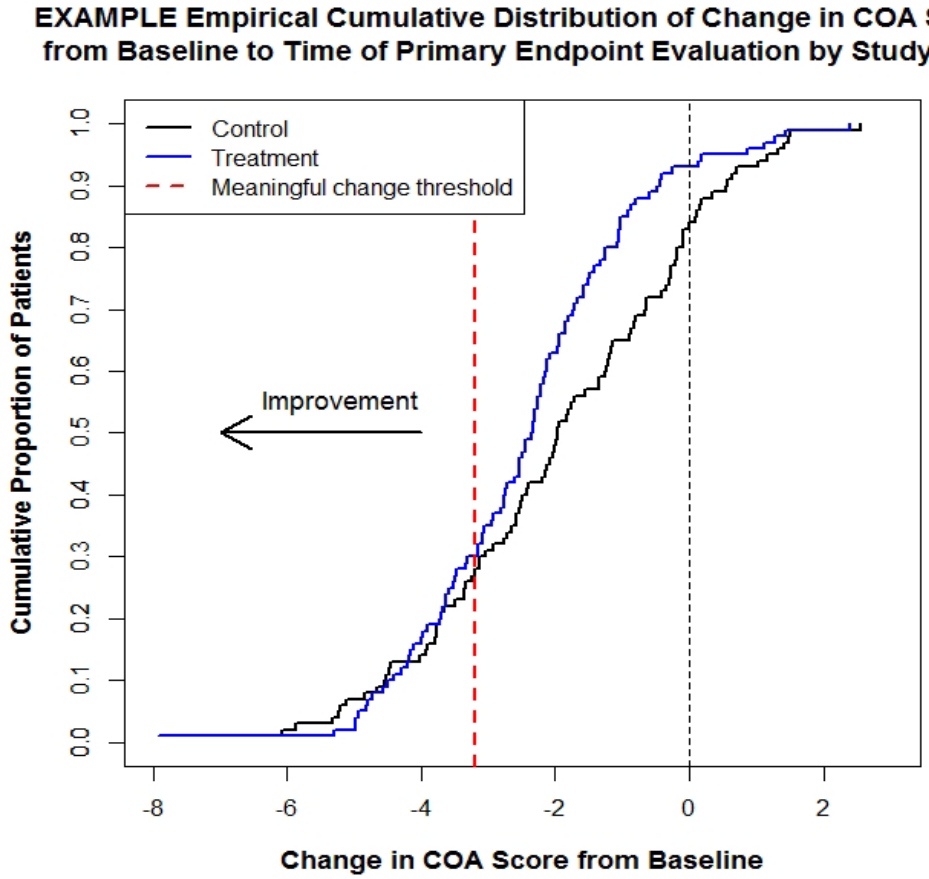
Figure 4: An eCDF Curve by Treatment Arm Showing Consistent Separation Between Two Treatment Arms



Abbreviation: eCDF = empirical cumulative distribution function

Figure 5 provides an example of an eCDF curve where the treatment effect does not occur in the range patients consider to be clinically meaningful. Of note, the eCDF does not take in to account estimation uncertainty and is not a test.

Figure 5: An eCDF Curve Where Treatment Effect Is Not in Range Considered Clinically Meaningful by Patients



Abbreviations: eCDF = empirical cumulative distribution function; COA = clinical outcome assessment

IV. ADDITIONAL CONSIDERATIONS

Key Messages in This Section

- Appropriately positioned COAs intended to support approval and/or labeling claims are in the endpoint testing hierarchy.
- A trial's protocol and SAP should state each COA-based endpoint as a specific clinical trial objective.
- Address multiplicity concerns and plans for handling missing data at both the instrument and patient level.
- Short list of formatting and submission considerations applicable to COA data.

When planning a study, confirm the following:

1. Each COA-based endpoint is stated as part of a specific clinical trial objective
2. COAs intended to support meaningful outcomes to patients (i.e., labeling claims or other communications) are fit-for-purpose and sensitive to detect clinically meaningful changes
3. Clinical trial duration is adequate to support COA objectives
4. Frequency and timing of COA administration is appropriate given patient population, clinical trial design and objectives, and demonstrated COA measurement properties
5. How blinding or masking will be implemented (e.g. assessor blinding)
6. Plans for instrument administration are consistent with instrument's user manual, or if different are well-developed, communicated to all study sites, and documented
7. Procedures for training are well-described
8. Content and scoring information are clearly delineated in the clinical trial protocol
9. Plans for COA scoring are consistent with those used during instrument development
10. COA-based endpoints intended to support approval and/or labeling claims are appropriately positioned in the endpoint testing hierarchy
11. Plans for multiplicity adjustment
12. Plans for handling missing data at both the instrument (e.g. person skips an item but answers other items on a PRO) and patient (e.g. patient does not provide any responses for a PRO at a study visit) level
13. Procedures include assessment of COA-based endpoint before or shortly after a patient withdraws from the clinical trial

14. Plans for COA measurement after discontinuation from treatment are driven by the research questions

15. Description of how between-group differences will be portrayed (e.g., cumulative distribution function)

16. Data collection, data storage, and data handling and transmission of procedures, including electronic COAs, are specified

Both SPIRIT (Calvert et al, 2018) and CONSORT (Calvert et al, 2013) consensus documents have been published with extensive details on what PRO information should be included in trial protocols and manuscripts. This information is extensible for most COAs.

A. Other Study Design Considerations

Blinding: Patients' and/or clinicians' knowledge of treatment assignment may lead to changes in how they report information on a COA, or how they engage with PerfO ***tasks*** (e.g., amount of encouragement given to patients when measuring walking distance). The protocol should specify who will evaluate the COA-based endpoints, outcomes, or measurements in relation to the subjects (e.g., the investigator or an independent evaluator/rater) as well as who the intended reporter of patient information will be (e.g., clinicians, patients, caregivers) and to what extent blinding will be maintained among the investigators, evaluators/raters and reporters (e.g., clinicians, patients, caregivers).

Considerations When Using a Nonrandomized or Nonconcurrent Control: When considering the use of COAs to support endpoints in an externally controlled trial, it is important to establish comparability of the COAs both within each of the treatment and external control groups and between the treatment and external control groups. It will be essential to use well-defined and reliable COAs across comparator arms, 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 case report forms), to allow meaningful comparison of changes over time.

These considerations apply to clinical trials, as well as natural history studies (see FDA draft guidances *Rare Diseases: Natural History Studies for Drug Development* (FDA, 2019) and *Rare Diseases: Common Issues in Drug Development* (FDA, 2019), and FDA final guidance *Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices* (FDA, 2017)), 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 ICH guidance for industry *E10 Choice of Control Group and Related Issues in Clinical Trials* (ICH E10).

Computerized Adaptive Testing (CAT): We are asked questions about the use of CAT during trials. We encourage people to submit to the docket content they would like to see in the guidance.

B. Formatting and Submission Considerations

Regardless of how a COA is administered in a given study, COA data collected and submitted to FDA to support a regulatory medical product application are subject to all the same regulations and submission requirements as other types of study data, such as, but not limited to, the following:

- ICH guidelines, such as *M8 Electronic Common Technical Document (eCTD)*
- The Electronic Code of Federal Regulations, Title 21, Chapter 1 (21 eCFR, Chapter 1)—with particular attention given to Parts 11, 21, 312.57, and 312.62(b, c)
- FDA guidance *Use of Electronic Records and Electronic Signatures in Clinical Investigations Under 21 CFR Part 11 – Questions and Answers* (June 2017)
- FDA guidance *Computerized Systems Used in Clinical Investigations* (May 2007)
- FDA guidance *Electronic Source Data in Clinical Investigations* (September 2013)
- FDA guidance *Providing Regulatory Submissions in Electronic Format—Standardized Study Data* (December 2014)
- FDA guidance *Providing Regulatory Submissions in Electronic Format — Submissions Under Section 745A(a) of the Federal Food, Drug, and Cosmetic Act* (December 2014)
- FDA guidance *Providing Regulatory Submissions in Electronic Format – Certain Human Pharmaceutical Product Applications and Related Submissions Using the eCTD Specifications* (January 2019)
- The FDA Data Standards Catalog and other data standards, accessible here

Electronic devices used to administer COAs in studies conducted to support a regulatory medical product application have special development, testing, and deployment considerations like any ***digital health technology***, including a need for ***usability studies***. The following FDA guidances and related discussion documents have more information:

- FDA PFDD G3 Discussion Document
- FDA guidance *Contents of a Complete Submission for Threshold Analyses and Human Factors Submissions to Drug and Biologic Applications* (September 2018)
- FDA guidance *Comparative Analyses and Related Comparative Use Human Factors Studies for a Drug-Device Combination Product Submitted in an ANDA* (January 2017)
- FDA guidance *Applying Human Factors and Usability Engineering to Medical Devices* (February 2016)

- FDA guidance *Human Factors Studies and Related Clinical Study Considerations in Combination Product Design and Development* (February 2016)
- FDA guidances with digital health content

APPENDIX 1: CASE STUDY OF ESTIMAND FRAMEWORK

Key Messages in This Section

- Example of a randomized, concurrently controlled trial in which patients are randomized to one of two treatment arms: the current standard of care plus the investigational medical product; or the current standard of care plus placebo.
- The trial goal is to collect and analyze data to provide compelling evidence of the investigational product's efficacy in improving progression-free survival (PFS) and on a secondary endpoint of physical function measured using a PRO to support a labeling claim.
- The Case Study shows how the trial's research objective and scientific research question drive the approaches used within the attributes of the estimand framework presented earlier in this document: Target Study Population, Endpoint of Interest, Intercurrent Events, and Population-Level Summary.
- It also shows how the previous considerations drive the nature of the trial's SAP.

This Case Study exemplifies employment of the estimand framework when considering physical function in certain breast cancer patients that have progression on first line (standard of care) therapy. Breast cancer has heterogeneous disease symptoms and many women will be asymptomatic at baseline even in the second line setting. For this example, second line prior studies have shown a median overall survival (OS) time of 2-2.5 years with second line hormone therapy alone and a median PFS time of approximately 10-12 months. OS is defined as the time from randomization to date of death. PFS is defined as the time from randomization to date of first progression of disease or death due to any cause. This is a randomized controlled trial where patients are randomized in a 1:1 ratio to the following treatment arms:

- Treatment: Standard of care + oral targeted investigational agent
- Control: Standard of care + placebo

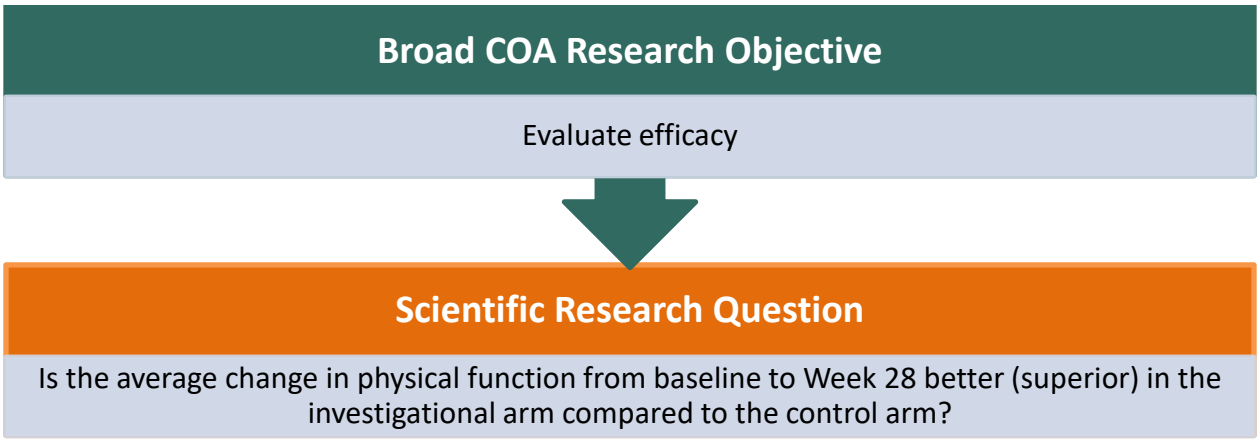
The primary efficacy endpoint is PFS, which is expected to show 6- to 8-month benefit with the addition of targeted therapy. OS may be impacted due to crossover. Symptomatic toxicities including diarrhea, fatigue, and rash are expected to be greater in the investigational arm. The population is generally high functioning (Eastern Cooperative Oncology Group (ECOG) 0 or 1) and is generally asymptomatic from disease at baseline.

A. Example Research Objective

The secondary endpoint's focused research objective is to use physical function assessed using a PRO to support a labeling claim. In this case, we would like to make conclusions by comparing the treatment arms. Therefore, a hypothesis test should be prespecified, and a correction for multiple testing is needed to control for Type I error.

1. *Define COA Scientific Research Question* A Priori

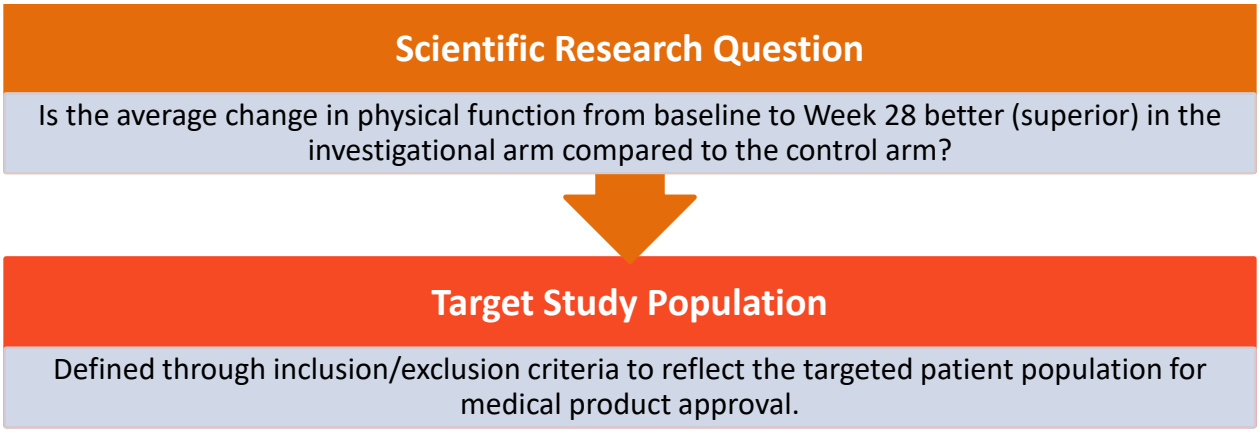
In the broad research objective, we prespecify that we intend to look at a superior benefit in physical function for the investigational arm. Based on this, we define our scientific research question as follows:



We would like to treat PRO endpoints with the same rigor as we would see with efficacy endpoints in oncology such as OS and PFS, particularly when we want to support a labeling claim. Often in oncology trials the sample size may be very small, leading to wide confidence intervals that do not demonstrate superiority. In addition, the COA may not be sensitive to change.

2. *Define Target Study Population Based on the Research Question* A Priori

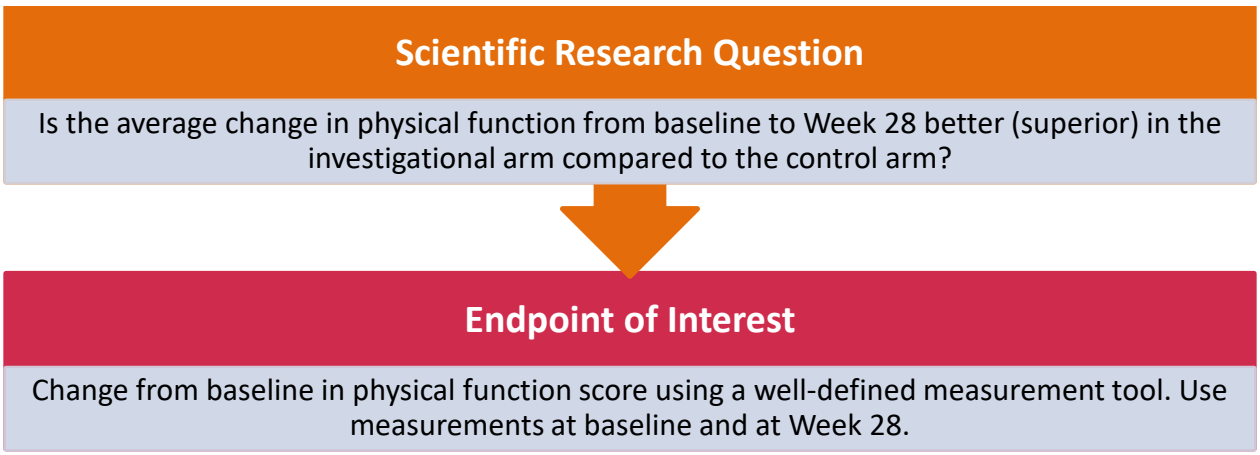
Since we are aiming to compare the two treatment arms for a labeling claim, we are defining our study population based on inclusion/exclusion criteria to reflect the targeted patient population for medical product approval.



If assessing tolerability (not efficacy) of the product while a patient is on treatment is of interest, we may want to include only patients who received at least one dose of the product, regardless of randomization.

3. Define Endpoint of Interest Based on the Research Question A Priori

Based on our scientific research question, we aim to collect change from baseline in physical function score at Week 28 assuming we already have a well-defined measurement tool.



We are looking at Week 28, which is around a 6-month time point in which the cumulative effects of the product in terms of both efficacy and toxicity have equilibrated.

Table 2 presents considerations in defining the COA-based endpoint of interest.

Table 2: Considerations When Defining a COA-Based Endpoint

Concepts (Examples)	Measurement Tool Qualities	Endpoint Type	Analysis Time Point
• Physical function • Pain	• Well-defined • Reliable • Validated • Sensitive	• Time-to-event • Proportion with event at time *t* • Continuous summary score at time *t* • Overall PRO score over time • Response patterns/profiles (longitudinal)	• Specific time point • Over time (specify time frame)

Abbreviations: COA = clinical outcome assessment; PRO = patient-reported outcome

Considerations for defining the endpoint include the concepts of interest. We chose physical function as an example, but another concept we might be interested in is pain. For each PRO endpoint, we look at all the measurement properties to check the instrument well-defined. Since our variable is change from baseline, statisticians may refer to this as a "continuous summary score at time *t*." If we were interested in time-to-deterioration, the endpoint type would have

been time-to-event. Other possibilities include proportions, overall PRO score over time, and response profiles.

We specified our analysis time point at Week 28. In addition, we might be interested in analyzing data over a specific time frame. For example, it may be of interest to analyze data at each PRO assessment time point while a patient is on treatment.

4. Address Intercurrent Events in Alignment with the Research Question

Based on our research question, some examples of intercurrent events that may impact interpretation include death, progression, and discontinuation. These events and the way they are handled will impact the estimate of the treatment effect. We specify that after date of death, we cannot collect or include physical function assessments in our analyses. We do not expect a high proportion of death to occur at the time of the analysis. For patients who discontinue treatment, progress, start physical therapy, initiate subsequent therapy or experience any other intercurrent event, we continue to collect physical function assessments regardless of these intercurrent events and will include them in our analysis.

Scientific Research Question

Is the average change in physical function from baseline to Week 28 better (superior) in the investigational arm compared to the control arm?

Intercurrent event	Handling of intercurrent event
• Death	Physical function not collected after intercurrent event occurs.
• Discontinuation of treatment • Disease progression • Physical therapy • Subsequent therapy	Physical function collected and included in analysis regardless of whether intercurrent event occurs.

Table 3 presents a list of additional intercurrent events that may impact interpretation of physical function. It is crucial to list intercurrent events and how they are handled in the analysis so that there is a clear understanding between regulators and sponsors of what is being estimated.

Table 3: Considerations When Addressing Intercurrent Events

Intercurrent Events (Examples)	Handling Intercurrent Events
• Death • Progression • Discontinuation due to adverse event • Taking subsequent therapy beyond discontinuation • Use of rescue medication or therapy • Analgesic use • Hospitalization • Nonadherence	• Prespecify handling of intercurrent events in alignment with research question • There are multiple ways to handle intercurrent events

5. *Define Population-Level Summary Based on Research Question* A Priori

Since our research question is looking at mean change from baseline between the two treatment arms, we chose the population-level summary to be the difference between the two arms in mean change from baseline to Week 28.

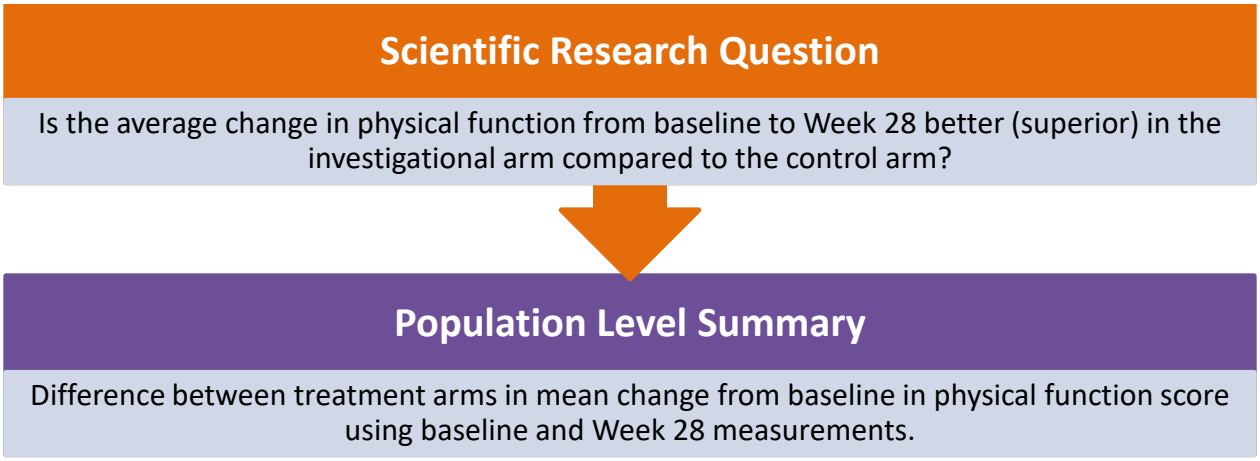


Table 4 presents considerations when defining the COA population-level summary.

Table 4: Considerations When Defining a COA Population-Level Summary

Population-Level Summary (Examples)	Clinical Relevance

• Median time to event, hazard ratio
• Proportion of patients with event at time *t*
• Mean change at time *t*
• Mean overall PRO score over time (e.g., mean area under the curve)
• Mean longitudinal profile
	• **Clinically relevant thresholds**
	  • Within-patient change
	• **Estimate**
	  • Within-group mean change
	  • Between-group difference

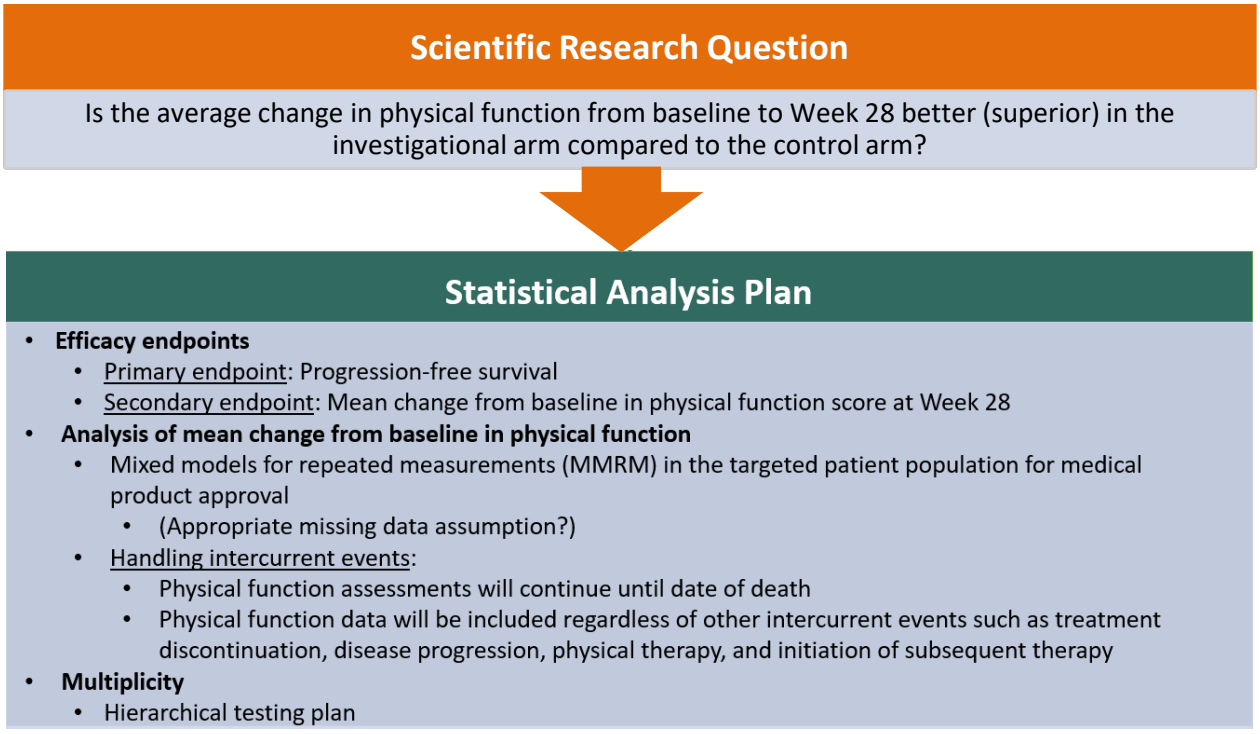
Abbreviations: COA = clinical outcome assessment; PRO = patient-reported outcome

Since we are looking at a magnitude, we chose mean change from baseline as the population level summary. If we were doing a time-to-event analysis, perhaps a hazard ratio and median time to event summary measure would have been used.

Next, we want to evaluate clinical relevance of our estimates. Once we have results, we want to know what these numbers mean and how they apply to the ***patient perspective***. FDA is interested in evaluating within-patient change to interpret clinical meaningfulness, a topic in Section III of this Discussion Document. Other estimates that are important in interpretation of clinical relevance include the within-group and between-group difference of mean change from baseline.

6. Prespecify Statistical Analysis Plan

Our primary endpoint is PFS with a secondary endpoint of mean change from baseline in physical function score at Week 28. We will analyze this endpoint using a mixed model for repeated measurements (MMRM) in the ITT population to obtain a least squares (LS) mean change from baseline in physical function score at Week 28 for each treatment arm, difference from control arm and their associated 95% confidence intervals.



We defined how we handle intercurrent events where patients are assumed missing after death, progression, or treatment discontinuation. Note that the MMRM assumes patients who drop out behave similarly to other patients in the same treatment group, who had similar covariate and COA data prior to dropping out. If, for example, a patient discontinues because of toxicity, this assumption may not be reasonable. The estimated treatment effect may be biased, leading to

uncertainty regarding information. Similar issues arise when death occurs prior to the landmark analysis date.

In this example trial, MMRM may be reasonable because we do not expect a high proportion of death to occur and will continue to collect physical function assessments regardless of progression or treatment discontinuation. Suitable supplementary analyses should be performed to challenge the assumptions of the prespecified analysis by incorporating reasons for missingness in the analysis.

B. Summary of Decisions Made in This Case Study

We applied the estimand framework to an example research objective to support a labeling claim in a second line advanced cancer trial. A summary of decisions made for the estimand based on the research question is given in Table 5.

Table 5: Summary of Estimand Decisions Made

Estimand Attributes	Decisions Based on Research Question
Target population	Defined through inclusion/exclusion criteria to reflect the targeted patient population for approval.
Endpoint of interest	Change from baseline in physical function score using well-defined measurement tool. Use measurements at baseline and at Week 28.
Handling of intercurrent events	
Death	Physical function data not collected after this intercurrent event occurs.
Disease progression; Treatment discontinuation; Physical therapy; Initiation of subsequent therapy	Physical function collected and analyzed regardless of whether these intercurrent events occur.
Population-level summary	Difference between treatment arms in mean change from baseline in physical function score using baseline and Week 28 measurements.

Abbreviations: ITT = intent-to-treat; CI = confidence interval; LS = least squares

This case study is not an endorsement of any singular study design, outcome or analysis; rather, it is meant to demonstrate application of the estimand framework on a COA-based endpoint.

APPENDIX 2: EXAMPLE FROM GENE THERAPY

In December 2017, Luxturna (voretigene neparvovec-rzyl), a gene therapy delivered though subretinal injection, was approved by FDA for the treatment of patients with confirmed biallelic RPE65 mutation-associated retinal dystrophy (information page for Luxturna BLA (biologics license application) 125610 (FDA, 2018), and information page for the October 12, 2017, Cellular, Tissue and Gene Therapies Advisory Committee Meeting (FDA, 2018)), a condition that leads to visual function decline with age, resulting in total blindness in young adulthood. There is no approved pharmacological treatment for this condition, which affects approximately 1,000 to 3,000 patients in the United States. The phase 3 trial that provided the primary evidence of efficacy of Luxturna received the 2017 David Sackett Trial of the Year Award from the Society of Clinical Trials (Evans, 2018).

The open-label, two-center trial randomized 31 eligible subjects in a 2:1 ratio to the Luxturna intervention group or the control (nonintervention) group. Primary and key secondary efficacy endpoints were measured after one year for both groups. The control group was then crossed over to receive the Luxturna intervention. After another year of follow-up, the same efficacy outcomes were collected for crossed-over control subjects and subjects in the original intervention group.

This trial used a novel performance outcome assessment (PerfO) as the primary efficacy endpoint. Biallelic mutations in the RPE65 gene cause a progressive retinal dystrophy characterized by decreased light sensitivity, constricted visual fields, and impaired visual acuity, resulting in poor functional vision, defined as the ability to conduct vision dependent activities of daily living independently. Because traditional mobility metrics do not address the effects of illumination on speed and accuracy of navigation in a standardized and quantitative manner, to evaluate the effect of Luxturna on functional vision the sponsor developed and validated a novel PerfO of mobility, the multiluminance mobility test (MLMT) (Chung et al, 2018), targeted specifically at the treatment effect on retinal dystrophy.

In MLMT, a patient navigated a marked path along a 5-foot by 10-foot obstacle course relying on vision, in varying environmental illuminations, including very low light levels. There were seven light levels, ranging from 1 Lux to 400 Lux, each assigned a score code going from 6 to 0, respectively (Table 6). The patient's MLMT score corresponded to the lowest light level at which the patient completed the course accurately and at a reasonable pace. A score of -1 was assigned to patients who could not pass MLMT at a light level of 400 lux, the highest light level tested.

Table 6: MLMT Illuminance Level, Score Code, and Real-World Examples

Illuminance (Lux)	Score Code	Corresponding Environment
1	6	Moonless summer night; or indoor nightlight
4	5	Cloudless summer night with half-moon; or outdoor parking lot at night
10	4	60 min after sunset in a city setting; or a bus stop at night
50	3	Outdoor train station at night; or inside of illuminated office building stairwell
125	2	30 min before cloudless sunrise; or interior of shopping mall, train or bus at night
250	1	Interior of elevator, library or office hallway
400	0	Office environment; or food court

Adapted from Chung et al, 2018.
Abbreviation: MLMT = multiluminance mobility test

The primary efficacy endpoint was the MLMT score change from the Baseline visit to the Year 1 visit. A positive score change indicated that the patient was able to complete the MLMT at a lower light level. The trial showed that Luxturna treatment led to a clinically meaningful and statistically significant improvement in the ability to navigate independently in lower light conditions compared with control (see Figure 6).

In many aspects, this trial reflects the challenges and opportunities with using a novel PerfO endpoint in a registration trial (Richardson et al, 2019), especially in the context of a rare disease and a new therapeutic class. These challenges and opportunities are also addressed in two recent FDA draft guidance documents on human gene therapies for rare diseases (*Human Gene Therapy for Rare Diseases* (FDA, 2018), and retinal disorders (*Human Gene Therapy for Retinal Disorders* (FDA, 2018)), respectively. In what follows, we summarize the salient features of the Luxturna trial with regard to the use of the MLMT endpoint to demonstrate the efficacy of Luxturna, and general considerations on using a novel PerfO as the primary efficacy endpoint.

- After a phase 1 trial, the sponsor identified the need to develop a novel clinically meaningful PerfO endpoint specific to the treatment effect of Luxturna on the target patient population, and went on to develop and validate the MLMT, in discussion with FDA. This illustrates the importance of carefully designed and conducted early-phase trials in informing the design of late-phase trials.
- The phase 3 trial used a randomized concurrent control group, instead of a single-arm design, despite the limited number of patients potentially eligible for the trial. In general, a lack of adequate information on the natural history of a condition, coupled with a high diversity in clinical manifestations and rates of progression, would call for a randomized concurrent-controlled trial, instead of a single-arm trial, to provide the primary evidence of efficacy. Using a novel PerfO endpoint further adds to the importance of using a randomized concurrent control for comparison with the investigational product.

- The cross-over component, together with a 2:1 randomization ratio, not only potentially increased enrollment, but also provided additional data to strengthen the efficacy conclusion, which was primarily based on the primary analysis comparing the MLMT score change between the two groups one year after randomization. The MLMT score change in the control group one year after crossing-over to receive the intervention showed similar improvement to that observed in the original intervention group one year after randomization (Russell, Bennett, Wellman, et al, 2017a). This design also allowed the observation of the maintenance of the treatment effect in the original intervention group two years after randomization.

- The trial was designed to be open-label, due to various considerations, some of which are listed in FDA draft guidance of gene therapy for retinal disorders (*Human Gene Therapy for Retinal Disorders* (FDA, 2018)). However, it was also designed with considerable focus on mitigating potential biases on endpoint evaluation.

  - MLMT evaluation was masked to the evaluator. Audio and video recordings of MLMT were independently graded by two trained reviewers and an adjudicator, if needed, at a separate time and location from the testing. The reviewers were affiliated with an independent reading center and were masked to treatment group by receiving coded video files that did not reference date or group assignment.

  - To mitigate learning effects, the MLMT used 12 different configurations of the obstacle course of comparable difficulties. Each test was randomly assigned one of the 12 configurations.

- The Package Insert (see information page for Luxturna BLA 125610 (FDA, 2018)) states that "An MLMT score change of two or greater is considered a clinically meaningful benefit in functional vision." In this trial, this threshold of 2 seems to refer to both the difference in the medians between the two trial groups and the within-patient change. In general, however, the between-group difference and a meaningful within-patient change are two distinct concepts. It may be challenging to reach a consensus on the threshold for a meaningful within-patient change, especially for an endpoint based on an ordinal scale.

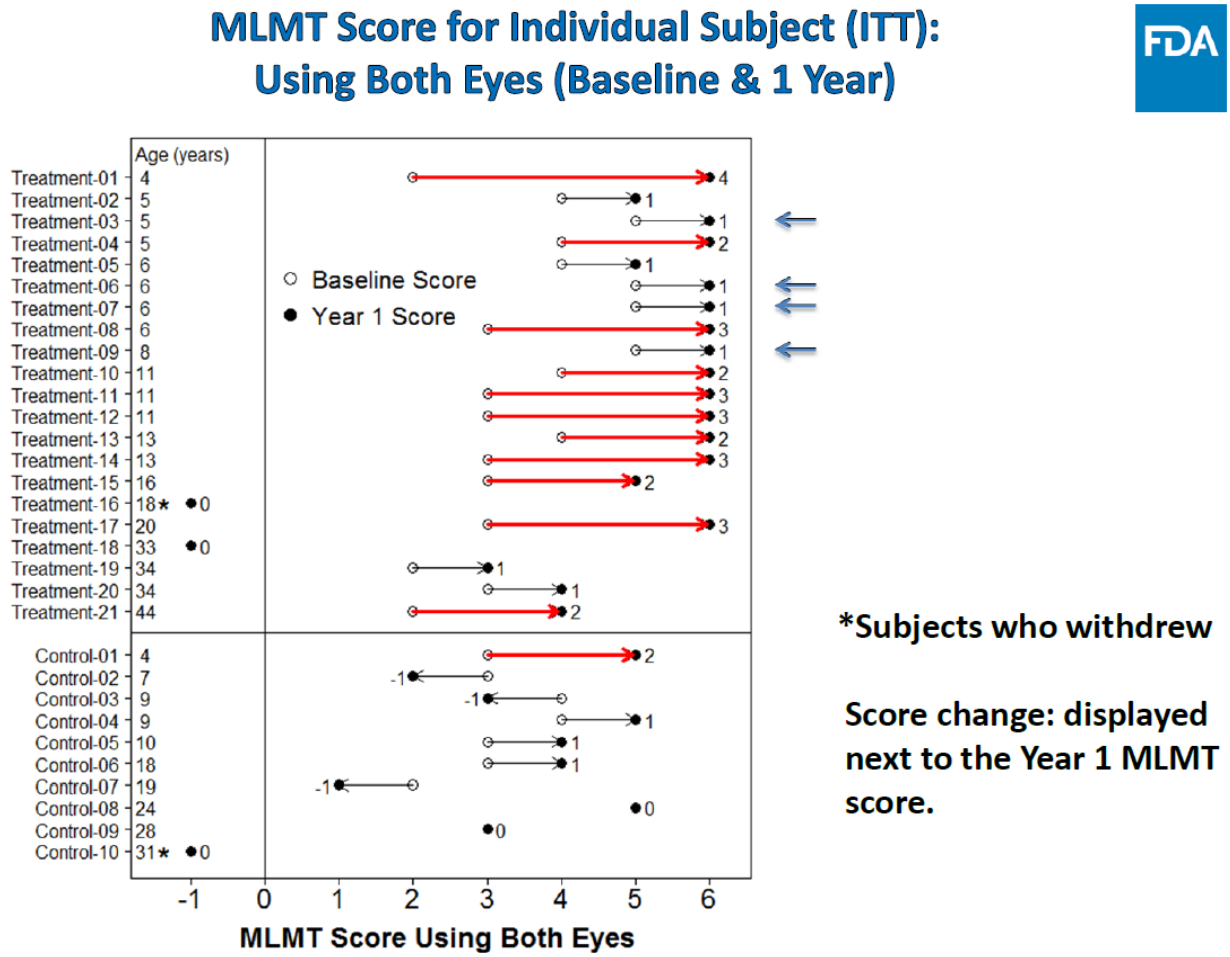
  - Patients entering the trial with a MLMT score of 5 at most could improve by one light level to 6, the highest attainable light score. This ***ceiling effect*** precludes a demonstration of attaining the meaningful within-patient change of at least 2, but it is also important to include these patients to provide data for a broad target population. In the Luxturna trial, all four (4) patients in the Luxturna group with a baseline MLMT score of 5 improved to a score of 6 at the Year 1 visit, consistent with the efficacy result in other patients.

- In the evaluation of treatment efficacy, FDA, applicant, and Advisory Committee also considered supportive evidence from the secondary endpoints (information page for Luxturna BLA 125610 (FDA, 2018), and information page for the October 12, 2017, Cellular, Tissue and Gene Therapies Advisory Committee Meeting (FDA, 2018)). Russell and colleagues (Russell, Bennett, Wellman, et al, 2017b) considered the unavailability of traditional bilateral best-corrected visual acuity data to be a limitation of the trial. In the case of a primary endpoint that is novel to most clinicians, and/or when the primary endpoint does not comprehensively capture the potential impact of a treatment on the disease, it is important to include secondary endpoints that are directly

interpretable to clinicians and that characterize the treatment effect more fully. Care should be taken in the study design and conduct to collect good quality data on the secondary endpoints, ideally with the same care as afforded the primary efficacy endpoint.

- The primary endpoint, score change in the MLMT test, is on an ordinal scale, because the log-unit illuminance levels are not evenly spaced (Table 6) (information page for Luxturna BLA 125610 (FDA, 2018), and information page for the October 12, 2017, Cellular, Tissue and Gene Therapies Advisory Committee Meeting (FDA, 2018)). Some of the statistical analyses used for this endpoint have an interpretation only for variables of an interval scale, e.g., a mean difference between the two trial groups and the corresponding confidence interval. Other analyses have an interpretation for ordinal variables as well, e.g., median and Wilcoxon rank sum tests. For this particular example, it is unclear whether treating the primary endpoint as an interval-scale variable is reasonable. While the endpoint scores do not correspond to evenly spaced log-illuminance level, they are aligned to real-world ambiance illumination that one can relate to. In general, statistical methods, including choice of effect parameters and effect size estimators, should correspond to the scale of the endpoint.

Figure 6: MLMT Scores in Phase 3 Trial



Excerpted from FDA's presentation at the advisory committee meeting (2).
Abbreviations: MLMT = multiluminance mobility test; ITT = intent-to-treat

References Specific to Appendix 2

Chung DC, McCague S, Yu ZF, et al. Novel mobility test to assess functional vision in patients with inherited retinal dystrophies. *Clin Exp Ophthalmol*. 2018;46(3):247-259.

Evans S. The 2017 David Sackett Trial of the Year Award. Society of Clinical Trials Newsletter, June 2018, Volume 29, #1.

FDA. Information page for Luxturna BLA 125610, including package insert and approval supporting documents. Available at: https://www.fda.gov/vaccines-blood-biologics/cellular-gene-therapy-products/luxturna. Content current as of 07/26/2018. Accessed 11/1/2019.

FDA. Information page for the Cellular, Tissue and Gene Therapies Advisory Committee meeting, October 12, 2017, on the Luxturna BLA discussion. Includes package insert, briefing documents, presentations, and summary minutes. Available at: https://www.fda.gov/advisory-committees/cellular-tissue-and-gene-therapies-advisory-committee/2017-meeting-materials-

cellular-tissue-and-gene-therapies-advisory-committee. Content current as of 03/23/2018. Accessed 11/1/2019.

FDA Guidance for Industry. Human Gene Therapy for Rare Diseases. Draft, July 2018. Available at: https://www.fda.gov/media/113807/download. Accessed 11/1/2019.*

FDA Guidance for Industry. Human Gene Therapy for Retinal Disorders. Draft, July 2018. Available at: https://www.fda.gov/media/124641/download. Accessed 11/1/2019.*

Richardson E, Burnell J, HR Adams HR, et al. Developing and implementing performance outcome assessments: evidentiary, methodologic, and operational considerations. *Ther Innov Regul Sci*. 2019;53(1):146-153.

Russell S, Bennett J, Wellman JA, et al. Phase 3 trial update of voretigene neparvovec in biallelic RPE65-mediated inherited retinal disease. Presentation given at the American Academy of Ophthalmology meeting; 2017a; New Orleans.

Russell S, Bennett J, Wellman JA, et al. Efficacy and safety of voretigene neparvovec (AAV2-hRPE65v2) in patients with RPE65-mediated inherited retinal dystrophy: a randomised, controlled, open-label, phase 3 trial. *Lancet*. 2017b;390(10097):849-860. (Including a 17-page "Supplementary Appendix" accessed online.)

* When finalized, this guidance will represent FDA's current thinking on this topic.

APPENDIX 3: REFERENCES

American Educational Research Association; American Psychological Association; National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: American Educational Research Association, 2014.

American Psychological Association (2018). Retrieved July 16, 2019, from APA Dictionary of Psychology. Available at: https://dictionary.apa.org.

Calvert M, Blazeby J, Altman DG, et al. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO Extension. *JAMA*. 2013;309(8):814-822.

Calvert M, Kyte D, Mercieca-Bebber R, et al: the SPIRIT-PRO Group. Guidelines for inclusion of patient-reported outcomes in clinical trial protocols: the SPIRIT-PRO Extension. *JAMA*. 2018;319(5):483-494.

Duke Margolis Center for Health Policy. Report of an event on April 4, 2017. Clinical outcome assessments: establishing and interpreting meaningful within-patient change. Available at: https://healthpolicy.duke.edu/events/clinical-outcome-assessments-establishing-and-interpreting-meaningful-within-patient-change. Accessed 11/1/2019.

FDA. Discussion Document for the Patient-Focused Drug Development Public Workshop on Guidance 3: Select, Develop or Modify Fit-for-Purpose Clinical Outcome Assessments. Prepared for a Patient-Focused Drug Development Guidance Public Workshop held on October 15-16, 2018. Available at: https://www.fda.gov/media/116277/download. Accessed 11/1/2019.

FDA. Guidance for Industry. Multiple Endpoints in Clinical Trials. Draft, January 2017. Available at: https://www.fda.gov/media/102657/download. Accessed 11/1/2019.

FDA. Guidance for Industry. Non-Inferiority Clinical Trials to Establish Effectiveness. November 2016. Available at: https://www.fda.gov/media/78504/download. Accessed 11/1/2019.

FDA Guidance for Industry. Rare Diseases: Natural History Studies for Drug Development. Draft, March 2019. Available at: https://www.fda.gov/media/122425/download. Accessed 11/1/2019.

FDA Guidance for Industry. Rare Diseases: Common Issues in Drug Development. Draft, January 2019. Available at: https://www.fda.gov/media/119757/download. Accessed 11/1/2019.

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. Available at: https://www.fda.gov/media/99447/download. Accessed 11/1/2019.

Goldberg TE, Harvey PD, Wesnes KA, Snyder PJ, Schneider LS. Practice effects due to serial cognitive assessment: implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimers Dement (Amst)*. 2015;1(1):103-111.

ICH Guideline E8(R1). General Considerations for Clinical Studies. May 2019. Available at: https://database.ich.org/sites/default/files/E8-R1_EWG_Draft_Guideline.pdf. Accessed 11/1/2019.

ICH Guideline E9(R1). Estimands and Sensitivity Analysis in Clinical Trials. June 2017. Available at: https://database.ich.org/sites/default/files/E9-R1_EWG_Draft_Guideline.pdf. Accessed 11/1/2019.

ICH E10. Choice of Control Group and Related Issues in Clinical Trials. July 2000. Available at: https://database.ich.org/sites/default/files/E10_Guideline.pdf. Accessed 11/1/2019.

Jones RN. Practice and retest effects in longitudinal studies of cognitive functioning. *Alzheimers Dement (Amst)*. 2015;1(1);101-102.

Lachin JM. Worst-rank score analysis with informatively missing observations in clinical trials. *Control Clin Trials*. 1999;20(5):408-422.

National Research Council (US) Panel on Handling Missing Data in Clinical Trials. Washington, DC: National Academies Press (US), 2010.

Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Belmont, CA: Wadsworth Cengage Learning, 2002.

Song MK, Ward SE. Assessment effects in educational and psychosocial intervention trials: an important but often-overlooked problem. *Res Nurs Health.* 2015;38(3):241-247.

Test Design and Development. In A. E. Association, A. P. Association, & N. C. Education, *Standards for Educational and Psychological Testing* (pp. 75-93). Washington, DC: American Educational Research Association, 2014.

APPENDIX 4: GLOSSARY

This glossary defines terms that will be used in the series of methodological patient-focused drug development (PFDD) FDA guidance documents that are required by the 21st Century Cures Act, and part of commitments made by FDA under the sixth authorization of the Prescription Drug User Fee Act (PDUFA VI). The goal of this glossary is to provide standardized nomenclature and terminologies related to patient-focused medical product development. As appropriate, definitions from existing federal resources (e.g., Biomarkers, EndpointS, and Other Tools (BEST) Resource)[11] have been incorporated into this glossary. External resources were also used to define terms and are cited.

Ability to Detect Change: Evidence that a COA can identify differences in scores over time in individuals or groups who have changed with respect to the measurement concept.

Alternate Forms (also referred to as parallel forms or equivalent forms): 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" (Test Design and Development, 2014).

Benefit: Benefits are the favorable effects of a medical product. Types of benefit include clinical benefit (see clinical benefit). Benefits may also include important characteristics of the medical product, such as convenience (e.g., a more convenient dosing regimen or route of administration) that may lead to improved patient compliance, or benefits that affect those other than the patient. (Source: International Conference on Harmonisation (ICH) guideline *Revision of M4E Guideline on Enhancing the Format and Structure of Benefit-Risk Information in ICH* (Efficacy – M4E(R2)), available at http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/CTD/M4E_R2_Efficacy/M4E_R2Step_4.pdf; ANSI/AAMI/ ISO 14971: 2007/(R)2016 Medical devices—Application of risk management to medical devices.)

Caregiver: A person who helps a patient with daily activities, health care, or any other activities that the patient is unable to perform because of illness or disability, and who understands the patient's health-related needs. This person may or may not have decision-making authority for the patient and is not the patient's health care provider.

Case Report Form: A form used throughout clinical trials to record data collected from subjects in the trial. The form captures all the information specified in the trial's protocol for each subject. All data recorded on the form must be verifiable from original source documentation.

Ceiling Effect: A ceiling effect can occur at the item level or at the scale score level. An item level ceiling effect is observed when a large concentration of participants endorses the highest response category within an item. A scale score level ceiling effect is observed when a large concentration of participants' scores fall at or near the upper limit of the scale score of the instrument. Either situation may occur when the upper extreme of the concept(s) assessed by

[11] Available at https://www.ncbi.nlm.nih.gov/books/NBK338448/

item response categories or by the scale score of the instrument does not sufficiently match the level of the upper extreme of the target patient population.

Clinical Benefit: A positive clinically meaningful effect of an intervention (i.e., a positive effect on how an individual feels, functions, or survives). (Source: BEST (Biomarkers, EndpointS, and other Tools) Resource)

Clinical Outcome: A positive clinically meaningful effect of an intervention (i.e., a positive effect on how an individual feels, functions, or survives). (Source: BEST (Biomarkers, EndpointS, and other Tools) Resource)

Clinical Outcome Assessment (COA): Assessment of a clinical outcome can be made through report by a clinician, a patient, a nonclinician observer, or through a performance- based assessment. Types of COAs include: patient-reported outcome (PRO) measures, clinician-reported outcome (ClinRO) measures, observer-reported outcome (ObsRO) measures, and performance outcome (PerfO) measures. (Source: BEST (Biomarkers, EndpointS, and other Tools) Resource)

Clinician-Reported Outcome (ClinRo): A measurement based on a report that comes from a trained health-care professional after observation of a patient's health condition. Most ClinRO measures involve a clinical judgment or interpretation of the observable signs, behaviors, or other manifestations related to a disease or condition. ClinRO measures cannot directly assess symptoms that are known only to the patient (e.g., pain intensity). (Source: BEST (Biomarkers, EndpointS, and other Tools) Resource)

Clinical Study: Research according to a protocol involving one or more human subjects to evaluate biomedical or health-related outcomes, including interventional studies and observational research. (Source: https://www.federalregister.gov/documents/2016/09/21/2016-22129/clinical-trials-registration-and-results-information-submission#p-1195)

Cognitive Interviews: A qualitative research process used to determine whether concepts and items are understood by respondents in the same way that instrument developers intend. Cognitive interviews involve incorporating follow-up questions in a field test interview to gain a better understanding of how respondents interpret questions/tasks asked of them. In this method, respondents are often asked to think aloud and describe their thought processes as they answer the instrument questions. Respondents should reflect the target population who will be responding to the instrument during the study.

Competing Risks: A competing risk is an event whose occurrence precludes the occurrence of the primary event of interest.

Concept (also referred to as concept of interest): In a regulatory context, the concept is the aspect of an individual's clinical, biological, physical, or functional state, or experience that the assessment is intended to capture (or reflect). (Source: BEST (Biomarkers, EndpointS, and other Tools) Resource)

Construct Validity: Evidence that relationships among items, domains, and concepts conform to a priori hypotheses concerning logical relationships that should exist with other measures or characteristics of patients and patient groups**.**

Content Validity: Evidence from qualitative research demonstrating that an instrument measures the concept of interest, including evidence that the items and domains of an instrument are appropriate and comprehensive relative to its intended measurement concept, population, and use. Testing other measurement properties will not replace or rectify problems with content validity.

Context of Use: A statement that fully and clearly describes the way a medical product development tool is to be used and the medical product development-related purpose of the use. (Source: BEST (Biomarkers, EndpointS, and other Tools) Resource)

Digital Health Technologies (DHTs): Use of computing platforms, connectivity, software and/or sensors for healthcare and related uses. These technologies span a range of products, from general wellness applications to medical devices. These products are also used as diagnostics, therapeutics or adjuncts to medical products (devices, drugs, and biologics). They may also be used to develop or study medical products.

Disease Burden (also referred to as burden of disease): The impacts, direct and indirect, of the patient's health condition that have a negative effect on his or her health, functioning, and overall well-being. Disease burden includes but is not limited to the physical and physiologic impacts of the disease and its symptoms; co-morbidities; emotional and psychological effects of the disease, its management, or its prognosis; social impacts; effects on relationships; impacts on the patient's ability to care for self and others; time and financial impacts of the disease and its management; and considerations of the impacts on the patient's family.

Domain: A sub-concept represented by a score of an instrument that measures a larger concept comprised of multiple domains. For example, psychological function is the larger concept containing the domains subdivided into items describing emotional function and cognitive function.

Endpoint: A precisely defined variable intended to reflect an outcome of interest that is statistically analyzed to address a particular research question. A precise definition of an endpoint typically specifies the type of assessments made; the timing of those assessments; the assessment tools used; and possibly other details, as applicable, such as how multiple assessments within an individual are to be combined. (Source: BEST (Biomarkers, EndpointS, and other Tools) Resource)

Estimand: 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. (Source: ICH E9(R1))

Fit-for-Purpose: A conclusion that the level of validation associated with a tool is sufficient to support its context of use. (Source: BEST (Biomarkers, EndpointS, and other Tools) Resource)

Generalizability: The extent to which study findings can be reliably extended to the target population of interest.

Instrument or Tool: An assessment system comprising three essential components: (1) materials for measurement; (2) an assay for obtaining the measurement; and (3) method and/or criteria for interpreting those measurements. (Source: BEST (Biomarkers, EndpointS, and other Tools) Resource)

Item: An individual question, statement, or task (and its standardized response options) that is evaluated or performed by the patient to address a particular concept.

Learning Effect: See ***Practice Effect***

Measurement Properties: All the attributes relevant to the application of a COA including the content validity, construct validity, reliability, and ability to detect change. These attributes are specific to the measurement application and cannot be assumed to be relevant to all measurement situations, purposes, populations, or settings in which the instrument is used.

Multicomponent Endpoint: A within-patient combination of two or more components. In some cases, multiple aspects of a disease may appropriately be combined into a single endpoint, but subsequent analysis of the aspects or components is generally important for an adequate understanding of the drug's effect. In this type of endpoint, an individual patient's evaluation is dependent upon observation of all the specified components in that patient. A single overall rating or status is often determined according to specified rules.

Observer-Reported Outcome (ObsRO): A measurement based on a report of observable signs, events, or behaviors related to a patient's health condition by someone other than that patient or a health professional. Generally, ObsROs are reported by a parent, caregiver, or someone who observes the patient in daily life, and ObsROs are particularly useful for patients who cannot report for themselves (e.g., infants or individuals who are cognitively impaired). An ObsRO measure does not include medical judgment or interpretation. (Source: BEST (Biomarkers, EndpointS, and other Tools) Resource)

Observational Research: A type of nonexperimental social science research technique in which a researcher directly observes ongoing phenomena in a natural setting. In health sciences, this can include, but is not limited to, observing behaviors and disease signs (tremors) in real-world settings and in real-time.

Patient: Any individual with or at risk of a specific health condition, whether the individual currently receives any therapy to prevent or treat that condition. Patients are the individuals who directly experience the benefits and harms associated with medical products.

Practice Effect: Any change or improvement that results from practice or repetition of task items or activities, including repeated exposure to an instrument.

Patient Experience Data: Defined in Title III, section 3001, of the 21st Century Cures Act of 2016, as amended by section 605 of the Food and Drug Administration Reauthorization Act of 2017, and includes data that are collected by any persons and are intended to provide information about patients' experiences with a disease or condition. Patient experience data can be interpreted as information that captures patients' experiences, perspectives, needs, and priorities related to but not limited to (1) the symptoms of their condition and its natural history; (2) the impact of the conditions on their functioning and quality of life; (3) their experience with treatments; (4) input on which outcomes are important to them; (5) patient preferences for outcomes and treatments; and (6) the relative importance of any issue as defined by patients.

Patient-Focused (also referred to as patient-centered): Ensuring that patients' experiences, perspectives, needs, and priorities are meaningfully incorporated into decisions and activities related to their health and well-being.

Patient-Focused Drug Development (also referred to as patient-focused medical product development): A systematic approach to help ensure that patients' experiences, perspectives, needs, and priorities are captured and meaningfully incorporated into the development and evaluation of medical products throughout the medical product life cycle.

Patient Perspective: A type of patient experience data that specifically relates to patients' attitudes or points of view about their condition or its management. Patient perspectives may include, but are not limited to, perceptions, goals, priorities, concerns, opinions, and preferences.

Patient Preference: A statement of the relative desirability or acceptability to patients of specified alternatives or choice among outcomes or other attributes that differ among alternative health interventions. (Source: FDA guidance for industry *Patient Preference Information – Voluntary Submission, Review in Premarket Approval Applications, Humanitarian Device Exemption Applications, and De Novo Requests, and Inclusion in Decision Summaries and Device Labeling*)

Patient-Reported Outcome (PRO): A measurement based on a report that comes directly from the patient (i.e., study subject) about the status of a patient's health condition without amendment or interpretation of the patient's response by a clinician or anyone else. A PRO can be measured by self-report or by interview, provided that the interviewer records only the patient's response. Symptoms or other unobservable concepts known only to the patient (e.g., pain severity or nausea) can only be measured by PRO measures. PROs can also assess the patient perspective on functioning or activities that may also be observable by others. (Source: BEST (Biomarkers, EndpointS, and other Tools) Resource)

Performance Outcome (PerfO): A measurement based on a standardized task performed by a patient that is administered and evaluated by an appropriately trained individual or is independently completed.

Qualitative Research Methods: Methods associated with the gathering, analysis, interpretation, and presentation of narrative information (e.g., spoken or written accounts of experiences,

observations, and events). Qualitative research methods may also include direct observations (e.g., nonverbal communication and behaviors).

Recall Period: The period of time patients, caregivers, or clinicians are asked to consider in responding to a COA item or task. Recall can be momentary (real time) or retrospective of varying lengths.

Reliability: The ability of a COA to yield consistent, reproducible estimates.

Reporter: In research studies designed to collect patient experience data, the reporter is the individual, group of individuals, or entity providing patient experience data. Reporters may be patients, parents, sexual/romantic partners, caregivers, physicians, or other healthcare professionals. Selection of an appropriate reporter in a given research study will depend on the definition of the target patient population of interest. If a patient in the target population can be reasonably expected to reliably self-report, then one would expect the patient herself/himself to be the reporter in that research study.

Research Protocol: A document that describes the background, rationale, objectives, design, methodology, statistical considerations, and organization of a clinical research project. (Source: University of California San Francisco, 2017) A research protocol guides the study and associated data collection and analysis in a productive and standardized manner.

Response Scale: The system of numbers or verbal anchors by which a value or score is derived for an item. Examples include verbal rating scale (VRS), numeric rating scale (NRS), and visual analog scale (VAS).

Risks: Risks are adverse events and other unfavorable effects associated with a medical product. Risks include drug interactions, risks identified in the nonclinical data, risks to those other than the patient (e.g., fetus, those preparing and administering the medical product), and risks based on pharmacologic class or current knowledge of the product. Factors such as potential misuse, abuse, or diversion of the product may also be considered. (Source: ICH guidelines *Efficacy M4E(R2)*)

Score: A number derived from a patient's, caregiver's, or clinician's response to items or tasks in an instrument. A score is computed based on a prespecified, appropriate scoring algorithm and is subsequently used in statistical analyses of clinical trial results. Scores can be computed for individual items, domains, or concepts, or as a summary of items, domains, or concepts.

Scoring Algorithm: A set of prespecified rules to assign numerical value or values to quantify the responses to the instrument. A scoring algorithm may create a single score from a single item or multiple items (e.g., domain score).

Side Effects (also referred to as adverse reactions): Unwanted or unexpected events or reactions to a medical product (Source: https://www.fda.gov/drugs/drug-information-consumers/finding-and-learning-about-side-effects-adverse-reactions)

Sign: Any observable evidence of a disease, health condition, or treatment-related effect. Signs are usually observed and interpreted by the clinician but may be noticed and reported by the patient.

Symptom: Any experience of a disease, health condition, or treatment-related effect that can be known and confirmed only by the patient, and therefore is most reliably assessed by direct patient report.

Target population (also referred to as target patient population, underlying population, or intended population): The group of individuals (patients) about whom one wishes to make an inference.

Task: See ***item***

Treatment Burden (also referred to as burden of treatment): The impacts of a specific treatment or treatment regimen that have a negative impact on a patient's health, functioning, or overall well-being. Treatment burden includes but is not limited to side effects, discomfort, uncertainty about treatment outcomes, dosing and route of administration, requirements, and financial impacts.

Usability Studies: Studies conducted to demonstrate that the device can be used by the intended users without serious errors or problems, for the intended uses and under the expected use conditions.[12]

[12] See guidance for industry *Applying Human Factors and Usability Engineering to Medical Devices*. Definition derived from Human Factors Validation Testing.