U.S. FOOD & DRUG ADMINISTRATION

PATIENT-FOCUSED DRUG DEVELOPMENT
GUIDANCE PUBLIC WORKSHOP

Methods to Identify What is Important to Patients & Select, Develop or Modify Fit-for-Purpose Clinical Outcomes Assessments

Workshop Date: October 15-16, 2018

Attachment to Discussion Document for Patient-Focused Drug Development Public Workshop on Guidance 3:

SELECT, DEVELOP OR MODIFY FIT-FOR-PURPOSE CLINICAL OUTCOME ASSESSMENTS

APPENDICES

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

Appendix 1: Information on a COA Reviewed by the FDA

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

If the COA information is provided electronically, it should be placed in section 5.3.5.3 of the electronic common technical document.[1]

I. Instrument (review cannot begin without a copy of the proposed instrument and its scoring algorithm):
   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. Timing, administration mode (e.g., self-, clinician-, or interviewer-administered) and data collection method (e.g., paper or pencil, electronic)
      2. The scoring algorithm
      3. Training method and materials

II. <u>Context of Use</u>
   A. Identify the targeted study population, including a definition of the disease and selection criteria for clinical trials (e.g., baseline symptom severity, patient clinical and demographic characteristics, language/culture subgroups)
   B. Identify the targeted study design.
   C. Identify the targeted study objectives.
   D. Identify the endpoint definition and positioning (i.e., planned set of primary and secondary endpoints with testing hierarchy), if known.
      1. Relationships (known and hypothesized) among all clinical trial endpoints, both COA and non-COA.
      2. Hierarchy of all COA and non-COA endpoints intended to support claims corresponding with the planned data analyses.

III. <u>The COA's Conceptual Framework</u>

Diagram of hypothesized (proposed) or final COA conceptual framework showing relationship of items/tasks to domains and domains to total score. Ensure that the COA's

[1] See the ICH guidance for industry *M2 eCTD: Electronic Common Technical Document Specification*

conceptual framework corresponds to the clinical trial endpoints described in the clinical trial protocol and proposed as labeling claims.

IV. 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 C)
    2. Open-ended patient interviews (include transcripts in Appendix C)
    3. Cognitive interviews (include transcripts in Appendix C)

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

    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.

V. 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. Psychometric analysis plan to evaluate instrument measurement properties
    - Item descriptive statistics including frequency distribution of both item response and overall scores, floor and ceiling effect, and percentage of missing response
    - Inter-item relationships and dimensionality analysis (e.g., factor analysis or principal component analysis and evaluation of conceptual framework).
    - Item inclusion and reduction decisions, identification of subscales (if any), and modification to conceptual framework
    - Preliminary scoring algorithm (e.g. include information about evaluation of measurement model assumptions, applicable goodness-of-fit statistics). The scoring algorithm should also include how missing data will be handled.
    - Reliability
        - Test-retest (e.g., intra-class correlation coefficient)
        - Inter-rater (e.g. kappa coefficient)
        - Intra-rater (e.g., intra-class correlation coefficient)
        - Internal consistency (e.g. Cronbach's alpha)

    - Construct validity
        - Convergent and discriminant validity (e.g., association with other instruments assessing similar concepts)
        - Known groups analysis (e.g., difference in scores between subgroups of subjects with known status)
    - Score reliability in the presence of missing item-level and if applicable scale-level data
    - Final instrument, conceptual framework, provisional scoring algorithm for exploratory use, and plans for further revision and refinement

C. Summary of testing results for each domain or summary score proposed as support for claims:
    1. Descriptive statistics
    2. Reliability
    3. Construct validity
    4. Ability to detect change

VI. <u>Interpretation of Scores</u>

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

B. Threshold(s) (e.g., a range of score change) that constitutes a clinically meaningful within-patient change (improvement and worsening) in scores in the target patient clinical trial population

VII. 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

VIII. Data Collection Mode

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

   If electronic data collection is used to assess COA 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 modes

C. User manual for each additional data collection mode

IX. 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 in the modified instrument are fit-for-purpose for the new context of use.

X. COA-Specific Plans Related to Clinical Trial Design and Data Analysis

A. Clinical trial protocol. Ensure in the protocol that:
   - Each COA 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
   - Plans for instrument administration are consistent with instrument's user manual
   - Plans for COA scoring are consistent with those used during instrument development
   - Procedures include assessment of COA endpoint before or shortly after a patient withdraws from the clinical trial

- Frequency and timing of COA assessments are appropriate given patient population, clinical trial design and objectives, and demonstrated COA measurement properties
- Clinical trial duration is adequate to support COA 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 COAs, 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)

XI. Key References

List and attach all relevant published and unpublished documents

Appendix A — User Manual
Appendix B — Item Tracking Matrix
Appendix C — Transcripts (upon request)

Appendix 2: Examples of Response Option Types

Table 1. Examples of Response Option Types

Type	Description	Potential Limitations
Checklists	• A response scale that allows respondents to provide multiple answers to a single item (i.e., respondent can check off all the choices that apply to them). • Checklists are commonly shown with square checkboxes Checklists can generate categorical data.	• Provides limited information • Checklists may not cover all of the possible responses; in these instances, free text may be needed • The use of checklists can impact data analysis, so careful consideration is needed when analyzing data from a multi-option variable
Numeric rating scale	• A response scale with numeric labels from which respondents are asked to choose from an ordered set of response options, for example, 1 to 10, coupled with anchors (words). Anchors can be put at the endpoints or at each point on the scale. Numeric rating scales can generate interval data.	• Potential decreased validity with lower extremes of age
Pictorial scale	• A response scale with a set of pictures applied to a set of response options (numeric or verbal labels). • 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. Pictorial scales can generate ordinal and/or interval data.	• May not account for cultural and ethnic differences • Cannot be administered verbally
Verbal rating scale	• A response scale with verbal labels from which respondents	• Limited number of response categories

Type	Description	Potential Limitations
	are asked to choose from verbal descriptors, for example, "None/Mild/Moderate, Severe." Verbal rating scales can generate ordinal data.	• Decreased validity in illiterate patients • Although distances between verbal descriptors on verbal rating scales appear equidistant, the actual observed distances may vary. • Only rank-order inferences can be made about the relative differences between two or more ratings.
Visual analog scale (VAS)	• A response scale represented by a line of fixed length (usually 100 mm) coupled with anchors (words) at the endpoints, which respondents indicate a position along the line between two endpoints. Anchors can also be positioned along the line (i.e., anchored or categorized VAS). • VAS can generate interval and/or ratio data.	• False sense of precision • Cannot be administered verbally • Higher rates of missing data (Dworkin et al., 2005; Hawker et al., 2011) • Inconsistencies with the length of VAS line (paper photocopying/printing differences, zooming in/out on electronic VAS line)

Appendix 3: Measurement Properties Considered in the Review of COAs used in Clinical Trials

Table 2. Measurement Properties Considered in FDA Review of COAs

Measurement Property	Type	What Is Examined	FDA Review Consideration
Reliability	Test-retest reliability	Consistency of scores over time when no change is expected in the concept of interest	• Time periods of assessment • Statistics and/or figures demonstrating the degree of agreement between scores (e.g., intra-class coefficient ≥0.70) • Does the study design, disease condition (e.g., acute), or treatment effect (e.g., rapid acting) allow assessment of test-retest reliability?
	Intra-rater reliability	Consistency of ratings from the same rater to multiple patients who are identified as the same in the concept of interest	• Time periods of assessment • Statistics and/or figures demonstrating the degree of agreement between ratings
	Inter-rater reliability	Agreement of ratings from the multiple raters to the same patient or patients who are identified as the same in the concept of interest	• Time periods of assessment • Statistics and/or figures demonstrating the degree of agreement between ratings
	Internal consistency	Extent to which items composing a scale measure the same concept	• Statistics and/or figures demonstrating the degree of internal consistency among items (e.g., Cronbach's alpha >0.70)
Validity	Content validity	Evidence that the COA 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	• Derivation of all items • Literature review • Stakeholder input (e.g., patients, clinicians, caregivers) • Interview or focus group transcripts • Items derived from the transcripts

Measurement Property	Type	What Is Examined	FDA Review Consideration
		concept, population, and use. Testing other measurement properties will not replace or rectify problems with content validity.	• Composition of patients used to develop content • Cognitive interview transcripts to evaluate respondent 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 (convergent validity and/or discriminant validity) • Degree to which the COA score can distinguish among groups hypothesized *a priori* to be different (known groups analysis)
Ability to detect change		Evidence that a COA score 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

Appendix 4: Examples of Generic PGIS and PGIC Scales[2]

Examples of Patient Global Impression of Severity (PGIS) Scales:[3]

Please choose the response below that
best describes the severity of your
<SYMPTOM/OVERALL STATUS/ETC.>
over the past week.

□ None
□ Mild
□ Moderate
□ Severe

Please choose the response below that
best describes the severity of your
<SYMPTOM/OVERALL STATUS/ETC.>
over the past week.

□ None
□ Mild
□ Moderate
□ Severe
□ Very severe

Example of a Patient Global Impression of Change (PGIC) Scale:

Please choose the response below that
best describes the overall change in your
<SYMPTOM/OVERALL STATUS/ETC.>
since you started taking the study medication.

□ Much better
□ A little better
□ No change
□ A little worse
□ Much worse

[2] Note: Global scales can be used for other types of COAs, however the instructions and question stems would need to be modified appropriately.

[3] The appropriateness of the PGIS scale used depends on the context of use (e.g., patient population). For example, sponsors should explore whether patients in the target patient population believe that going from "very severe" to "severe" would be considered a meaningful improvement.

263

Appendix 5: Observer-Reported Outcome Assessment Use Throughout the Medical Product Lifecycle to Support Patient-Focused Outcome Measurement

I. INTRODUCTION

This discussion document attachment is intended to provide additional considerations for the development and implementation of an observer-reported outcome (ObsRO) that are not discussed in the Guidance 3 discussion document. This document will focus on input reported by observers other than clinicians or trained health care professionals. For general principles that can be broadly applied across all COAs, please refer to the Guidance 3 discussion document.

II. BACKGROUND

While patient input is critical and provides meaningful information on clinical outcomes, there are instances where patient input cannot be obtained or reported reliably (e.g., young children, individuals with cognitive problems) and other stakeholder input is needed (e.g., clinician or other trained health care professional and/or primary caregiver(s)) to report and understand what is most valuable to assess in patients.

An ObsRO is a type of clinical outcome assessment that assesses observable signs, events or behaviors related to a patient's health condition and is reported by someone other than the patient or a health professional (e.g., a parent, caregiver, or someone who observes the patient in daily life). ObsROs are particularly useful for patients who cannot report for themselves (e.g., infants or individuals who are cognitively impaired). An ObsRO instrument does not rely on medical judgment or interpretation.

Example

What are some examples of ObsRO instruments?

- Rating scales completed by a caregiver, such as:
    - Acute Otitis Media Severity of Symptoms scale (AOM-SOS), a measure used to assess signs and behaviors related to acute otitis media in infants
    - Face, Legs, Activity, Cry, Consolability scale (FLACC), a measure used to assess signs and behaviors related to pain
- Counts of events recorded by a caregiver (e.g., observer-completed log of seizure episodes)

III. OBSERVER-REPORTED OUTCOME ASSESSMENTS IN MEDICAL PRODUCT DEVELOPMENT

A. General Considerations for ObsROs

Who should report on the patient experience? FDA generally recommends that the patient directly report on their experience with their disease or condition, unless the patient cannot reasonably be expected to reliably self-report (e.g., young children, individuals with cognitive problems, such as Alzheimer's disease, etc.). In such cases a parent, caregiver, or someone who observes the patient in daily life may report on patient experience if it is observable (e.g., signs of disease or condition, events, behaviors, etc.).[4]

Who the ***reporter*** is (i.e., the person who will be providing the patient experience information) may vary from patient to patient *within* the target population. Every effort should be made to ensure that all observer-reported assessments for a given subject are completed by the same individual throughout the study.

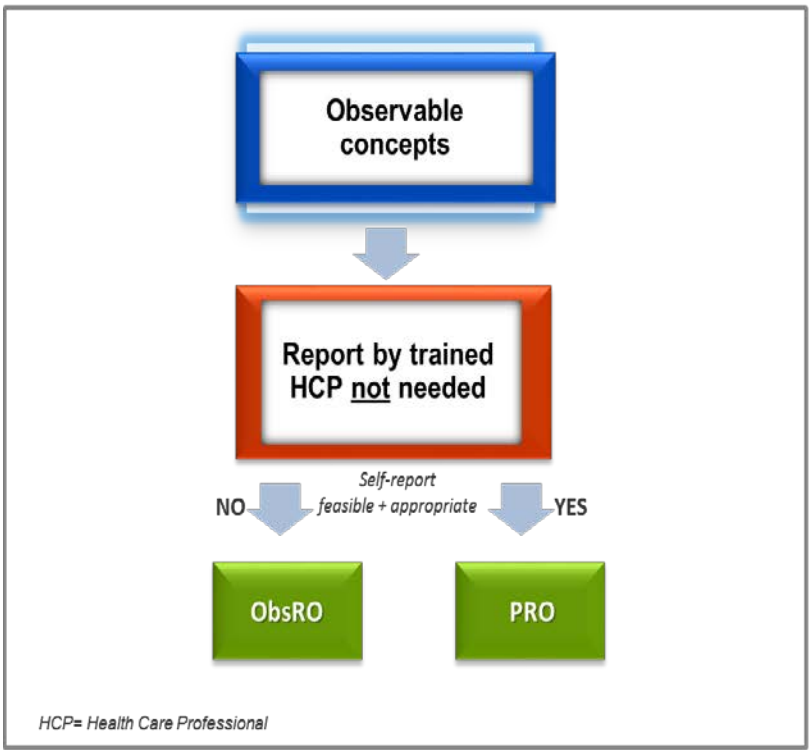
Factors to consider when determining if self-report is feasible for patients include (but are not limited to):

- Age
- Level of cognitive development (e.g., reading ability, numeracy)
- Communication skills (e.g., verbal ability)
- Health literacy
- Insight
- Disease of interest or concomitant illness affecting cognitive ability (including level of consciousness/awareness)

If the concept of interest does not require report by a trained health-care professional, can be adequately captured only by observation in daily life (outside of a health care setting), and the patient cannot report for him- or herself, then an ObsRO should be considered. See **Figure 1**.

[4] FDA. (2015). Clinical Outcome Assessment (COA): Glossary of Terms. Retrieved March 11, 2018, from https://www.fda.gov/drugs/developmentapprovalprocess/drugdevelopmenttoolsqualificationprogram/ucm370262.htm

Figure 1. Process to Select ObsRO[5]



When to consider using both a PRO and ObsRO instrument? In studies that include a wide range of patient age or disease severity groups (e.g., varying levels of cognitive impairment in a progressive disease), it may be necessary to administer both a PRO and an ObsRO instrument using similar forms of instruments measuring the same concepts.

Some key considerations for using both a PRO and ObsRO instrument include but are not limited to the following:

- Conduct qualitative research and assess measurement properties in each subgroup to determine whether the measurement concepts are the same and understood and interpreted similarly.
- Establish criteria to determine when multiple reporters are needed (e.g., determine the minimal age limit at which children can provide reliable responses; determine minimal cognitive function at which individuals can provide reliable responses)
- Engage with subject matter experts in specific disease areas to determine the appropriateness of self-report or an ObsRO and use of multiple reporters in the target population. (FDA, 2015).
- Assess feasibility to use both a PRO and ObsRO in clinical trials, particularly multinational trials.

[5] This selection process may also be relevant for selection of COAs derived from mobile health technologies (e.g., activity monitors, sleep monitors).

Example

Scenario: There is a rare itch condition that can manifest in infancy or early childhood, in which case, symptoms progress very quickly. However, some children show initial signs of the disease (e.g., even as late as the teen years) and their condition progresses more slowly. Nearly all children with this condition will require treatment before age 30.

What type of COA should be used to measure itch in this population (ObsRO vs. PRO)?

Since children ages 6 months to 7 years are not able to reliably and validly self-report on their itch symptoms, we rely heavily upon observer reports (e.g., parent, caregiver) of observable signs of itching in their children. An ObsRO instrument would best capture scratching events in younger children (as they often co-sleep with their children and closely monitor them or get reports on their itching throughout the day). A PRO instrument, on the other hand, would still be the best measure of itch symptoms (intensity) among older children (ages 8 and above) who can reliably and validly self-report on their own symptoms.

In instances where you have a wide range of ages in a clinical trial, both the ObsRO and PRO can be administered if a similar concept is being measured comparably across both instruments (e.g., scratching frequency). That is, the ObsRO can be administered across all ages and the PRO can be administered among older children and used as supportive data to help interpret the results of the ObsRO.

A digital monitoring device could also be potentially used for exploratory purposes to monitor the scratching experience with this condition.

What is the difference between an ObsRO and Proxy-reported outcome[6] instrument? An ObsRO instrument is limited to the assessment of observable signs and symptoms that can be reported from the perspective of a parent or caregiver. A proxy-reported outcome instrument is not an ObsRO instrument but is an assessment in which someone other than the patient reports on patient symptom experiences as if he or she is the patient. Proxy-reported outcome instruments are discouraged when measuring concepts that are only known by the patient (e.g., symptoms) because they do not necessarily reflect how patients feel and function in daily life. FDA acknowledges that there are some instances where it is impossible to collect valid and reliable self-report data from the patient. However, in these instances, it is recommended that an ObsRO instrument be used, rather than a proxy-reported outcome instrument.

[6] https://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm370262.htm

Example

Examples of an ObsRO item versus a Proxy-reported item

ObsRO Items

- "Based on what you observed/saw, please rate how fussy your child has been today"
- "Based on what you observed or what your child told you, how often did your child itch (i.e., scratch) from the time your child woke up today until now."

Proxy-reported Outcome Items

- "How severe was your child's pain from the time he/she woke up until right now?"
- "Please rate your child's tiredness over the past 24 hours."
- "My child felt wheezy and out of breath because of his/her asthma."
- "My child felt sad when he/she had pain."

IV. EVALUATION OF AN OBSERVER-REPORTED OUTCOME

A. Evidence of Content Validity

The ObsRO instrument should capture all the important and relevant aspects of the overall concept it is supposed to measure. Evidence may come from literature review, input from content area and measurement experts, and qualitative research in individuals who are representative of the population of responders (either the patient or the observer) who will be completing the assessment in clinical trials. Additionally, if the patient can discuss their experience with the disease or condition, sponsors can obtain direct input from the patient about the aspects of their condition that are important to them to inform the choice of concepts to be measured. Where possible, input from multiple sources (children or patients with impaired ability to communicate; parents or caregivers; clinicians, other experts) regarding significant signs, symptoms, and effects on daily living is important to consider when determining the appropriate concept(s) of interest to be measured.

When behavioral manifestations of the same symptom vary by age or symptom severity group when a condition changes over time, it is very important to conduct interviews with clinicians with expertise in the disease or condition, caregivers, and/or patients to understand how to measure such symptoms over time or in study participants of different ages. This information can be used to adapt COAs for these variations. FDA recommends that sponsors engage in early discussions with the Agency whether the disease is sufficiently similar across the age groups to use common COAs.

EXAMPLE

***Scenario*:** Eosinophilic esophagitis is a chronic disease with signs and symptoms that differ by age. In infants, food refusal is commonly observed. Children often suffer from gastro-esophageal reflux-like symptoms: vomiting, dysphagia, and abdominal pain. Adolescents experience mostly dysphagia for solids and food impaction. Because of the different clinical presentations by age it will be important to measure the appropriate concept during the patient's experience over time with this condition.

Please refer to the Guidance 3 discussion document **Section VIB of Guidance 3 discussion document** for further discussion about general considerations when generating evidence of content validity.

1. Item and Content Generation

ObsRO instructions and items should be designed in a way that observers (e.g., parents, caregivers) understand that they are to report only on observable signs, behaviors, and verbalizations made by the patient. Development of training and a standardized approach to assessment for the observer will also be critical. The same observer should complete the assessments throughout the trial.

Please refer to **Section VIB.3 of Guidance 3 discussion document** for further discussion about general considerations when generating COA item and content.

2. Recall Period (if applicable)

It is important to assess, through cognitive interviews, whether parents or caregivers fully understand the recall period of an ObsRO instrument consistently across respondents. Refer to **Section VIB.3 of Guidance 3 discussion document** for further details on considerations for selecting appropriate recall periods for COAs.