Clinical Outcome Assessments: Establishing and Interpreting Meaningful Within-Patient Change
April 4, 2017
Meeting Summary

Background

Over the last several decades, there has been growing recognition of the importance of incorporating the patient perspective into medical product development. A key component of this effort has been the development and implementation of fit-for-purpose clinical outcome assessments (COAs) that can accurately and reliably measure meaningful treatment outcomes in a clinical trial. COAs are generally divided into four broad categories, depending on how the assessment is conducted and reported: patient-reported outcomes (PROs), clinician-reported outcomes (ClinROs), observer-reported outcomes (ObsROs), and performance outcomes (PerfOs). While most registration trials utilize COAs to support primary or secondary endpoints, a continuing challenge with their use has been the interpretation of meaningful within-patient change over time – specifically, how to establish the threshold of change in those study endpoints that can be interpreted as being clinically meaningful. The usefulness and relevance of a COA endpoint relies on its ability to accurately capture the effectiveness of an intervention and communicate those results to patients and other stakeholders. The degree to which an observed change in a COA endpoint can be interpreted as clinically meaningful is of particular importance to providers and patients who are making treatment decisions.

Deriving thresholds for interpreting meaningful within-patient change for COAs is a complex issue for which there is little scientific consensus and many ongoing challenges. Patients, for example, may have different expectations for treatment outcomes, particularly when their disease is heterogeneous in terms of symptom severity or the particular symptoms that a given patient experiences. What they regard as ‘meaningful’ may also depend on the stage of their disease. There are also ongoing questions related to how best to choose and apply the most appropriate method for deriving thresholds to interpret change, as well as whether and how emerging methods might be applied in certain contexts.

The Regulatory Perspective on Meaningful Change

The U.S. Food and Drug Administration (FDA) published guidance for one type of COA – patient-reported outcomes – in 2009, which includes some considerations and recommendations for interpreting meaningful within-patient change. In addition to statistical evidence of treatment effect, FDA generally recommends the use of evidence-based responder definitions and meaningful change estimates to support the interpretation of the COA endpoints. The guidance underlines the importance of examining individual-level responses to treatments for interpretation and communication of treatment benefit.∗

∗Similar recommendations for establishing meaningful change appear in later FDA guidances as well, including the 2012 Irritable Bowel Syndrome Guidance, the 2013 Alzheimer’s Disease Draft Guidance, and the 2014 Analgesic Indications Draft Guidance.
The guidance also highlights one primary approach, anchor-based methods, to derive meaningful within-patient change. Though these methods have been widely used, there are some limitations to this approach, as well as challenges with evaluating the direction of the meaningful change (specifically how to derive thresholds that interpret stability or worsening on the COA, not just improvement). In addition to anchor-based methods, the guidance also mentions the use of distribution-based methods and cumulative distribution function (CDF) displays to provide supportive and supplemental information to anchor-based approaches. However, distribution-based methods alone are not sufficient for establishing meaningful change scores.

There has also been substantial interest in identifying alternative approaches that may be useful or supportive in interpreting clinically meaningful change. Examples of these novel approaches include, but are not limited to, bookmarking/standard-setting, scale-judgment, and exit-interview methods. Each of these methods has advantages and limitations that may make them appropriate for use, depending on the COA, therapeutic area, or type of intervention. However, these methods are still relatively new and require further research and stakeholder discussion regarding their suitability for use in clinical trials. Furthermore, most of the work to date on methods for interpreting meaningful within-patient change has focused on PROs. While some principles can be applied broadly across all COA types, there are context-specific issues that still need to be addressed for ClinROs, ObsROs, and PerfOs.

**Meeting Objectives**

In light of these ongoing issues, and under a cooperative agreement with the U.S. Food and Drug Administration (FDA), the Duke-Margolis Center for Health Policy convened an expert workshop on April 4, 2017 entitled “Clinical Outcome Assessments: Establishing and Interpreting Meaningful Within-Patient Change” to advance the discussion on meaningful within-patient change. The objectives of the workshop were to: 1) explore and discuss methodologies and best practices surrounding meaningful within-patient change, and 2) identify specific recommendations on methodologies used to derive and interpret meaningful within-patient change with use of COA endpoints in medical product development.

This workshop provided an opportunity for representatives from across academia, industry, government, and patient advocacy to engage and discuss the challenges associated with interpreting meaningful within-patient change, explore three emerging approaches, and determine how and when different methods may be applicable across the four types of COA. The day’s discussion was divided into four sessions, one dedicated to anchor-based methods and the others each dedicated to an emerging method. The following text represents a summary of the meeting, including areas of consensus as well as areas that require additional research.

**Addressing Ongoing Challenges with Anchor-Based Methods to Establish Meaningful Within-Patient Change**

Anchor-based methods examine the associations between the concept of interest targeted by a COA and the concept measured by independent anchoring measure(s), often other COAs. The anchor measure(s) are used as external criteria to define patients who have experienced a meaningful change in their condition. The meaningful change estimate of the COA endpoint of interest can then be derived from the group of patients who are identified as having experienced meaningful change based on the anchor measure. Multiple anchor measures may be used to provide a range of meaningful change estimates rather than a single estimate. To be useful, the selected anchor(s) should be plainly
understood in context, easier to interpret than the COA itself, and sufficiently correlated to the targeted COA.\(^5\)

As noted above, anchor-based methods are a widely used approach for determining meaningful within-patient change, and they can be reliably applied across COA types. However, participants agreed that successful application of anchor-based methods relies on the appropriateness of the anchoring measures. If adequate anchors are not available, if there is not a sufficient number of patients who experienced meaningful change in the study, or if there is not sufficient clinical trial experience to draw from, it may not be possible to derive a threshold (or the threshold that is derived may not be adequate). It is also possible that different anchors may provide substantially different meaningful change threshold estimates. These factors will affect the reliability and interpretability of the derived threshold. When feasible, other methods may be explored to compliment the anchor-based methods especially when no adequate anchor is available. The participants agreed that relying solely on the anchor-based method may hamper the development and study of other methods that may be more appropriate for a given context.

Discussion also focused on a simple global impression question that has been widely used as an anchor, the Patient Global Impression of Change (PGIC). PGIC anchors can suffer from recall error because they are most highly correlated with a patient’s current state. For example, if a patient is feeling better on a day when they are questioned about their impression of change, they are more likely to say there has been improvement in their symptoms over time even if that particular day was an outlier in terms of their symptom or disease severity. To avoid issues with recall error, many suggested using Patient Global Impression of Severity (PGIS) anchors instead. However, PGIS anchors have their own limitations, as a single item may lack measurement precision. Because of limitations inherent in all anchors, this discussion underlined the notion that using multiple anchors will provide the most reliable results.

**Statistical Considerations for Anchor-Based Methods**

A number of statistical challenges emerged from the discussion. Experts noted that researchers often treat measures as continuous data when they are actually ordinal. For example, Patient Global Impression of Change (PGIC) with response options such as ‘very much improved’, ‘improved’, and ‘no change’ are ordinal in nature. Handling these data as continuous outcomes may violate certain assumptions, rendering the result uncertain. The general advice is to test the assumptions prior to conducting analyses to ensure that these assumptions are met or acceptable.

Participants also reinforced the importance of examining individual-level data to establish meaningful change. A mean group-level score difference (e.g., score difference between the ‘a little improved’ group and the ‘no-change’ group) does not capture or reflect meaningful treatment benefit to patients. However, some participants noted that group-level methods are often misused to define responder thresholds at the individual level. Participants underscored that individual level changes could only be observed in longitudinal studies. Therefore, while cross-sectional data can be used to look at group-level estimates, only longitudinal data can be utilized to calculate individual-level change.
Exploring the Applicability of the Bookmarking/Standard-Setting Method and the Scale-Judgment Method to Establish Meaningful Within-Patient Change

The first discussion of emerging methods focused on the bookmarking/standard-setting method. In this approach, patients and experts are presented with clinical vignettes of a disease (selected across the score scale to represent the full range of disease severity) in order to reach a consensus on thresholds for severity levels to interpret change. The following session was dedicated to another emerging method – the scale-judgment method. In the scale-judgment method, panels of judges evaluate vignettes comprised of completed questionnaires that meet certain conditions to determine whether the amount of change specified by the responses before and after a treatment indicate an important difference.

While there are nuances in the application of each of these methods, participants noted significant similarities between the two approaches. Both methods include groups of patients, experts, or judges who evaluate hypothetical clinical vignettes to determine whether a change was meaningful to patients, and both employ item response theory (IRT) methodology as part of their analysis (this approach is discussed in greater detail in the next section). Both methods also have advantages over anchor-based methods, as they do not require an anchor to carry out the analysis and they incorporate patient input directly.

However, these methods require a well-defined and reliable instrument with an established score scale, and may not be feasible for use with multi-concept instruments. These methods may also require a large sample size if modern psychometric based scoring models such as an IRT are used. Additionally, selection of the vignettes is dependent on the quantitative rules specified by the designers of each study. Accounting for differences in disease areas, disease progression, and whether the disease is chronic or episodic, may impact whether a set of vignettes or questions adequately reflects patients’ or responders’ experiences with a disease. It was unclear to participants whether all of these considerations could be sufficiently addressed through pre-selected vignettes. One possible solution is to directly involve patients in the development of the vignettes so that they are representative of patients’ actual experiences.

Statistical Considerations for Bookmarking/Standard-Setting and Scale-Judgment Methods

Item response theory is a probability-based, psychometric modeling technique that assumes that responses on a set of items or questions are related to an unmeasured “trait”. IRT modeling is complex and relies on certain assumptions in order to work properly. Its application usually requires persons with special training and the use of specific statistical software. Considering these constraints, some participants suggested that more conventional methods for statistical analysis could be used instead of IRT-based methods. However, further applications of these methods are needed to determine what types of models are most effective.

There were also statistical concerns stemming from where a patient falls on a scale in terms of disease severity. In one of the examples presented, patients at the end of the distribution required more of a change to be meaningful, while patients whose severity fell in the middle of the distribution required less of a change. Such results can be difficult to analyze, as they make it challenging to identify a threshold.
Another area of concern was the operational challenge of applying these methods, particularly in a clinical trial setting. The unidimensional IRT does not allow researchers to analyze more than a single domain at a time, while patients may experience a variety of symptoms associated with their disease. Thus, there was uncertainty over how to develop a reasonable set of representative vignettes for a multi-symptom disease, especially for a heterogeneous population. There were also concerns with the patient burden in rating these vignettes. Comprehension of items and the ability to assess meaningfulness might become more challenging as the number of vignettes and categories increase. In situations where responders are asked to consider too many items, they may simply use heuristics and only focus on one or two items most important to them. This could lead to paradoxical or inconsistent judgments from responders, skewing the results. When designing studies, researchers should take steps to reduce the complexity and difficulty for patients evaluating these vignettes.

The applicability of these methods in certain populations may also be challenging. Pediatric patients, for example, may not be able to self-report or may have different priorities than parents or observers, and it can be difficult to weigh these competing priorities. It was also noted that these methods might not be useful for populations with cognitive impairments who cannot make judgments or understand the nuances of the vignettes. Lastly, participants were skeptical of utilizing this approach for rare disease populations where there could be a great deal of heterogeneity. Overall, while participants generally expressed enthusiasm for these methods and their potential applications, there was consensus that more research and testing is needed.

**Exploring the Applicability of Exit-Interviews to Establish Meaningful Within-Patient Change**

In the exit-interview approach, patients are formally asked a series of questions at the end of a clinical trial regarding the impact of the treatment. These interviews allow patients to describe, in their own words, the meaningfulness of treatment-related changes. Exit interviews also enable researchers to detect unanticipated treatment benefits, compare pre-study expectations with clinical outcomes, and identify characteristics of new or rare patient populations. Because exit interviews are more qualitative than the other methods for interpreting meaningful within-patient change, they are complementary to other quantitative-oriented methods to determine clinical meaningfulness.

**Statistical and Operational Considerations for Exit-Interviews**

Overall, participants generally viewed the exit-interview method favorably, as it provides additional information that can enrich researchers’ understanding of the patient experience with a disease, informs researchers on the impact of treatment on daily life, and may aid in the interpretation of other clinical data. Participants felt that exit-interviews could provide beneficial supplementary information to researchers when interpreting meaningful within-patient change, although some of the issues raised would need to be addressed to implement them into a clinical trial.

For example, it was noted that exit-interviews may introduce selection bias, as the volunteers who choose to participate in these interviews may not reflect the full range of patient experiences in a trial. Recall error could also impact patients’ perceptions of their symptoms and disease progression over time, as the interviews happen at the end of the trial. It was suggested that adding interviews both before and immediately following the trial could mitigate this concern. It was also emphasized that
patients often try to discern whether they are receiving the treatment, which could potentially affect the neutrality of the interviews. Furthermore, there was discussion of how the questions should be chosen, whether they adequately cover a range of symptoms to reflect the variety of patient experiences, and whether the questions could also focus on slowing or non-worsening of disease progression rather than just improvement. Lastly, because interviews ask the patient specifically about their treatments, participants wondered if exit-interviews were appropriate across the range of COA types, or should be limited to PROs.

**Major Takeaways and Areas for Future Research**

It was generally agreed that more consensus is needed on how “meaningful” is defined when interpreting COA results. Participants agreed that a minimal amount of change that is noticeable does not imply that the change was clinically meaningful to patients. However, there was disagreement over whether the change needed to significantly impact patients’ daily living in order to be considered meaningful. Some posited that to be meaningful, a change only needed to be one that patients noted as significant and one they cared about. Another factor that had not been widely considered by many participants prior to the discussion was the benefit/risk calculus, specifically, how much risk a patient is willing to tolerate in order to reach a given level of change. Often the focus of a treatment is the benefit to the patient, while the risk to the patient is overlooked. Developing a more formalized definition of “meaningful” change may be necessary in order to move this field forward.

Participants emphasized throughout the day’s discussion that patient engagement is of paramount importance. They also stressed that patients should be engaged as early as possible, ideally in the pre-competitive space, in order to minimize the burden on patients and patient groups.

There was also general agreement that there is no “gold-standard” for methods for interpreting meaningful within-patient change. While anchor-based methods have been the most frequently used, researchers may need to utilize multiple methods to establish meaningful within-patient change. However, multiple methods often result in a range of values for what is a clinically meaningful benchmark. Therefore, there is a need for a triangulation of evidence, which consists of examining a range of values to converge on an appropriate value or range of values likely to represent meaningful change in the outcome of interest. Descriptive analyses (such as CDF displays) can be helpful for understanding where these ranges fall. As part of this triangulation of evidence, it is essential to keep in mind the disease severity of patients, and to consider not just improvement, but also non-worsening of symptoms. Lastly, it is critical to start with an outcome assessment that is reliable and valid before utilizing it to ascertain whether a change is meaningful.

**Priorities for Future Discussion**

Some areas for future research and discussion include better characterizing the context in which to employ each of these methods in a clinical trial. Each of these methods have statistical, operational, and logistical challenges that need to be further addressed, and it will be important to evaluate when it is appropriate and worth the additional effort to include these emerging methods. The list of methods chosen for this workshop was not exhaustive, and there may be other emerging methods that can be useful for interpreting meaningful within-patient change.
There were also open questions regarding how to leverage each of these methods across COA types. Participants generally accepted that anchor-based methods can be used across all COA types, but it was not as evident when the emerging methods might be appropriate for other COAs. If a patient is being asked directly about whether they believe a change is meaningful or not, participants questioned whether it would be possible to utilize anything other than a PRO. Further testing of each of these and other emerging methods will be needed to clarify these issues.

Lastly, participants expressed confidence in the potential for utilizing technology to help address some of these questions. There may be opportunities to use technology, such as wearable fitness trackers or mobile device apps, to gather clinical data through both passive and active avenues. These data could help researchers as they try to identify study designs and study endpoints that can tackle the question of meaningful change.

---


