Guidance for Industry

Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims

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

December 2009
Clinical/Medical

Guidance for Industry

Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims

Additional copies are available from:

Office of Communications, Division of Drug Information
Center for Drug Evaluation and Research
Food and Drug Administration
10903 New Hampshire Ave., Bldg. 51, rm. 2201
Silver Spring, MD 20993-0002
Tel: 301-796-3400; Fax: 301-847-8714; E-mail: druginfo@fda.hhs.gov
http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm

or

Office of Communication, Outreach, and Development, HFM-40
Center for Biologics Evaluation and Research
Food and Drug Administration
1401 Rockville Pike, Suite 200N, Rockville, MD 20852-1448
Tel: 800-835-4709 or 301-827-1800; E-mail: ocod@fda.hhs.gov
http://www.fda.gov/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/default.htm

or

Office of Communication, Education, and Radiation Programs
Division of Small Manufacturers, International, and Consumer Assistance, HFZ-220
Center for Devices and Radiological Health
Food and Drug Administration
1350 Piccard Drive, Rockville, MD 20850-4307
DSMICA E-mail: dsmica@cdrh.fda.gov
DSMICA Fax: 301-443-8818
(Tel) Manufacturers Assistance: 800-638-2041 or 301-443-6597
(Tel) International Staff: 301-827-3993
http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/default.htm

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

December 2009
Clinical/Medical

TABLE OF CONTENTS

Contains Nonbinding Recommendations

Guidance for Industry[1] Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims

This guidance represents the Food and Drug Administration's (FDA's) current thinking on this topic. It does not create or confer any rights for or on any person and does not operate to bind FDA or the public. You can use an alternative approach if the approach satisfies the requirements of the applicable statutes and regulations. If you want to discuss an alternative approach, contact the FDA staff responsible for implementing this guidance. If you cannot identify the appropriate FDA staff, call the appropriate number listed on the title page of this guidance.

I. INTRODUCTION

This guidance describes how the Food and Drug Administration (FDA) reviews and evaluates existing, modified, or newly created ***patient-reported outcome (PRO) instruments*** used to support ***claims*** in approved medical product labeling.[2] A PRO instrument (i.e., a ***questionnaire*** plus the information and documentation that support its use) is a means to capture PRO data used to measure ***treatment benefit*** or risk in medical product clinical trials. This guidance does not address the use of PRO instruments for purposes beyond evaluation of claims made about a medical product in labeling. This guidance also does not address disease-specific issues. Guidance on clinical trial endpoints for specific diseases can be found on various FDA Web sites.[3]

By explicitly addressing the review issues identified in this guidance, sponsors can increase the efficiency of their discussions with the FDA during the medical product development process, streamline the FDA's review of PRO instrument adequacy and resultant PRO data collected

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

[2] *Labeling*, as used in this guidance, refers to the information about an FDA-approved medical product intended for the clinician to use in treating patients. See 21 CFR 201.56 and 201.57 for regulations pertaining to prescription drug (including biological drug) labeling. Section 201.56 specifically describes the need for labeling that is not false or misleading. See 21 CFR part 801 for medical device labeling. See 21 CFR 606.122 for blood and blood products for transfusion.

[3] See the following FDA Web sites:
http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm (CDER),
http://www.fda.gov/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/default.htm (CBER), and
http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/default.htm (CDRH).

during a clinical trial, and provide optimal information about the patient perspective for use in making conclusions about treatment effect at the time of medical product approval. PRO instrument development is an iterative process and we recognize there is no single correct way to develop a PRO instrument. Different strategies and methods can be used to address FDA review issues.

The Glossary defines many of the terms used in this guidance. Words or phrases found in the Glossary appear in ***bold italics*** at first mention. Specifically, we encourage sponsors to familiarize themselves with the terms ***conceptual framework of a PRO instrument***, ***endpoint model***, and ***content validity***.

FDA's guidance documents, including this guidance, do not establish legally enforceable responsibilities. Instead, guidances describe the Agency's current thinking on a topic and should be viewed only as recommendations, unless specific regulatory or statutory requirements are cited. The use of the word *should* in Agency guidances means that something is suggested or recommended, but not required.

II. BACKGROUND

A PRO is any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else. The outcome can be measured in absolute terms (e.g., severity of a ***symptom***, ***sign***, or state of a disease) or as a change from a previous measure. In clinical trials, a PRO instrument can be used to measure the effect of a medical intervention on one or more ***concepts*** (i.e., the *thing* being measured, such as a symptom or group of symptoms, effects on a particular function or group of functions, or a group of symptoms or functions shown to measure the severity of a health condition).

Generally, findings measured by a well-defined and reliable PRO instrument in appropriately designed investigations can be used to support a claim in medical product labeling if the claim is consistent with the instrument's documented measurement capability. The amount and kind of evidence that should be provided to the FDA is the same as for any other labeling claim based on other data. Use of a PRO instrument is advised when measuring a concept best known by the patient or best measured from the patient perspective. A PRO instrument, like physician-based instruments, should be shown to measure the concept it is intended to measure, and the FDA will review the evidence that a particular PRO instrument measures the concept claimed. The concepts measured by PRO instruments that are most often used in support of labeling claims refer to a patient's symptoms, signs, or an aspect of functioning directly related to disease status. PRO measures often represent the effect of disease (e.g., heart failure or asthma) on health and functioning from the patient perspective.

Claims generally appear in either the Indications and Usage or Clinical Studies section of labeling, but can appear in any section. Regardless of the labeling section, PRO instrument evaluation principles described here apply.

III. EVALUATION OF A PRO INSTRUMENT

The evaluation of a PRO instrument to support claims in medical product labeling includes the following considerations:

- The population enrolled in the clinical trial
- The clinical trial objectives and design
- The PRO instrument's conceptual framework
- The PRO instrument's ***measurement properties***

Because the purpose of a PRO measure is to capture the patient's experience, an instrument will not be a credible measure without evidence of its usefulness from the target population of patients. Sponsors should provide documented evidence of patient input during instrument development and of the instrument's performance in the specific application in which it is used (i.e., population, condition). An existing instrument can support a labeling claim if it can be shown to reliably measure the claimed concept in the patient population enrolled in the clinical trial.

A. Endpoint Model

Sponsors should define the role a PRO ***endpoint*** is intended to play in the clinical trial (i.e., a primary, key secondary, or exploratory endpoint) so that the instrument development and performance can be reviewed in the context of the intended role, and appropriate statistical methods can be planned and applied. It is critical to plan these approaches in what can be called an endpoint model.

Figures 1 and 2 show examples of endpoint models. In Figure 1, a PRO symptom assessment is a secondary endpoint with a physiologic measure as the primary endpoint intended to support an indication for the treatment of Disease X. In this case, the clinical trial would need to succeed on the physiologic endpoint before success could be attained on the secondary endpoints. In Figure 2, a PRO symptom assessment is the primary clinical trial endpoint intended to support an indication for the treatment of symptoms associated with Disease Y and the physical performance and limitation measures would be the key secondary endpoints. PRO instrument adequacy depends on its role and relationships with other clinical trial endpoints as depicted in the endpoint model. The endpoint model explains the exact demands placed on the PRO instrument to attain the evidence to meet the clinical trial objectives and support the targeted claims corresponding to the concepts measured.

Figure 1. Endpoint Model: Treatment of Disease X

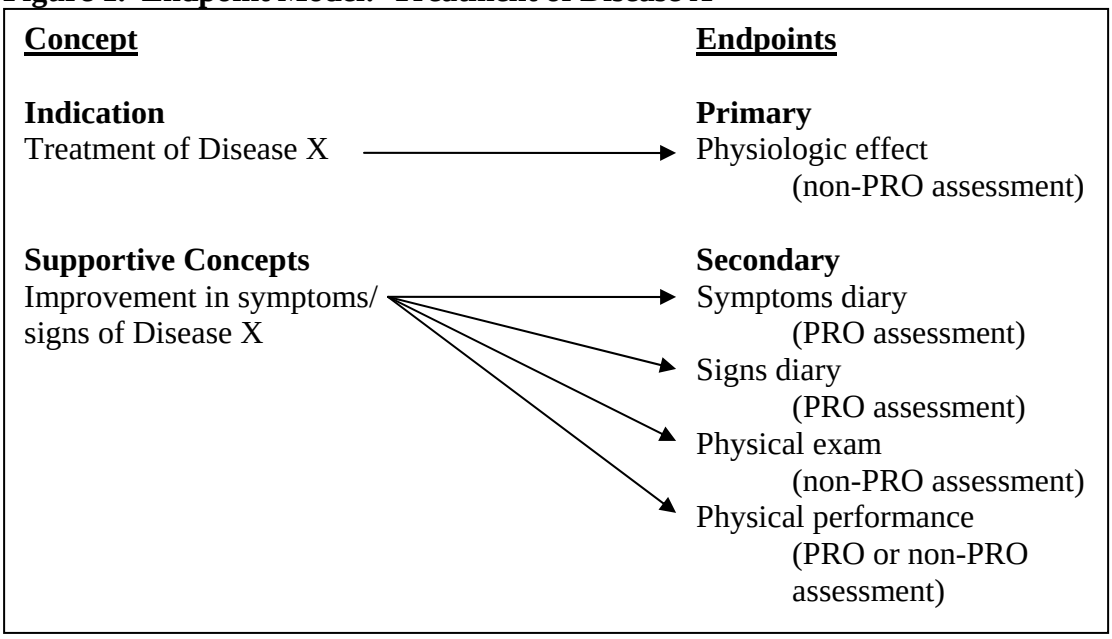
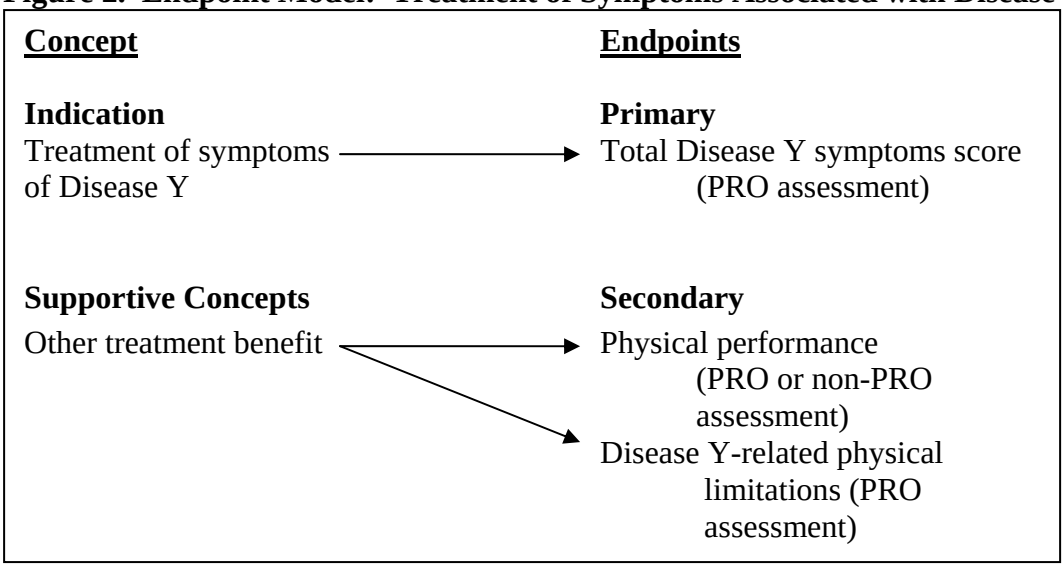


Figure 2. Endpoint Model: Treatment of Symptoms Associated with Disease Y



To help specify potential labeling claims and to facilitate communication with the FDA about the specific clinical trials designed to assess the planned concepts, sponsors can use a ***target product profile (TPP)***, which is a clinical development program summary in the context of prescribing information goals (i.e., targeted labeling claims).[4,5]

[4] See the draft guidance for industry and review staff *Target Product Profile — A Strategic Development Process Tool*. When final, this guidance will represent the FDA's current thinking on this topic. For the most recent version of a guidance, check the FDA Drug guidance Web page at http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm.

[5] Although the TPP process is used for drug and biologic approvals, the concept of beginning with the desired claims and designing the clinical trials to assess these claims is similar for medical devices.

B. Choice of PRO Instrument

Early in medical product development, sponsors planning to use a PRO instrument in support of a labeling claim are encouraged to determine whether an adequate PRO instrument exists to assess and measure the concepts of interest. If it does not, a new PRO instrument can be developed. In some situations, the new instrument can be developed by modifying an existing instrument.

The adequacy of any PRO instrument, whether existing, modified, or newly developed, as a measure to support medical product labeling claims depends on whether its characteristics (see this section), conceptual framework (see section III.C.), content validity (see section III.D.), and other measurement properties (see section III.E.) are satisfactory. The FDA will review documentation of PRO instrument development and testing in conjunction with clinical trial results to determine whether a labeling claim is substantiated. The Appendix lists the type of PRO information sponsors should provide to the FDA to facilitate instrument review.

Characteristics of PRO instruments that are reviewed by the FDA include the following:

- Concepts being measured
- Number of ***items***
- Conceptual framework of the instrument
- Medical condition for intended use
- Population for intended use
- Data collection method
- Administration mode
- Response options
- ***Recall period***
- Scoring
- Weighting of items or ***domains***
- Format
- Respondent burden
- Translation or cultural adaptation availability

We encourage instrument developers to make their instruments and related development history available and accessible publicly. When development history is not available, sponsors generally should provide documentation of content validity with an application (i.e., evidence that the instrument measures what it is intended to measure), including open-ended patient input from the appropriate population. Content validity is discussed in more detail in section III.D., Content Validity. In addition, we anticipate empiric evidence of an instrument's other measurement properties, discussed in more detail in section III.E., Reliability, Other Validity, and Ability to Detect Change.

We suggest that an instrument's measurement properties be well established before enrollment begins for confirmatory clinical trials. Therefore, sponsors should begin instrument development and evaluation early in medical product development, and engage the FDA in a discussion about a new or unique PRO instrument before confirmatory clinical trial protocols are finalized.

Requests for FDA input should be addressed to the review division responsible for the medical product in question. For the FDA to provide useful early input, sponsors should provide their labeling goals, a hypothesized PRO instrument conceptual framework, and the relationship of the PRO endpoints to other clinical trial endpoints in preliminary endpoint models for the planned confirmatory trials.

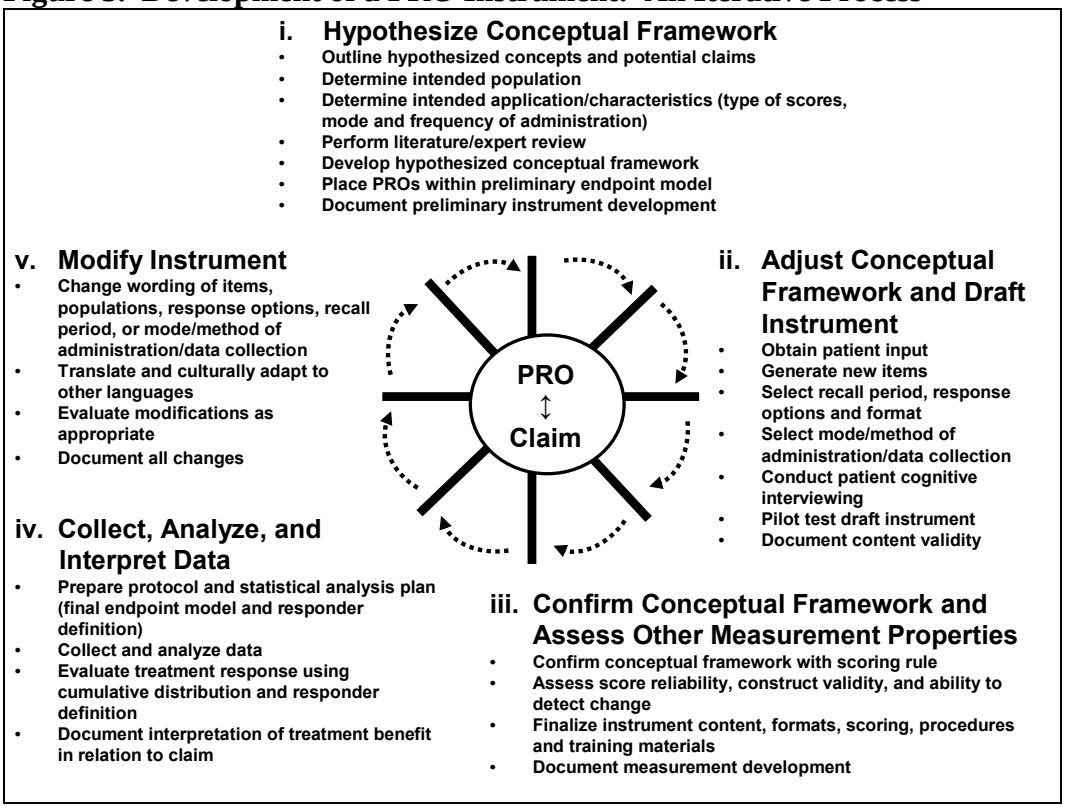
If the measurement goal is to support a complex, multidomain concept, PRO instruments that measure a simple concept may not be adequate to substantiate the complex claim. For example, PRO-based evidence of improved symptoms alone will only support claims specific to improvement of the symptoms and would not support a general claim related to improvement in a patient's ability to function or the patient's psychological state. In addition, a complex, multidomain claim cannot be substantiated by instruments that do not adequately measure the individual component domain concepts adequately.

PRO instruments can be used to measure important safety concerns if those concerns represent symptoms or signs that are best captured from the patient perspective. The principles for PRO instrument development are not different for this application.

Claims representing general concepts often are not supported, even though the PRO instrument was developed to measure the general concepts, because the instrument may not distinguish adverse side effects of treatment that affect the general concept and that may not be known at the time the clinical trials are designed. If adverse effects are captured, PRO instruments should aim to measure the adverse consequences of treatment separately from the effectiveness of treatment. As with any clinical trial evaluating FDA-regulated medical products, all adverse events detected with a PRO instrument should be included in the clinical trial report.

Figure 3 summarizes the iterative process used in developing a PRO instrument for use in clinical trials. FDA review of the developmental process documentation is discussed in more detail in section III.C., Conceptual Framework of a PRO Instrument, through section III.G., PRO Instruments Intended for Specific Populations.

Figure 3. Development of a PRO Instrument: An Iterative Process



C. Conceptual Framework of a PRO Instrument

The adequacy of a proposed instrument to support a claim depends on the conceptual framework of the PRO instrument. The conceptual framework explicitly defines the concepts measured by the instrument in a diagram that presents a description of the relationships between items, domain (subconcepts), and concepts measured and the ***scores*** produced by a PRO instrument.

1. Concepts Measured

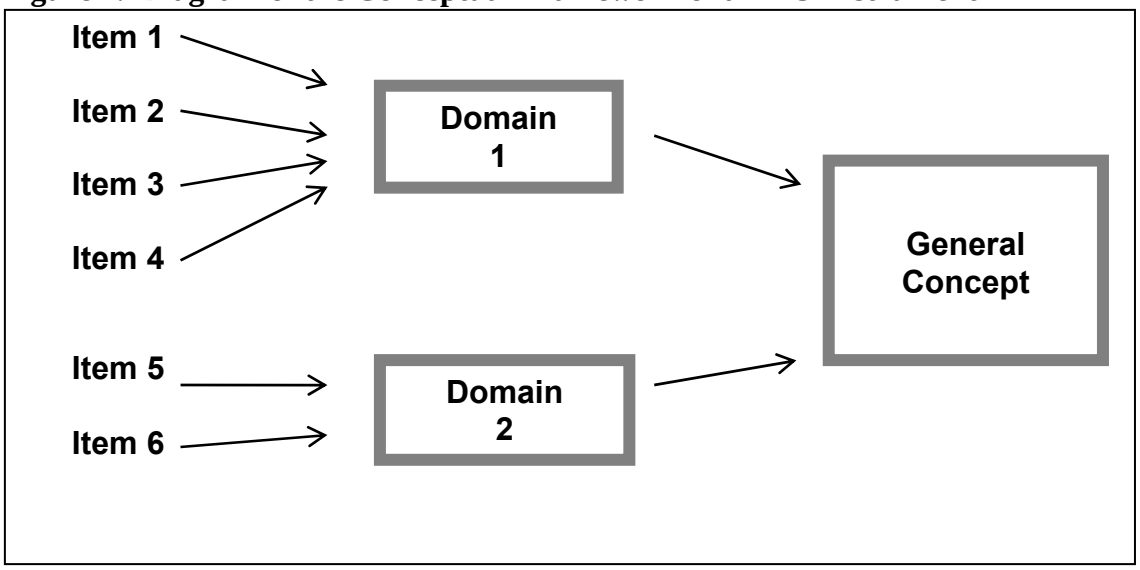
One fundamental consideration in the review of a PRO instrument is the adequacy of the item generation process to support the final conceptual framework of the instrument. In some cases, the question of what to measure may be obvious given the condition being treated. For example, to assess the effect of treatment on pain, patients from the target population are queried about pain severity using a single-item PRO instrument. Generally, when it is not obvious, instrument developers initially can hypothesize a conceptual framework to support the measurement of the concept of interest drafting the domains and items to be measured based on literature reviews and expert opinion. Subsequently, patient interviews, focus groups, and qualitative ***cognitive interviewing*** ensures understanding and completeness of the concepts contained in the items. (See section III.D.1., Item Generation.)

The conceptual framework of a PRO instrument will evolve and be confirmed over the course of instrument development as a sponsor gathers empiric evidence to support item grouping and scores. When used in a clinical trial, the PRO instrument's conceptual framework should again be confirmed by the observed relationships among items and domains.

Documentation of the instrument development process should reveal the means by which the items and domains were identified. The exact words used to represent the concepts measured by domain or total scores should be derived using patient input to ensure the conclusions drawn using instrument scores are valid.

For measures of general concepts, we intend to review how individual items are thought to be associated with each other, how items are associated with each domain, and how domains are associated with each other and the general concept of interest based on the conceptual framework of the PRO instrument. The diagram in Figure 4 depicts a generic example of a conceptual framework of a PRO instrument where Domain 1, Domain 2, and General Concept each represent related but separate concepts. Items in this diagram are aggregated into domains. The final framework is derived and confirmed by measurement property testing.

Figure 4. Diagram of the Conceptual Framework of a PRO Instrument



The conceptual framework of a PRO instrument may be straightforward if a single item is a reliable and valid measure of the concept of interest (e.g., pain intensity). If the concept of interest is general (e.g., physical function), a single-item PRO instrument does not provide a useful understanding of the treatment's effect because a stand-alone single item does not capture the domains of the general concept. For this reason, single-item questions about general concepts that include multiple items or domains rarely provide sufficient evidence to support claims about that general concept. For example, in clinical trials of functional disorders defined by clusters of specific symptoms and signs, a PRO instrument consisting of a single-item global question usually would be inadequate as an endpoint to support labeling claims and would be uninformative about the effects on each specific symptom and sign. Instead, the effect of treatment on each of the appropriate symptoms and signs should be adequately measured.

The conceptual framework for PRO instruments intended to measure a general concept will be complex because identifying all of the appropriate domains and items of the general concept can be difficult. Multidomain PRO instruments can be used to support claims about a general concept if the PRO instrument has been developed to measure the important and relevant

domains of the general concept contained in the claim. However, the complex nature of multidomain PRO instruments often raises significant questions about how to interpret and report results in a way that is not misleading. For example, if improvement in a score for a general concept (e.g., symptoms associated with a certain condition) is driven by a single responsive item (e.g., pain intensity improvement) whereas other important items (e.g., other symptoms) did not show a response, a general claim about the general concept (e.g., improvements in symptoms associated with the condition) cannot be supported. However, that single responsive item or domain may support a claim specific to that item or domain.

We intend to examine the final version of an instrument in light of its development history, including documentation of the complete list of items generated and the reasons for deleting or modifying items, as illustrated in Table 1. We will determine from empiric evidence provided whether the PRO instrument's final conceptual framework (e.g., the hypothesized relationships among items, domains, and concepts measured) is confirmed in the appropriate study population and is consistent with the endpoint model of the planned clinical trials.

Table 1. Common Reasons for Changing Items during PRO Instrument Development

Item Property	Reason for Change or Deletion
Clarity or relevance	• Reported as not relevant by a large segment of the target population • Generates an unacceptably large amount of missing data points • Generates many questions or requests for clarification from patients as they complete the PRO instrument • Patients interpret items and responses in a way that is inconsistent with the PRO instrument's conceptual framework
Response range	• A high percent of patients respond at the floor (response scale's worst end) or ceiling (response scale's optimal end) • Patients note that none of the response choices applies to them • Distribution of item responses is highly skewed
Variability	• All patients give the same answer (i.e., no variance) • Most patients choose only one response choice • Differences among patients are not detected when important differences are known
Reproducibility	• Unstable scores over time when there is no logical reason for variation from one assessment to the next
Inter-item correlation	• Item highly correlated (redundant) with other items in the same concept of interest
Ability to detect change	• Item is not sensitive (i.e., does not change when there is a known change in the concepts of interest)
Item discrimination	• Item is highly correlated with measures of concepts other than the one it is intended to measure • Item does not show variability in relation to some known population characteristics (i.e., severity level, classification of condition, or other known characteristic)
Redundancy	• Item duplicates information collected with other items that have equal or better measurement properties
Recall period	• The population, disease state, or application of the instrument can affect the appropriateness of the recall period

2. Intended Population

Using documentation of the process described in Figure 3 and of the measurement properties as described in Table 2, we plan to compare the patient population studied in the PRO instrument development process to the population enrolled in the clinical trial to determine whether the instrument is applicable for that population. See the Appendix for a description of the types of information sponsors should provide for FDA discussion and review of PRO instruments.

Specific measurement considerations posed by pediatric, cognitively impaired, or seriously ill patients are discussed in section III.G., PRO Instruments Intended for Specific Populations.

Table 2. Measurement Properties Considered in the Review of PRO Instruments Used in Clinical Trials

Measurement Property	Type	What Is Assessed?	FDA Review Considerations
Reliability	Test-retest or intra-interviewer reliability (for interviewer-administered PROs only)	Stability of scores over time when no change is expected in the concept of interest	• Intraclass correlation coefficient • Time period of assessment
	Internal consistency	• Extent to which items comprising a scale measure the same concept • Intercorrelation of items that contribute to a score • Internal consistency	• Cronbach's alpha for summary scores • Item-total correlations
	Inter-interviewer reliability (for interviewer-administered PROs only)	Agreement among responses when the PRO is administered by two or more different interviewers	• Interclass correlation coefficient
Validity	Content validity	Evidence that the instrument measures the concept of interest including evidence from qualitative studies 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.	• Derivation of all items • Qualitative interview schedule • Interview or focus group transcripts • Items derived from the transcripts • Composition of patients used to develop content • Cognitive interview transcripts to evaluate patient understanding
	Construct validity	Evidence that relationships among items, domains, and concepts conform to *a priori* hypotheses concerning logical relationships that should exist with measures of related concepts or scores produced in similar or diverse patient groups	• Strength of correlation testing *a priori* hypotheses (discriminant and convergent validity) • Degree to which the PRO instrument can distinguish among groups hypothesized *a priori* to be different (known groups validity)
Ability to detect change		Evidence that a PRO instrument can identify differences in scores over time in individuals or groups (similar to those in the clinical trials) who have changed with respect to the measurement concept	• Within person change over time • Effect size statistic

D. Content Validity

Content validity is the extent to which the instrument measures the concept of interest. Content validity is supported by evidence from qualitative studies that the items and domains of an instrument are appropriate and comprehensive relative to its intended measurement concept, population, and use. Content validity is specific to the population, condition, and treatment to be studied. For PRO instruments, items, domains, and general scores reflect what is important to patients and comprehensive with respect to patient concerns relevant to the concept being assessed. Documentation of patient input in item generation as well as evaluation of patient understanding through cognitive interviewing can contribute to evidence of content validity. Evidence of other types of validity (e.g., ***construct validity***) or reliability (e.g., consistent scores) will not overcome problems with content validity because we evaluate instrument adequacy to measure the concept represented by the labeling claim. It is important to establish content validity before other measurement properties are evaluated.

When evaluating the utility of an existing instrument or developing a new PRO instrument, sponsors are encouraged to support the adequacy of the instrument's content validity by documenting the following development processes and instrument attributes.

1. Item Generation

Item generation should include input from the target patient population to establish the items that reflect the concept of interest and contribute to its evaluation. The population will help generate item wording, evaluate the completeness of item coverage, and perform initial assessment of clarity and readability. PRO instrument items can be generated from literature reviews, transcripts from focus groups, or interviews with patients, clinicians, family members, researchers, or other sources. We may review whether appropriate individuals and sources were used and how information gleaned from those sources was used in the PRO instrument development process. We will also review whether open-ended patient interviews provide a full understanding of the patient's perspective of the concept of interest.

Item generation generally incorporates the input of a wide range of patients with the condition of interest to represent variations in severity and in population characteristics such as age, sex, ethnicity, and language groups in accordance with the anticipated clinical trial design.

Without adequate documentation of patient input, a PRO instrument's content validity is likely to be questioned. We will review documentation to determine that the items cover all aspects of the concept important to patients, and that ***saturation*** has been reached. Saturation is reached at the point when no new relevant or important information emerges and collecting additional data will not likely add to the understanding of how patients perceive the concept of interest and the items in the questionnaire.

Documentation provided to the FDA to support content validity should include all item generation techniques used, including any theoretical approach; the populations studied; source of items; selection, editing, and reduction of items; cognitive interview summaries or transcripts;

pilot testing; importance ratings; and quantitative techniques for item evaluation. Table 1 lists common reasons for changing items.

If items are not generated in all language groups included in the clinical trials, the appropriateness of the content should be addressed in cognitive interviewing in each language group tested. An ***item tracking matrix*** may be helpful to document the changes or deletions in items and the reasons for those changes.

With existing instruments, it cannot be assumed that the instrument has content validity if patients were not involved in instrument development. New qualitative work similar to that conducted when developing a new instrument can provide documentation of content validity for existing instruments if patient interviews or focus groups are conducted using open-ended methods to elicit patient input. Such qualitative testing of existing instruments is particularly important if a review of the instrument content gives cause for concern. For example, if symptoms known to be common to the population to be studied in the clinical trial are missing from a measure meant to capture important symptoms in that population, we will question the instrument's content validity. We cannot provide recommendations for the number or size of the individual patient interviews or focus groups for establishing content validity. The sample size depends on the completeness of the information obtained from analysis of the transcripts. Generally, the number of patients is not as critical as interview quality and patient diversity included in the sample in relation to intended clinical trial population characteristics.

Items that ask patients to respond hypothetically may cause patients to respond on the basis of their desired condition rather than on their actual condition and therefore are not recommended for clinical trials. For example, in assessing the concept *ability to perform daily activities*, it is more appropriate to ask whether or not the patient performed specific activities (and if so, with how much difficulty) than whether or not the patient perceived that he or she can perform daily activities, because patients may report they are able to perform a task even when they never do the task.

When using multi-item instruments, it is important that all items be relevant to most of the patients in the clinical trial. Using the example in the previous paragraph, it would be severely disadvantageous to use a measure with items that include activities most of the clinical trial patients would not perform. Doing so would yield a *bias toward the null*, or a tendency to show no effect of treatment, even if the treatment were effective. In such cases, a negative response (or indication of little to no activity) is not useful. Use of *not applicable* response options creates problems with scoring. Skip patterns may create difficulties in administration.

2. Data Collection Method and Instrument Administration Mode

Sponsors should consider the data collection method and all procedures and protocols associated with the instrument administration mode, including instructions to interviewers, instructions for self-administration, or instructions for supervising self-administration. We will review data quality control procedures specific to the data collection method or instrument administration mode along with case report forms or screen shots of electronic PRO instruments. Administration modes can include self-administration, interview, or a combination of both. Data

collection methods can include paper-based, computer-assisted, and telephone-based assessments. We intend to review the comparability of data obtained when using multiple data collection methods or administration modes within a single clinical trial to determine whether the treatment effect varies by method or mode. If a patient diary or some other form of unsupervised data entry is used, we plan to review the clinical trial protocol to determine what steps are taken to ensure that patients make entries according to the clinical trial design and not, for example, just before a clinic visit when their reports will be collected.

3. Recall Period

Sponsors should also evaluate the rationale and the appropriateness of the recall period for a PRO instrument. To this end, it is important to consider patient ability to validly recall the information requested. The choice of recall period that is most suitable depends on the instrument's purpose and intended use; the variability, duration, frequency, and intensity of the concept measured; the disease or condition's characteristics; and the tested treatment. When evaluating PRO-based claims, we intend to review the clinical trial protocol to determine what steps were taken to ensure that patients understood the instrument recall period. In many cases, what is of real interest is not the integrated effect over a short time period (e.g., 2-week period), but the effect at regular intervals (e.g., 2, 4, and 6 weeks), similar to how measurements might be made every 2 weeks in a blood pressure trial. In that case, patients can be asked to report on recent status. Note also that any problems created by differential recall are likely to add noise and obscure treatment effects.

PRO instruments that call for patients to rely on memory, especially if they must recall over a long period of time, compare their current state with an earlier period, or average their response over a period of time, are likely to undermine content validity. Response is likely to be influenced by the patient's state at the time of recall. For these reasons, items with short recall periods or items that ask patients to describe their current or recent state are usually preferable. If detailed recall of experience over a period of time is necessary, we recommend the instrument use appropriate methods and techniques for enhancing the validity and reliability of retrospectively reported data (e.g., ask patients to respond based on their worst (or best) experience over the recall period or make use of a diary for data collection).

4. Response Options

It is also important to consider whether the response options for each item are consistent with its purpose and intended use. Table 3 describes some of the various types of item response options that are typically seen in PRO instruments.

Table 3. Response Option Types

Type	Description
Visual analog scale (VAS)	A line of fixed length (usually 100 mm) with words that anchor the scale at the extreme ends and no words describing intermediate positions. Patients are instructed to indicate the place on the line corresponding to their perceived state. The mark's position is measured as the score.
Anchored or categorized VAS	A VAS that has the addition of one or more intermediate marks positioned along the line with reference terms assigned to each mark to help patients identify the locations between the scale's ends (e.g., half-way).
Likert scale	An ordered set of discrete terms or statements from which patients are asked to choose the response that best describes their state or experience.
Rating scale	A set of numerical categories from which patients are asked to choose the category that best describes their state or experience. The ends of rating scales are anchored with words but the categories are numbered rather than labeled with words.
Recording of events as they occur	Specific events are recorded as they occur using an event log that can be included in a patient diary or other reporting system (e.g., interactive voice response system).
Pictorial scale	A set of pictures applied to any of the other response option types. Pictorial scales are often used in pediatric questionnaires but also have been used for patients with cognitive impairments and for patients who are otherwise unable to speak or write.
Checklist	Checklists provide a simple choice between a limited set of options, such as *Yes*, *No*, and *Don't know*. Some checklists ask patients to place a mark in a space if the statement in the item is true. Checklists are reviewed for completeness and nonredundancy.

Item response options generally are considered appropriate when:

- Wording used in responses is clear and appropriate (e.g., anchoring a ***scale*** using the term *normal* assumes that patients understand what is normal for the general population).

- The item response options are appropriate for the intended population. For example, patients with visual impairment may find a VAS difficult to complete.

- Responses offer a clear distinction between choices (e.g., patients may not distinguish between *intense* and *severe* if both are offered as response choices to describe their pain).

- Instructions to patients for completing items and selecting responses for the items are adequate.

- The number of response options is justified empirically (e.g., using qualitative research, initial instrument testing, or existing literature).

- Responses for an item are appropriately ordered and represent similar intervals.

- Responses for items avoid potential ceiling or floor effects (e.g., it may be necessary to introduce more responses to capture worsening or improvement so that fewer patients respond at the response continuum top or bottom).

- Responses do not bias the direction of responses (e.g., bias exists if possible responses are weighted toward the severity spectrum's mild end with two severity options for *mild* and only one each for *moderate* and *severe*).

5. Instrument Format, Instructions, and Training

Results obtained using a PRO instrument can vary according to the instructions given to patients or the training given to the interviewer or persons supervising PRO data collection during a clinical trial. Sponsors should consider all PRO instrument instructions and procedures contained in publications and user manuals provided by developers, including procedures for reviewing completed questionnaires and procedures used to avoid missing data or clarify responses.

It is important that the PRO instrument format used in the clinical trial be consistent with the format that is used during the instrument development process. *Format* refers to the exact questionnaire, diary, or interview script appearance used to collect the PRO data. Format is specific to the administration mode and the data collection method. We plan to review the specific format used in the clinical trial including the order and numbering of items, the presentation of response options in single response or grid formats, the grouping of items, patterns for skipping questions, and all instructions to interviewers or patients.

We recommend that the user manual provided by a developer during the PRO instrument development process specify how to incorporate the instrument into a clinical trial in a way that minimizes administrator burden, patient burden, missing data, and poor data quality. The user manual should explain to investigators and interviewers critical principles of PRO administration.

6. Patient Understanding

When the initial and subsequent drafts of an instrument are prepared, sponsors are encouraged to examine all items and procedures in a pilot test of whether patients understand the items and instructions included in the PRO instrument. This examination should include documentation that the concepts represented in the PRO instrument's conceptual framework are confirmed, that the response options and recall period are appropriately comprehended, and that the instrument's readability is adequate for the intended population. The FDA's evaluation of these procedures is likely to include a review of a cognitive interviewing report containing the script used in patient cognitive interviews, the interview transcripts, the readability test used (if applicable), the ***usability testing*** process description (if applicable), the cognitive interviews analysis, and the actions taken to delete or modify items, response scales, or patient instructions in response to the cognitive interview or pilot test results. Evidence from the patient cognitive interview studies (i.e., the interview schedule, transcript, and listing of all concepts elicited by a single item) can be used to determine when a concept is adequately captured. Repeating cognitive interviews can help confirm content validity.

7. Scoring of Items and Domains

For each item, numerical scores generally should be assigned to each answer category based on the most appropriate scale of measurement for the item (e.g., nominal, ordinal, interval, or ratio

scales). We will review the distribution of item responses to ensure that the response choices represent appropriate intervals.

A scoring algorithm creates a single score from multiple items. We will review the evidence that the summary score is appropriate. Equally weighted scores for each item are appropriate when the responses to the items are independent. If two items are dependent, their collected information is less than two independent items and they are over-weighted when they are treated as two equally weighted items. Over-weighting also may be a concern when the number of response options or the values associated with response options varies by item. The same weighting concerns apply with added complexity when combining domain scores into a single general score. Using qualitative research or defined statistical techniques, sponsors should justify the method chosen to combine items to create a score or to combine domain scores to create a general score.

When empirically determined patient preference ratings are used to weight items or domains, we intend to review the composition of samples and the process used to determine the preference weights. Because preference weights are often developed for use in resource allocation (e.g., as in cost-effectiveness analysis that may use predetermined community weights), it is tempting to use those same weights in the clinical trial setting to demonstrate treatment benefit. However, this practice is discouraged unless the preference weights' relationship to the intended clinical trial population is known and found adequate and appropriate.

Total scores combining multiple domains should be supported by evidence that the total score represents a single albeit complex concept. As described earlier in section III.C., Conceptual Framework of a PRO Instrument, the instrument's final conceptual framework documents the concept represented by each score. If a score is intended to support a targeted claim, the concept measured will match the targeted claim language. Generally, we discourage claims expressed in terms of domain or instrument titles because they often do not represent the concept measured.

8. Respondent and Administrator Burden

Undue physical, emotional, or cognitive strain on patients generally decreases the quality and completeness of PRO data. Factors that can contribute to respondent burden include the following:

- Length of questionnaire or interview
- Formatting
- Font size too small to read easily
- New instructions for each item
- Requirement that patients consult records to complete responses

- Privacy of the setting in which the PRO is completed (e.g., not providing a private space for patients to complete questionnaires containing sensitive information about their sexual performance or substance abuse history)

- Inadequate time to complete questionnaires or interviews

- Literacy level too high for population

- Questions that patients are unwilling to answer

- Perception by patients that the interviewer wants or expects a particular response

- Need for physical help in responding (e.g., turning pages, holding a pen, assistance with a telephone or computer keyboard)

The degree of respondent burden that is tolerable for instruments in clinical trials depends on the frequency and timing of PRO assessments in a protocol and on patient cognition, illness severity, or treatment toxicity. For example, if the questionnaire contains instructions to skip one or more questions based on response to a previous question, respondents may fail to understand what to do and make errors in responding or find the assessment too complicated to complete. Sponsors should consider missing data and the refusal rate as possible indications of inappropriate respondent burden or inappropriate items or response options.

E. Reliability, Other Validity, and Ability to Detect Change

Once the instrument's content validity has been established, we intend to consider the following additional measurement properties during FDA review of a PRO instrument: reliability, construct validity, and ability to detect change. We plan to review the measurement properties that are specific to the documented PRO instrument's conceptual framework, confirmed scoring algorithm, administration procedures, and questionnaire format in light of the clinical trial's objectives, design, enrolled population, and statistical analysis plan (SAP). We also plan to review whether the population and medical conditions included in any sample used to develop or test a PRO instrument are appropriate for the planned clinical trials.

In addition, an adequate study to evaluate any specific measurement property of a PRO instrument should be designed to test a prespecified hypothesis. For example, if the study compares a new PRO measure to an existing measure of the same concept administered during the same interview or within a short time of each other to establish construct validity, the study should be designed to test the hypothesized level of correlation and the results should be discussed in light of that hypothesis.

1. Reliability

Because clinical trials measure change over time, the adequacy of a PRO instrument for use in a clinical trial depends on its reliability or ability to yield consistent, reproducible estimates of true treatment effect.

We will review documentation of tests to determine if reproducibility (e.g., test-retest reliability) has been demonstrated. Test-retest is most informative when the time interval chosen between the test and retest is long enough in stable patients to minimize memory effects. The time interval chosen depends on the variability of the state or experience being evaluated and on the potential for change in the condition or population over time that reflects actual change in the condition rather than variability in stable patients. Test-retest reliability can be tested over a variety of periods to satisfy different study protocols or even in different intervals between visits in the same protocol. We acknowledge that for remitting and relapsing or episodic diseases, test-retest reliability may be difficult or impossible to establish.

Internal consistency reliability tests (e.g., Cronbach's alpha) to determine agreement among responses to different questions, in the absence of test-retest reliability, may not constitute sufficient evidence of reliability for clinical trial purposes. However, as is true for other imperfections in testing, in general, flaws in reliability tend to increase the beta (Type II) error, and instruments demonstrating poor reliability are unlikely to give a false positive result.

When PRO instruments are interviewer-administered, we will review inter-interviewer reproducibility. Inter-interviewer reproducibility depends on instrument administration standardization and interviewer training on this standard.

2. Other Validity

In addition to content validity (discussed in section III.D., Content Validity), we will evaluate evidence of construct validity, and if appropriate, ***criterion validity***.

Construct validity is determined by 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.

We will review the construct validity of an instrument to determine whether the documented relationships between results gathered using the instrument and results gathered using other measures are consistent with pre-existing hypotheses concerning those relationships (i.e., discriminant and convergent validity). We will also review evidence that the instrument can differentiate between clinically distinct groups (i.e., known groups validity).

As stated earlier, single-item questions about general concepts are not useful to support claims; however, they can be useful to help assess the construct validity of multi-item measures of the same concept and to determine whether important items or domains of a general concept are missing. For example, when results using single-item general questions do not correlate with results using a multi-item questionnaire of the same general concept, this may be evidence that the questionnaire is not capturing all the important domains of the general concept.

Criterion validity is the extent to which the scores of a PRO instrument are related to a known *gold standard* measure of the same concept. In rare cases, we will also review the criterion validity of an instrument if a criterion measure is purported for the PRO concept assessed (e.g.,

comparing a new sleep scale to a clinical measure of polysomnography). However, for most PROs, criterion validity testing is not possible because the nature of the concept to be measured does not allow for a criterion measure to exist. This is true for any symptom measure where the symptom is known only to the patient. If a criterion measure is used, sponsors should provide rationale and support for that criterion. We will review the extent to which the PRO measure is correlated with the criterion measure as well as the sensitivity, specificity, and predictive value of the criterion measure.

3. Ability to Detect Change

We will review an instrument's ability to detect change using data that compare change in PRO scores to change in other similar measures that indicate that the patient's state has changed with respect to the concept of interest. A review of the ability to detect change includes evidence that the instrument is equally sensitive to gains and losses in the measurement concept and to change at all points within the entire range expected for the clinical trial population.

When patient experience of a concept is predicted to change, the values for the PRO instrument measuring that concept should change. If there is clear evidence that patient experience relative to the concept has changed, but the PRO scores do not change, then either the ability to detect change is inadequate or the PRO instrument's validity should be questioned. If there is evidence that PRO scores are affected by changes that are not specific to the concept of interest, the PRO instrument's validity may be questioned.

The ability of an instrument to detect change influences the sample size for evaluating the effectiveness of treatment. The extent to which the PRO instrument's ability to detect change varies by important patient subgroups (e.g., sex, race, or age) can affect clinical trial results. If important subgroup differences in ability to detect change are known, these documented differences can be taken into account in assessing results. In general, an inability to detect change tends to support the null hypothesis of no treatment effect.

F. Instrument Modification

The adequacy of an instrument's development and testing is specific to its intended application in terms of population, condition, and other aspects of the measurement context for which the instrument was developed. When a PRO instrument is modified, sponsors generally should provide evidence to confirm the new instrument's adequacy. That is **not** to say that every small change in application or format necessitates extensive studies to document the final version's measurement properties. Additional qualitative work may be adequate depending on the type of modification made. Examples of changes that can alter the way that patients respond to the same set of questions include:

- Changing an instrument from paper to electronic format
- Changing the timing of or procedures for PRO instrument administration within the clinic visit

- Changing the application to a different setting, population, or condition

- Changing the order of items, item wording, response options, or recall period or deleting portions of a questionnaire

- Changing the instructions or the placement of instructions within the PRO instrument

A small nonrandomized study may be adequate to compare the distribution of responses between versions of a questionnaire with different formats (e.g., changing a response scale from vertical to horizontal). If the PRO instrument will be used in a significantly different patient population (e.g., a different disease or age group), we may recommend using qualitative studies to confirm content validity in the new population. A small randomized study to ascertain the measurement properties in the new population may minimize the risk that the instrument will not perform adequately.

G. PRO Instruments Intended for Specific Populations

As previously mentioned, if multiple versions of an instrument will be used in a clinical trial, documentation should exist that the content validity and other measurement properties of those versions are similar to each other. Measurement of PRO concepts in children and adolescents, in patients who have cognitive impairment or are unable to communicate because of serious illness, and across culture or language groups introduces challenges in addition to those already mentioned. These challenges are discussed below.

1. Children and Adolescents

In general, the review issues related to the development process for pediatric PRO instruments are similar to the issues detailed for adults. Additional review issues for PRO instruments applied in children and adolescents include age-related vocabulary, language comprehension, comprehension of the health concept measured, and duration of recall. Instrument development within fairly narrow age groupings is important to account for developmental differences and to determine the lower age limit at which children can understand the questions and provide reliable and valid responses that can be compared across age categories. We discourage ***proxy-reported outcome*** measures for this population (i.e., reports by someone who is not the patient responding as if that person were the patient). For patients who cannot respond for themselves (e.g., infant patients), we encourage observer reports that include only those events or behaviors that can be observed. For example, observers cannot validly report an infant's pain intensity but can report infant behavior thought to be caused by pain.

2. Patients Cognitively Impaired or Unable to Communicate

We discourage proxy-reported outcome measures for this population. For patients who cannot respond for themselves (e.g., cognitively impaired), we encourage observer reports that include only those events or behaviors that can be observed.

3. Culture or Language Subgroups

Because many development programs are multinational, application of PRO instruments to multiple cultures or languages is common in clinical trials. Regardless of whether the instrument was developed concurrently in multiple cultures or languages or whether a fully developed instrument was adapted or translated to new cultures or languages, we recommend that sponsors provide evidence that the content validity and other measurement properties are adequately similar between all versions used in the clinical trial. We will review the process used to translate and culturally adapt the instrument for populations that will use them in the trial.

IV. CLINICAL TRIAL DESIGN

The same clinical trial design principles that apply to other endpoint measures also apply to PROs. Therefore, this section is not a comprehensive overview of those principles but rather focuses primarily on issues unique to PRO endpoints.

A. General Protocol Considerations

If the PRO measurement goal is to support labeling claims, PRO concept measurement should be stated as a specific clinical trial objective or hypothesis. It is important that the case report form in the protocol include the exact format and version of the specific PRO instrument to be administered. If an electronic version of the instrument will be used, the protocol can include screen shots or other similar instrument representations. In the process of considering the new drug application (NDA)/biologics license application (BLA)/medical device premarket approval (PMA) or NDA/BLA/PMA supplement, we intend to compare both the planned and actual PRO instrument used and its analysis.

1. Blinding and Randomization

Open-label clinical trials, where patients and investigators are aware of assigned therapy, are rarely adequate to support labeling claims based on PRO instruments. Patients who know they are in an active treatment group may overestimate benefit whereas patients who know they are not receiving active treatment may underreport any improvement actually experienced. For the same reasons, to prevent influencing patient perceptions, PRO instruments administered during a clinic visit should be administered before other clinical assessments or procedures.

In blinded clinical trials, patients should be blinded to treatment assignment throughout the trial. If the treatment has obvious effects, such as adverse events, the clinical trial may be at risk for unintentional unblinding. In these situations, sponsors can use PRO instrument administration techniques that may minimize the effects of possible unblinding, such as using response options that ask for current status, not giving patients access to previous responses, and using instruments that include many items about the same concept.

Suspicion of inadvertent unblinding can be a problematic review consideration for the FDA when assessing PRO endpoints. Therefore, when PRO instruments are included in a clinical

trial, we encourage sponsors to include a single item during or at the end of the trial to ask patients to identify the clinical trial arm in which they believe they participated.

The effect of intentional unblinding is important to consider in the interpretation of clinical trial results. There are certain situations, such as in the evaluation of some medical devices or administration of identifiable treatment regimens, where blinding is not feasible and other situations where there is no reasonable control group (and therefore no randomization). When a PRO instrument appears useful in assessing patient benefit in those situations, we encourage sponsors to confer with the appropriate review division.

2. Clinical Trial Quality Control

The quality of a clinical trial can be optimized at the design stage by specifying in the protocol procedures to minimize inconsistencies in trial conduct. We recommend a standardized order by which PRO and other clinical assessments are administered. Other examples of standardized instructions and processes that can appear in the protocol include:

- Training and instructions to patients for self-administered PRO instruments
- Interviewer training and interview format for PRO instruments administered in an interview format
- Instructions for the clinical investigators regarding patient supervision, timing and order of questionnaire administration during or outside the office visit, processes and rules for questionnaire review for completeness, and documentation of how and when data are filed, stored, and transmitted to or from the clinical trial site
- Plans for confirmation of the instrument's measurement properties using clinical trial data

3. Handling Missing Data

Sometimes patients fail to report for visits, fail to complete questionnaires, or withdraw from a clinical trial before its planned completion. The resulting missing data can introduce bias and interfere with the ability to compare effects in the test group with the control group because only a subset of the initial randomized population contributes, and these patient groups may no longer be comparable. Missing data is a major challenge to the success and interpretation of any clinical trial. The clinical trial protocol should describe how missing data will be handled in the analysis.

The protocol can increase the likelihood that a clinical trial will still be informative by establishing backup plans for gathering all treatment-related reasons for patients failing to report at scheduled times or withdrawing from a treatment or the clinical trial and by trying to minimize patient dropouts before trial completion. Patients should remain in the clinical trial, even if they have discontinued treatment, and should continue to provide PRO data. The protocol should also

establish a process by which PRO measurement is obtained before or shortly after patient withdrawal from treatment should early withdrawal be unpreventable.

B. Frequency of Assessments

The frequency of PRO assessment should correspond with the specific research questions being addressed, length of recall asked by the instrument's response options, demonstrated instrument measurement properties, the disease or condition's natural history, the treatment's nature, and planned data analysis. Some diseases, conditions, or clinical trial designs may necessitate more than one baseline assessment and several PRO assessments during treatment.

C. Clinical Trial Duration

The duration of PRO assessment depends on the PRO research questions being posed. It is important to consider whether the clinical trial's duration is of adequate length to support the proposed claim and assess a durable outcome in the disease or condition being studied. Generally, duration of follow-up with a PRO assessment should be the same as for other measures of effectiveness. However, the clinical trial duration appropriate for the PRO-related objective may not be the same duration as for other endpoints.

D. Design Considerations for Multiple Endpoints

A single hierarchy of endpoints as diagrammed in an endpoint model (see Figures 1 and 2 in section III.A., Endpoint Model) is determined by the trial's stated objectives and the clinical relevance and importance of each specific measure independently and in relationship to each other. We consider any endpoints that are not part of the prespecified hierarchy of primary and key secondary endpoints to be exploratory. Endpoints included for economic evaluation that are not intended for labeling claims should be designated as such, and will be regarded as exploratory. A PRO measurement can be the clinical trial's primary endpoint measure, a co-primary endpoint measure in conjunction with another PRO measure, other clinical endpoints or physician-rated measurements, or a secondary endpoint measure whose analysis is considered according to a hierarchical sequence. It is critical that the clinical trial protocol define the endpoint measures and the criteria for the statistical analysis and interpretation of results, including a specification of the conditions for a positive clinical trial conclusion, because determination of these criteria and conditions after data are unblinded will not be credible. Sponsors should avoid separate consideration of PRO endpoints from the clinical trial's primary objectives in terms of clinical trial design or data analysis. Sponsors also should avoid *cherry picking* or post hoc selective picking of PRO endpoint results for inclusion in proposed labeling.

E. Planning for Clinical Trial Interpretation Using a Responder Definition

Regardless of whether the primary endpoint for the clinical trial is based on individual responses to treatment or the group response, it is usually useful to display individual responses, often using an *a priori* ***responder definition*** (i.e., the individual patient PRO score change over a predetermined time period that should be interpreted as a treatment benefit). The responder definition is determined empirically and may vary by target population or other clinical trial

design characteristics. Therefore, we will evaluate an instrument's responder definition in the context of each specific clinical trial.

The empiric evidence for any responder definition is derived using anchor-based methods. Anchor-based methods explore the associations between the targeted concept of the PRO instrument and the concept measured by the anchors. To be useful, the anchors chosen should be easier to interpret than the PRO measure itself. For example, the number of incontinence episodes collected in incontinence diaries has been used to determine a responder definition for PRO instruments assessing the annoyance of incontinence. A 50 percent reduction in incontinence episodes might be proposed as the anchor for defining a responder on the PRO instrument. Confirmation of this anchor approach in early clinical trials can provide the basis for the proposed responder definition in the confirmatory trials.

Another anchor-based approach to defining responders makes use of patient ratings of change administered at different periods of time or upon exit from a clinical trial. These numerical ratings range from *worse* to *the same* and *better*. The difference in the PRO score for persons who rate their condition *the same* and *better* or *worse* can be used to define responders to treatment. Patient ratings of change are less useful as anchors when patients are not blinded to treatment assignment.

Another set of approaches to defining a responder are distribution-based methods that use, for example, the between-person standard deviation or the standard error of measurement to define a meaningful change on a scale. Distribution-based methods can be used to categorize these changes as small, moderate, and large and often can be combined with anchor-based estimates to provide confidence in the responder definition. Distribution-based methods for determining clinical significance of particular score changes should be considered as supportive and are not appropriate as the sole basis for determining a responder definition.

Alternatively, it is possible to present the entire distribution of responses for treatment and control group, avoiding the need to pick a responder criterion. Whether the individual responses are meaningful represents a judgment, but that problem is present with almost all endpoints except survival. Such cumulative distribution displays show a continuous plot of the percent change from baseline on the X-axis and the percent of patients experiencing that change on the Y-axis. This display type may be preferable to attempting to provide categorical definitions of *responders*. A variety of responder definitions can be identified along the cumulative distribution of response curve.

Guidance on interpretation considerations for a clinical trial's SAP is found in section V.E., Interpretation of Clinical Trial Results.

F. Specific Concerns When Using Electronic PRO Instruments

When PRO instruments are used, sponsors must ensure that FDA regulatory requirements are met for sponsor and investigator record keeping, maintenance, and access.[6] These responsibilities are independent of the method used to record clinical trial data and, therefore, apply to all types of PRO data including electronic PRO data. Sponsors are responsible for providing investigators with all information to conduct the investigation properly, for monitoring the investigation, for ensuring that the investigation is conducted in accordance with the investigational plan, and for permitting the FDA to access, copy, and verify records and reports relating to the investigation.

The principal record keeping requirements for clinical investigators include the preparation and maintenance of adequate and accurate case histories (including the case report forms and supporting data), record retention, and provision for the FDA to access, copy, and verify records (i.e., source data verification). The investigator's responsibility to control, access, and maintain source documentation can be satisfied easily when paper PRO instruments are used, because the patient usually returns the diary to the investigator who either retains the original or a certified copy as part of the case history. The use of electronic PRO instruments, however, may pose a problem if direct control over source data is maintained by the sponsor or the contract research organization and not by the clinical investigator. We consider the investigator to have met his or her responsibility when the investigator retains the ability to control and provide access to the records that serve as the electronic source documentation for the purpose of an FDA inspection. The clinical trial protocol, or a separate document, should specify how the electronic PRO source data will be maintained and how the investigator will meet the regulatory requirements.

In addition, the FDA has previously provided guidance to address the use of computerized systems to create, modify, maintain, archive, retrieve, or transmit clinical data to the FDA and to clarify the requirements and application of 21 CFR part 11.[7,8] Because electronic PRO data (including data gathered by personal digital assistants or phone-based interactive voice recording systems) are part of the case history, electronic PRO data should be consistent with the data standards described in that guidance. Sponsors should plan to establish appropriate system and security controls, as well as cyber-security and system maintenance plans that address how to ensure data integrity during network attacks and software updates.

Sponsors also should avoid the following:

- Direct PRO data transmission from the PRO data collection device to the sponsor, clinical investigator, or other third party without an electronic audit trail that documents all changes to the data after it leaves the PRO data collection device.

[6] For the principal record keeping requirements for clinical investigators and sponsors developing drugs and biologics, see 21 CFR 312.50, 312.58, 312.62, and 312.68. For medical devices, see 21 CFR 812.140 and 812.145.

[7] See the guidance for industry *Computerized Systems Used in Clinical Investigations* (http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm).

[8] See the guidance for industry *Part 11, Electronic Records; Electronic Signatures — Scope and Application* (http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm).

- Source document control by the sponsor exclusively.

- Clinical investigator inability to maintain and confirm electronic PRO data accuracy. The data maintained by the clinical investigator should include an audit trail to capture any changes made to the electronic PRO data at any point in time after it leaves the patient's electronic device.

- The existence of only one database without backup (i.e., risk of data corruption or loss during the trial with no way to reconstitute or verify the data).

- Ability of any entity other than the investigator (and/or site staff designated by the investigator) to modify the source data.

- Loss of adverse event data.

- Premature or unplanned access to unblinded data.

- Inability of an FDA investigator to inspect, verify, and copy the data at the clinical site during an inspection.

- An insecure system where records are easily altered.

- Direct PRO data transmission of important safety information to sponsors, clinical research organizations, and/or third parties, without ensuring the timely transmission of the data to the clinical investigator responsible for the patients.

V. DATA ANALYSIS

Incorporating PRO instruments as clinical trial endpoint measures introduces challenges in the analysis of clinical trial data. The most important of these challenges are discussed in the following sections.

A. General Statistical Considerations

The statistical analysis considerations for PRO endpoints are not unlike statistical considerations for any other endpoint used in medical product development.[9] Every protocol should describe the principal data analysis features in the statistical section with a detailed elaboration of the analysis in an SAP. We intend to determine the adequacy of clinical trial data to support claims in light of the prespecified method for endpoint analysis. We usually view unplanned or post hoc statistical analyses conducted after unblinding as exploratory and, therefore, unable to serve as the basis of a labeling claim of effectiveness.

[9] See the ICH guidance for industry *E9 Statistical Principles for Clinical Trials* (http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm).

B. Statistical Considerations for Using Multiple Endpoints

PROs in a clinical trial, like non-PRO clinical endpoints, can be primary or secondary endpoints. Primary endpoints are those endpoints on which the main benefit of a clinical trial's test treatment is judged rigorously. Primary endpoints are used to determine the clinical trial sample size and are the endpoints that will be tested statistically. They are clinically meaningful but may not be the most important endpoint because choice of clinical trial endpoints is a complex evolution of expected effect size, expected number of events, and other factors.

There are often multiple endpoints that would be of clinical interest. Analysis of multiple endpoints, where an effect on any of the endpoints will be considered evidence of effectiveness, can inflate the probability of false positive findings known as the Type I error rate, an inflation that can be controlled by a prospectively planned multiplicity adjustment. It is common to analyze secondary endpoints only after success on a primary endpoint. This can be done using a sequential analysis, testing additional endpoints in a defined sequence each at the usual alpha = 0.05 level of statistical significance. The analyses cease when a failure occurs. It is important that the clinical trial protocol specify all primary and secondary endpoints. The SAP should describe the planned primary analysis in detail noting whether the endpoint will be analyzed as a continuous variable (mean scores), dichotomous variable (success or failure), or some graded response; the primary and secondary endpoints; adjustments for multiplicity to control the overall Type I error rate; and the specific statistical methods planned. Sponsors should provide the FDA with the clinical trial's SAP for review.

Cases arise in clinical trials where a clinically meaningful treatment benefit depends on having two or more primary endpoints achieving statistical significance at a specified alpha level (e.g., alpha = 0.05). For example, a clinical trial may identify two endpoints with a decision rule that each should show that the treatment is better than control. Such a decision rule does not require multiplicity adjustment because the maximum Type I error rate (alpha) is actually reduced. However, this type of decision rule will increase the Type II error. Therefore, we recommend sizing the trial carefully for this situation.

There is no single best statistical procedure for multiplicity adjustment because the choice of procedure depends upon the clinical trial's objectives, the most important endpoints, the decision rule for declaring treatment benefit, and other considerations. Some of the statistical procedures that can be useful for a more efficient analysis approach include methods that prespecify a sequence or order of testing or a hierarchy of comparisons that should first be satisfied before others are considered for testing as described above. These methods can be less conservative than the conventional nonhierarchical type methods, such as Bonferroni, the step-down or step-up tests, and prospective alpha allocations schemes, which ignore the hierarchy of comparisons or their families. These conventional type methods should be used when a restriction on the order of testing is not warranted.

A multidomain PRO measure may successfully support a labeling claim based on one or a subset of the domains measured if an *a priori* analysis plan prespecifies the domains that will be targeted as endpoints and the method of analysis that will adjust for the multiplicity of tests for

the specific claim. The use of domain subsets as clinical trial endpoints presupposes that the PRO instrument was adequately developed and validated to measure the subset of domains independently from the other domains.

C. Statistical Considerations for Composite Endpoints

For a PRO instrument with multiple domains, combining the scores to calculate a general score creates a composite endpoint. Composite endpoints have a few advantages (e.g., they can reduce multiplicity problems), but their use for confirmatory clinical trials for specific claims of treatment benefit poses many difficulties and challenges.

Rules for interpretation of composite endpoints depend on substantial experience with the measure in the clinical trial setting. Therefore, development of a composite endpoint at the time the confirmatory clinical trial protocol is written depends on special considerations and substantial empirical evidence of the following: the components are of similar importance to patients, the more important and less important components are equally likely to occur with similar frequency, and the components are likely to have roughly similar treatment effects. Therefore, we discourage the use of a composite endpoint for confirmatory clinical trials when large variations are predicted to exist between its components.

Multiplicity problems arise when the multiple individual components of a composite endpoint are intended as possible claims. In general, individual components of a composite endpoint will not be adequate to support a labeling claim for the components unless the components are prespecified in the protocol as separate endpoints and all prespecified components are reported in labeling as suggested in current guidance.[10] The components of a composite endpoint will be shown in labeling to convey what *drove* a favorable result. Sequential testing approaches can be used to test the components of a composite. The components are tested only when there is a statistically significant treatment benefit for the composite.

D. Statistical Considerations for Patient-Level Missing Data

When the amount of missing data becomes large, clinical trial results can be inconclusive. As described in section IV., Clinical Trial Design, we encourage prespecified procedures in the clinical trial protocol to avoid missing data. We also encourage prespecified procedures for obtaining data on each patient at the time of early withdrawal from the clinical trial. If a measurement is taken at the time of withdrawal, this information can be handled according to rules established in the SAP. In clinical trials of terminal illness, it is critical to plan ahead for how missing data because of death will be handled. Missing data may occur because of the treatment received or the underlying disease and can introduce bias in the analysis of treatment differences and conclusions about treatment effect.

[10] See the guidance for industry *Clinical Studies Section of Labeling for Human Prescription Drug and Biological Products — Content and Format* (http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm).

Even with the best planning, data may be missing at the end of the clinical trial. The SAP should address plans for how the statistical analyses will handle missing data when evaluating treatment benefit and when considering patient success or patient response.

1. Missing Items within Domains

At a specific patient visit, a domain measurement may be missing some, but not all, items. One approach to handling this type of missing data is to define rules that specify the number of items that can be missing and still consider the domain as adequately measured. Rules for handling missing data should be specific to each PRO instrument and usually should be determined during the instrument development process. The SAP should specify all rules for handling missing data. For example, the SAP can specify the proportion of items that can be missing before a domain is treated as missing.

2. Missing Entire Domains or Entire Measurements

We will consider a variety of statistical strategies to deal with missing data because of a patient's early termination before planned completion of a trial. No single method is generally considered as preferred. All of these strategies are imperfect, as they involve strong or weak assumptions about what caused data to be missing, assumptions that usually cannot be verified from the data. Methods of missing data imputation should take the patient population, disease progression, and respondent burden into account. How to impute the missing data for a PRO endpoint and any related supportive endpoints should be addressed in the protocol and the SAP. In addition, the sensitivity analyses in analyzing the PRO endpoints should be proposed in the protocol and the SAP to assess the robustness of statistical estimation for endpoints with the missing data imputed. We recommend that in the protocol the sponsor propose two or more sensitivity analyses with different methods for missing data imputation.

E. Interpretation of Clinical Trial Results

Because statistical significance can sometimes be achieved for small changes in PRO measures that may not be clinically meaningful (i.e., do not indicate treatment benefit), we encourage sponsors to avoid proposing labeling claims based on statistical significance alone.

To demonstrate treatment benefit, we find it informative to examine the cumulative distribution function (CDF) of responses between treatment groups to characterize the treatment effect and examine the possibility that the mean improvement reflects different responses in patient subsets. To interpret the CDF, sponsors can apply the responder definition along the CDF curve at each level of response (see section IV.E., Planning for Clinical Trial Interpretation Using a Responder Definition).

Interpretation of PRO endpoints follows similar considerations as for all other endpoint types used to evaluate treatment benefit of a medical product.

GLOSSARY

Ability to detect change — Evidence that a PRO instrument can identify differences in scores over time in individuals or groups who have changed with respect to the measurement concept.

Claim — A statement of treatment benefit. A claim can appear in any section of a medical product's FDA-approved labeling or in advertising and promotional labeling of prescription drugs and devices.

Cognitive interviewing — A qualitative research tool used to determine whether concepts and items are understood by patients 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 patients interpret questions asked of them. In this method, respondents are often asked to *think aloud* and describe their thought processes as they answer the instrument questions.

Concept — The specific measurement goal (i.e., the *thing* that is to be measured by a PRO instrument). In clinical trials, a PRO instrument can be used to measure the effect of a medical intervention on one or more concepts. PRO concepts represent aspects of how patients function or feel related to a health condition or its treatment.

Conceptual framework of a PRO instrument — An explicit description or diagram of the relationships between the questionnaire or items in a PRO instrument and the concepts measured. The conceptual framework of a PRO instrument evolves over the course of instrument development as empiric evidence is gathered to support item grouping and scores. We review the alignment of the final conceptual framework with the clinical trial's objectives, design, and analysis plan.

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 the 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.

Criterion validity — The extent to which the scores of a PRO instrument are related to a known *gold standard* measure of the same concept. For most PROs, criterion validity cannot be measured because there is no gold standard.

Domain — A subconcept 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 — The measurement that will be statistically compared among treatment groups to assess the effect of treatment and that corresponds with the clinical trial's objectives, design, and data analysis. For example, a treatment may be tested to decrease the intensity of symptom Z. In this case, the endpoint is the change from baseline to time T in a score that represents the concept of symptom Z intensity.

Endpoint model — A diagram of the hierarchy of relationships among all endpoints, both PRO and non-PRO, that corresponds to the clinical trial's objectives, design, and data analysis plan.

Health-related quality of life (HRQL) — HRQL is a multidomain concept that represents the patient's general perception of the effect of illness and treatment on physical, psychological, and social aspects of life. Claiming a statistical and meaningful improvement in HRQL implies: (1) that all HRQL domains that are important to interpreting change in how the clinical trial's population feels or functions as a result of the targeted disease and its treatment were measured; (2) that a general improvement was demonstrated; and (3) that no decrement was demonstrated in any domain.

Instrument — A means to capture data (i.e., a questionnaire) plus all the information and documentation that supports its use. Generally, that includes clearly defined methods and instructions for administration or responding, a standard format for data collection, and well-documented methods for scoring, analysis, and interpretation of results in the target patient population.

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

Item tracking matrix — A record of the development (e.g., additions, deletions, modifications, and the reasons for the changes) of items used in an instrument.

Measurement properties — All the attributes relevant to the application of a PRO instrument 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.

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.

Proxy-reported outcome — A measurement based on a report by someone other than the patient reporting as if he or she is the patient. A proxy-reported outcome is not a PRO. A proxy report also is different from an observer report where the observer (e.g., clinician or caregiver), in addition to reporting his or her observation, may interpret or give an opinion based on the observation. We discourage use of proxy-reported outcome measures particularly for symptoms that can be known only by the patient.

Quality of life — A general concept that implies an evaluation of the effect of all aspects of life on general well-being. Because this term implies the evaluation of nonhealth-related aspects of life, and because the term generally is accepted to mean *what the patient thinks it is*, it is too general and undefined to be considered appropriate for a medical product claim.

Questionnaire — A set of questions or items shown to a respondent to get answers for research purposes. Types of questionnaires include diaries and event logs.

Recall period — The period of time patients are asked to consider in responding to a PRO item or question. Recall can be momentary (real time) or retrospective of varying lengths.

Reliability — The ability of a PRO instrument to yield consistent, reproducible estimates of true treatment effect.

Responder definition — A score change in a measure, experienced by an individual patient over a predetermined time period that has been demonstrated in the target population to have a significant treatment benefit.

Saturation — When interviewing patients, the point when no new relevant or important information emerges and collecting additional data will not add to the understanding of how patients perceive the concept of interest and the items in a questionnaire.

Scale — The system of numbers or verbal anchors by which a value or score is derived for an item. Examples include VAS, Likert scales, and rating scales.

Score — A number derived from a patient's response to items in a questionnaire. A score is computed based on a prespecified, validated 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.

Sign — Any objective 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 subjective evidence of a disease, health condition, or treatment-related effect that can be noticed and known only by the patient.

Target product profile (TPP) — A clinical development program summary in the context of labeling goals where specific types of evidence (e.g., clinical trials or other sources of data) are linked to the targeted labeling claims or concepts.

Treatment benefit — The effect of treatment on how a patient survives, feels, or functions. Treatment benefit can be demonstrated by either an effectiveness or safety advantage. For example, the treatment effect may be measured as an improvement or delay in the development of symptoms or as a reduction or delay in treatment-related toxicity. Measures that do not

directly capture the treatment effect on how a patient survives, feels, or functions are surrogate measures of treatment benefit.

Usability testing — A formal evaluation with documentation of respondents' abilities to use the instrument, as well as comprehend, retain, and accurately follow instructions.

APPENDIX: INFORMATION ON A PRO INSTRUMENT REVIEWED BY THE FDA

The following topics represent areas that should be addressed in PRO documents provided to the FDA for review. The extent of background information provided in each section will vary depending upon the PRO instrument used. Some sections may be less relevant for a particular PRO instrument application than others, or may be less complete for discussions in early stages of medical product development. Refer to the content of this guidance for additional information concerning the types of evidence needed in each of the following areas.

If the PRO information is provided electronically, it should be placed in section 5.3.5.3 of the electronic common technical document.[11]

I. Instrument (review cannot begin without a copy of the proposed instrument):

A. Exact version of the instrument proposed or used in the clinical trial (protocol) under review and all instructions for use. Include screen shots or interviewer scripts, if relevant.

B. Prior versions, if relevant.

C. Instructions for use: An instrument user manual can be provided as Appendix A and referenced here.
   1. Administration timing, method (e.g., paper or pencil, electronic), and mode (e.g., self-, clinician-, or interviewer-administered)
   2. The scoring algorithm
   3. Training method and materials used for questionnaire administration
      a. Patient training — summarize here and include a copy of all materials in Appendix A1
      b. Investigator training — summarize here and include a copy of all materials in Appendix A2
      c. Other training — summarize here and include a copy of all materials in Appendix A3

II. Targeted Claims or Target Product Profile (TPP)[12]

Include language describing all specific targeted labeling claims related to all clinical trial endpoint measures, both PRO and non-PRO, and specific to:

[11] See the ICH guidance for industry *M2 eCTD: Electronic Common Technical Document Specification* (http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm).

[12] See the draft guidance for industry and review staff *Target Product Profile — A Strategic Development Process Tool*. When final, this guidance will represent the FDA's current thinking on this topic. For the most recent version of a guidance, check the FDA Drug guidance Web page at http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm.

- Disease or condition with stage, severity, or category, if relevant
- Intended population (e.g., age group, sex, other demographics)
- Data analysis plan

III. Endpoint Model

A. Relationships (known and hypothesized) among all clinical trial endpoints, both PRO and non-PRO. These endpoints can include physiologic/lab/physical, caregiver, or clinician-reported measures in addition to PROs.

B. Hierarchy of all PRO and non-PRO endpoints intended to support claims corresponding with the planned data analyses.

IV. The PRO Instrument's Conceptual Framework

Diagram of hypothesized (proposed) or final PRO instrument conceptual framework showing relationship of items to domains and domains to total score. Ensure that the PRO instrument's conceptual framework corresponds to the clinical trial endpoints described in the clinical trial protocol and proposed as labeling claims.

V. Content Validity Documentation

Evidence that instrument captures all of the most clinically important concepts and items, and that items are complete, relevant (appropriate), and understandable to the patient. This evidence applies to both existing and newly created instruments and is specific to the planned clinical trial population and indication. Documentation includes:

A. Literature review and documentation of expert input

B. Qualitative study protocols, interview guides, and summary of results for:
   1. Focus group testing (include transcripts in Appendix C1)
   2. Open-ended patient interviews (include transcripts in Appendix C2)
   3. Cognitive interviews (include transcripts in Appendix C3)

C. Origin and derivation of items with chronology of events for item generation, modification, and finalization

   Item tracking matrix for versions tested with patients showing items retained and items deleted providing evidence of saturation. Summarize here and include complete materials under Appendix B.

D. Qualitative study summary that supports content validity for:
   1. Item content
   2. Response options
   3. Recall period
   4. Scoring

E. Summary of qualitative studies demonstrating how item pool was generated, reduced, and finalized. Specify type of study (i.e., focus group, patient interview, or cognitive interview) and characteristics of study population. Include full transcripts and datasets in Appendix C.

VI. Assessment of Other Measurement Properties

Assuming content validity is established in the intended population and application, evidence that the instrument is reliable, valid, and able to detect change. The same version of the instrument to be used in the clinical trial should be used to assess measurement properties.

A. Protocols for instrument testing

B. Summary of testing results for each domain or summary score proposed as support for claims:
   1. Reliability (internal; test-retest)
   2. Construct validity (convergent, discriminant, known-groups)
   3. Ability to detect change

VII. Interpretation of Scores

A. Summary of the logic and methods used to interpret the clinical meaningfulness of clinical trial results

B. Responder definition (i.e., definition of meaningful within-person change specific to the clinical trial population)

VIII. Language Translation and Cultural Adaptation

A. Process used to translate and culturally adapt the instrument for populations that will use them in the trial

B. Description of patient testing, language- or culture-specific concerns, and rationale for decisions made to create new versions.

C. Copies of translated or adapted versions

D. Evidence that content validity and other measurement properties are comparable between the original and new instruments

IX. Data Collection Method

A. Process used to develop data collection methods (e.g., electronic, paper) intended for use in the clinical trial

If electronic data collection is used to assess PRO endpoints, evidence that procedures for maintenance, transmission, and storage of electronic source documents comply with regulatory requirements.

B. Evidence that content validity and other measurement properties are comparable among all data collection methods

C. User manual for each additional data collection method

X. Modifications

Any change in the original instrument (e.g., wording of items, response options, recall period, use in a new population or indication)

A. Rationale for and process used to modify the instrument

B. Copy of original and new instruments

C. Evidence that content validity and other measurement properties are comparable between the original and modified instruments (including use in a new indication or population)

XI. PRO-Specific Plans Related to Clinical Trial Design and Data Analysis

A. Clinical trial protocol. Ensure in the protocol that:
   - Each PRO endpoint is stated as a specific clinical trial objective and multiplicity concerns are addressed
   - The clinical trial will be adequately blinded
   - Procedures for training are well-described for:
     – Patients
     – Interviewers
     – Clinical investigators
   - Plans for instrument administration are consistent with instrument's user manual
   - Plans for PRO instrument scoring are consistent with those used during instrument development
   - Procedures include assessment of PRO endpoint before or shortly after a patient withdraws from the clinical trial
   - Frequency and timing of PRO assessments are appropriate given patient population, clinical trial design and objectives, and demonstrated PRO measurement properties
   - Clinical trial duration is adequate to support PRO objectives
   - Plans are included for handling missing data
   - Plans are included for a cumulative distribution function comparison among treatment groups
   - Data collection, data storage, and data handling and transmission of procedures, including electronic PROs, are specified

B. Statistical analysis plan (SAP). Ensure the SAP includes:
   - Plans for multiplicity adjustment
   - Plans for handling missing data at both the instrument and patient level
   - Description of how between-group differences will be portrayed (e.g., cumulative distribution function)

XII. Key References

List and attach all relevant published and unpublished documents

Appendix A — User Manual

A1: Patient training

A2: Investigator training

A3: Other training

Appendix B — Item Tracking Matrix

Appendix C — Transcripts

C1: Focus groups

C2: Open-ended patient interviews

C3: Cognitive interviews