Qual Life Res (2018) 27:33–40
https://doi.org/10.1007/s11136-017-1616-3

SPECIAL SECTION: MEASURING WHAT MATTERS (BY INVITATION ONLY)

Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores

Cheryl D. Coon[1] · **Karon F. Cook**[2]

Accepted: 6 June 2017 / Published online: 15 June 2017

Abstract

Purpose Clinical outcome assessments (COAs) require evidence not only of reliability, validity, and ability to detect change, but also a definition of what constitutes a meaningful change on the instrument. The responder definition specifies the amount of change on the COA that may be interpreted as a treatment benefit and is critical for interpreting what constitutes a meaningful change on the COA scores. However, the literature that describes methods for developing and applying responder definitions can be difficult to navigate. Clear and concise guidelines regarding which methods to apply under what circumstances and how to interpret the results are lacking. This article provides a guide to the variety of available methods and issues that should be considered when establishing responder definitions for interpreting meaningful changes in COA scores.

Methods An overview is provided for selecting anchors, developing study designs, planning psychometric analyses, using psychometric results to set responder thresholds, and applying responder thresholds in demonstrating treatment efficacy.

Results There are a variety of anchor-based methods for consideration, but they all rely on a preference for strongly related and easily interpretable anchors. The benefits of applying multiple anchors and multiple analytic methods are discussed. The process of triangulation can synthesize results across multiple sources to gain confidence in a proposed responder definition. Though a link to meaningfulness from the patient's perspective is absent, distribution-based methods provide lower bound estimates of score precision and have a role in triangulation. Responder definitions are typically required within regulatory review, but their application may differ across clinical trial programs.

Conclusions By careful planning of anchor selection, study design, and psychometric methods, COA researchers can establish defensible responder thresholds that ultimately aid patients and clinicians in making informed treatment decisions.

Keywords Clinical outcome assessment · Patient-reported outcome · Score interpretation · Responder definition · Meaningful change

Introduction

Advancements in the science of clinical outcome assessments (COAs) have resulted in measures that yield increasingly precise scores. With such precision, it is easier to identify statistically significant differences between groups (such as between study arms in a clinical trial). However, statistical significance is not synonymous with clinical meaningfulness. The field of COAs, and more specifically the field of patient-reported outcome (PRO) measurement, has long recognized the distinction between *statistically significant* differences and *meaningful* differences between scores. Also recognized is the distinction between individual- and group-level guidelines for what constitutes meaningful differences. The distinction has real-world consequences. The proportion of clinical trial subjects who responded to treatment is helpful interpretive information for patients and their physicians when

✉ Cheryl D. Coon
ccoon@outcometrix.com

[1] Outcometrix, PO Box 890, Essex, MA 01929, USA

[2] Northwestern University, Chicago, IL, USA

reviewing a drug label to decide if the drug is the right treatment for them. The opportunity to report that information requires a definition of what it means to be a treatment responder on the COA. The purpose of this paper is to present current methods for setting thresholds for use in interpreting change in individual-level COA scores. While this topic is not new to the COA field, a succinct resource for researchers is needed to assist in study design, writing psychometric analysis plans, and interpreting results, particularly in the context of demonstrating treatment efficacy in clinical trials.

The need for strong anchors and appropriate study designs

A responder definition quantifies "the individual patient PRO score change over a predetermined time period that should be interpreted as a treatment benefit" [1]. There are a number of statistical methods for developing a responder definition. A key aspect of these methods is use of an external criterion (usually referred to as an "anchor" measure) to identify subjects who have experienced a treatment benefit on the outcome being assessed. Using these methods, a responder threshold identifies the magnitude of COA score change experienced by those who report a meaningful change based on the anchor.

The validity of anchor-based estimates of responder thresholds requires that scores on the anchor assess the same or a similar construct as that measured by the COA. If this assumption is not met, the responder threshold estimate is based on an extraneous concept, and, is therefore invalid. For example, suppose a COA measures disease severity and the anchor measures physical functioning. While physical functioning is certainly related to disease severity, for many conditions, an improvement in physical functioning is unlikely to correspond directly with a lessening of disease severity. The appropriateness of an anchor can be evaluated using both quantitative and qualitative methods. When possible, interviews with patients should be conducted to confirm that the anchor is interpretable and corresponds to what they consider meaningful change. Further, the anchor should "be easier to interpret than the PRO measure itself" [1].

To support the appropriateness of an anchor, the correlation between the anchor and the COA should be reported. While there is currently no consensus in the field regarding how strong the relationship should be, some psychometricians have suggested that the correlation be at least in the range of 0.30–0.40 [2–4]. However, stronger correlations lend greater confidence in the anchor's classifications [5].

Recently, consensus has grown for measuring anchors concurrently with the COAs for which they are being used as an external criterion. The patient global impression of change (PGIC) once was the most commonly used anchor for estimating responder thresholds. At the end of treatment, subjects rated the amount of change they had experienced since before the study began (e.g., "much improved" to "much worse"). However, research has called into question the ability of patients to accurately recall their pre-treatment after weeks, months, or even years [3, 6–8]. An alternative to the PGIC is the patient global impression of severity (PGIS). In this method, subjects rate their current condition (e.g., as "very severe," "severe," "moderate," "mild," or non-existent) both pre-treatment (at the start of the study) and at the end of treatment. The score difference between the two PGIS assessments serves as the estimate of how much subjects' conditions have changed. Again, cognitive debriefing on the anchor itself can help to justify the amount of change on the PGIS that indicates a meaningful change (e.g., a one-category improvement).

A self-reported anchor will likely produce the most accurate and relevant data for meaningful change because it is based on the patients' direct experiences of the symptoms and/or functioning associated with their conditions. In some therapeutic areas (e.g., schizophrenia), the clinician global impression of change (CGIC) scale may be substituted for the PGIC to obtain a clinical judgment of the patient's condition. However, unless there is impairment associated with the condition that would likely render the patient's feedback unreliable, a (suitable) patient-reported anchor is always preferable. This is true regardless of the type of COA being used to construct an endpoint because a patient-reported anchor links directly to the patient's experience.

COA developers should not feel limited to using a single anchor measure. In fact, using multiple anchors can be advantageous, particularly because the anchor will never be as valid as the COA itself. When a given anchor turns out to be problematic, perhaps having a weak correlation with the COA, other anchors may help refine the responder threshold estimate. Even when an anchor proves to be well correlated with the COA, additional anchors may build confidence around the proposed threshold. If one moderate-to-weak anchor points to a certain threshold, then researchers may wonder if that threshold is truly representative of meaningful treatment benefit. However, if several imperfect anchors all correspond to the same or proximate threshold location, researchers would have more confidence in applying that threshold or range of threshold values to the interpretation of the study results. If multiple anchors do not converge on the same threshold or range; however, then the results should be evaluated in light of the appropriateness of each anchor in relation to the COA (e.g., anchor wording, strength of correlation, reporter).

A requirement for use of anchor-based methods is that the anchor and COA be administered longitudinally so that changes in patients' conditions can be observed. Anchor-based methods require, at a minimum, a pre-treatment and a post-treatment assessment. An effective intervention also is required, otherwise, there would be no change on which to anchor. However, the longitudinal study should not be used for both establishing a responder definition and evaluating treatment efficacy using that responder definition, as this could bias results.

Methods for analyzing anchor-based data

Correlations and scatterplots

Analysis should begin with calculations and plots that help the researcher "get to know" the data. To gauge the strength and nature of the relationship between the COA and the chosen anchor, correlations can be calculated using methods appropriate for the types of data generated by the COA score, anchor score, and changes in those scores (e.g., interval level, ordinal level). As stated earlier, the relationship between a measure and an anchor should be strong enough to give confidence that the anchor can distinguish among subjects who have, and those who have not, experienced a treatment benefit. Initial analyses should include scatterplots of COA scores versus anchor scores at each assessment time point, as well as scatterplots of changes in COA scores versus changes in anchor scores from baseline to the end of treatment. Scatterplots allow examination of the spread of the observations within each anchor group. If change in COA score varies widely within a given anchor group, using the anchor to set a responder threshold would produce a large number of false positives and false negatives regardless of the analysis method (Fig. 1). While misclassifications are unavoidable, especially because the anchor can seldom be considered a "gold standard," scatterplots can evaluate the quality of an anchor prior to using it for establishing responder thresholds.

ANOVA

The traditional approach for analyzing anchor-based data is through analysis of variance (ANOVA). Mean (or median) changes in COA scores are computed for each anchor group and for the target anchor category that identifies the minimum change used to define the responder threshold (e.g., one-category improvement on the PGIS from pre-treatment to post-treatment). While this is a common and simple approach, by definition the method misclassifies about half of the target group (Fig. 2) [9]. Additional analyses should evaluate misclassifications to determine if the responder threshold should be adjusted. Thus, ANOVA-based methods are an appropriate first step, but researchers would be ill-advised to stop here, as there is much more to learn about their data.

Empirical cumulative distribution function plots

Researchers familiar with the FDA PRO Guidance are aware of cumulative distribution functions (CDFs) as an alternative to specifying a single responder threshold [1]. A CDF displays the probability of a variable (e.g., COA change score) taking on a value of X or greater at each point along the variable's continuum. A plot of the CDF for each treatment arm presents the probability of achieving each COA change score. However, these CDF plots are used for evaluating the efficacy of treatment and are not appropriate for setting a responder threshold on a COA.

ECDFs ("e" for empirical, indicating that it based on observed percentages rather than a fitted probability function), or, alternatively, probability density functions (PDFs), have recently been used for establishing responder

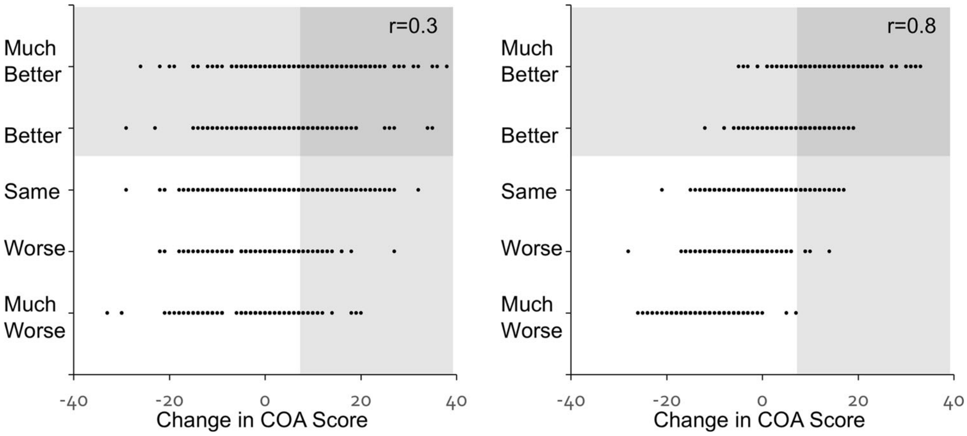

Fig. 1 Example scatterplots for weak (*left*; $r = 0.3$) and strong (*right*; $r = 0.8$) anchors against COA score changes. If the target anchor category is "Better", and the true responder threshold is 5, then there are more false classifications in the *top left* and *bottom right* quadrants (*light gray*) when the anchor is weak than when it is strong

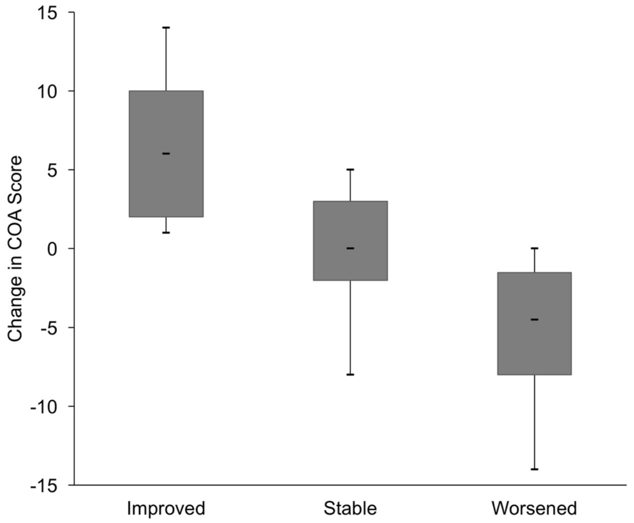

Fig. 2 Box-and-whisker plots of COA score changes in each anchor category. Note that while the center of the improved group is 6 points, if the responder threshold was set here, it would classify approximately half of the improved group as non-responders. In fact, the threshold could be set lower and still exclude the majority of the stable group

definitions. Instead of plotting one curve per study arm, the plot displays one ECDF curve per anchor category. ECDF plots display the distribution of COA score changes among subjects who experienced different levels of change (Fig. 3). This method, similar to ANOVA, graphically displays the center and spread of each anchor group's scores. ECDF plots display the impact of choosing various points along the COA score continuum as a responder threshold.

Plotting ECDFs is an exploratory method without established interpretation guidelines. One option is to start by identifying the point along the *x*-axis (i.e., change in COA score) that corresponds to 50% on the *y*-axis for the target anchor category (50% of the target group achieved that change score or higher). While this approach is similar to ANOVA in its focus of the center of the distribution, ECDF plots show the cumulative proportions observed in all anchor groups across the COA score continuum. Possible observations include lack of adequate separation between the curve for the target anchor group and the curve for the group reporting no change. This would suggest that the (absolute) magnitude of the threshold should be increased, perhaps looking at a higher cumulative proportion location instead (e.g., 75%).

Receiver operator characteristic (ROC) curves

ROC curves are another common anchor-based method. The sensitivity (the proportion of "true" responders according to the anchor that are correctly identified as responders) of each score change is plotted against one minus its specificity (the proportion of "true" non-responders according to the anchor that are correctly identified as non-responders) (Fig. 4). The point that maximizes both sensitivity and specificity (i.e., closest to the top left corner of the plot) is often selected as the responder threshold. However, the location of this point is not always

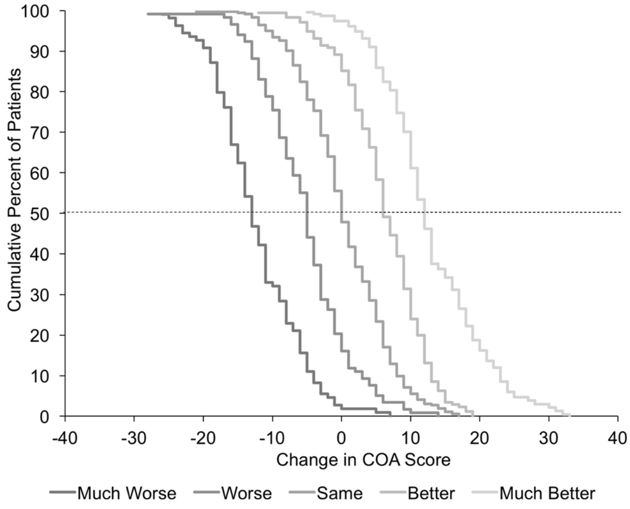

Fig. 3 ECDF curves for each anchor category along the COA score change continuum (*x*-axis). Half of the target anchor group ("better") reported score changes of 6 or higher, while that threshold was exceeded by only 17% of the stable group. Shifting the responder threshold lower to 4 points may be desirable, where 66% of the better group would be classified as responders, while only 29% of the stable group would be classified as responders

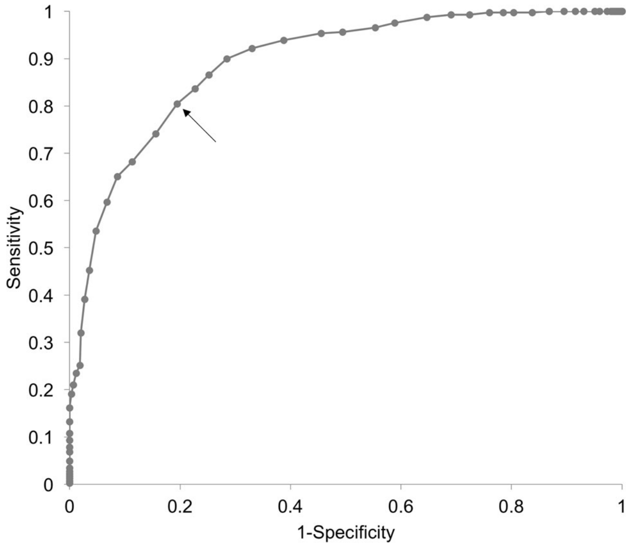

Fig. 4 ROC showing the tradeoff between sensitivity and specificity. Each point along the line is a different COA score change. The *arrow* indicates the point that maximizes both sensitivity and specificity (a change score of 5)

obvious, especially when the correlation between the anchor and COA is low. Further, this method gives equal weight to accurate identification of true responders and of true non-responders. While it would be ideal to maximize both, there are tradeoffs. The point that maximizes both sensitivity and specificity may not correspond to the most meaningful threshold or range along the COA change score scale. The researcher must balance costs and benefits within the context of use. For example, increasing the magnitude of the responder threshold may increase the risk of keeping an effective drug from coming to market, while setting the threshold too low could result in an ineffective drug being brought to market. In practice, the ROC may be more useful for evaluating candidate responder thresholds or ranges rather than as a method for identifying new ones.

Classification statistics

A method for evaluating the balance between a too high and a too low threshold is to consider other classification statistics. The positive predictive value (PPV) is calculated based on those subjects whose COA score change exceeded the responder threshold. It defines the probability that those subjects who exceeded the threshold were "true" treatment responders according to the anchor. The negative predictive value (NPV) is calculated based on those subjects whose COA score change was less than the responder threshold. It defines the probability that those subjects whose change was less than the threshold for "true" non-responders was based on the anchor. The PPV and NPV may be more relevant than sensitivity and specificity because these values correspond to a practical question: Given that a subject's COA score change exceeded the responder threshold (defined a priori), how confident can we be that this person truly experienced a treatment benefit? A plot that simultaneously considers all four classification statistics (sensitivity, specificity, PPV, NPV) can be useful in identifying a justifiable range of values for the responder threshold. It can be used to identify a location along the COA score change continuum at which the most relevant statistics are maximized and the impact of moving the threshold in either direction is evident (Fig. 5). Intersection of all lines in the same region of the y-axis would warrant confidence in a responder threshold selected from the area of the x-axis corresponding to this intersection. When the lines do not intersect closely, then the tradeoff between classification statistics depends on factors such as the therapeutic area and the drug's safety profile (e.g., NPV might be more important to consider for an orphan disease; PPV might be more important to consider for a drug with undesirable adverse effects).

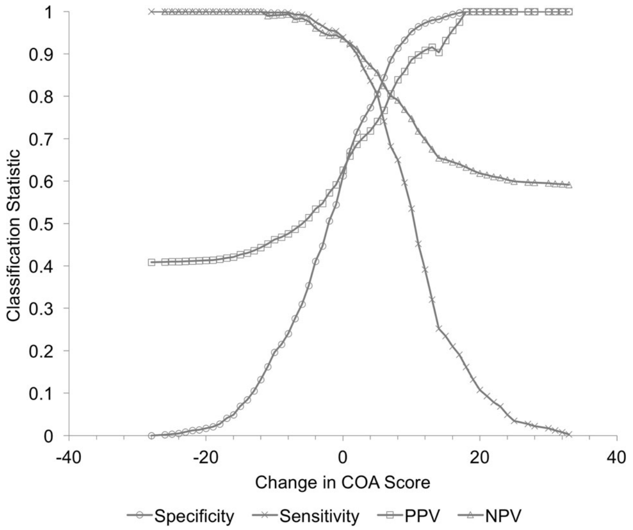

Fig. 5 Sensitivity, specificity, PPV, and NPV along the COA score change continuum. The four classification statistics happen to converge on a point between 7 and 8, establishing a potential range for the responder definition

Discriminant analysis

Discriminant analysis recently has been proposed as a method for defining responder thresholds [10, 11]. In this approach, the probability of a subject being in a particular anchor category given their COA score change (Fig. 6) is plotted. Similar to item response theory curves, the plot can be interpreted by identifying the location along the x-axis (the COA score change continuum) where membership in the target anchor category is most likely (i.e., where the

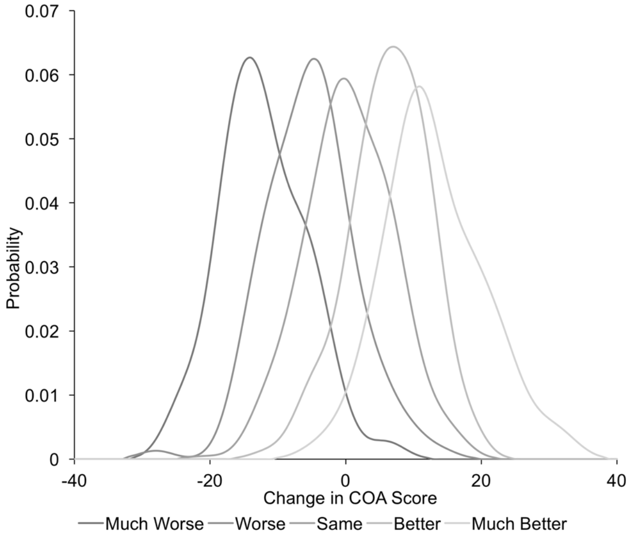

Fig. 6 Discriminant analysis plots showing the probability of anchor group membership by the COA score change. Membership in the target anchor group ("better") is most likely at a COA score change of 7. The responder threshold may be shifted as low as 3, below which point membership in the stable category becomes more likely

curve is highest). Such plots can be used to identify the location where adjacent curves intersect—the point along the *x*-axis where a lower anchor category (e.g., no change) ceases to be most likely and the target anchor category becomes most likely. The value is a lower bound estimate for the responder threshold. The plot is most easily interpretable when the target anchor category curve is steep. When the curves for anchor categories overlap and/or when the curves are less steep, the plots are less informative. Discriminant analysis may provide supplementary support for a specific responder threshold location, especially for researchers who prefer to think in terms of probabilities and fitted distributions.

The role of distribution-based analyses

According to the 2009 FDA PRO Guidance, anchor-based methods provide empirical evidence to justify the location of a responder threshold [1]. Evidence generated from distribution-based analyses; however, is considered only as supportive and supplementary because the results are not linked to the meaningfulness of score changes from the patient's perspective. In distribution-based methods, score changes are considered in the context of the variability and reliability of the scores themselves. For example, the standard error of measurement (SEM) adjusts the standard deviation of the COA scores at baseline by the reliability of the scores so that scores are evaluated in a metric that is similar to an effect size. COAs with higher reliability have a lower SEM, which allows for smaller COA score changes to be detected [12]. Thus, the SEM identifies the threshold below which a score change would be considered unreliably small and statistically indistinguishable from no change. As such, the SEM and other locations identified by distribution-based methods serve as lower bound responder threshold estimates. Distribution-based methods are useful for evaluating whether a proposed anchor-based threshold can be reliably measured by the COA, but they should not be used alone to set the responder threshold.

The role of score interpretation in regulatory review

The methods discussed above are appropriate for identifying responder threshold estimates from a range of perspectives (e.g., health authority, clinicians, patients) and environments (e.g., drug approval, reimbursement, clinical practice, population health). As this issue of Quality of Life Research is focused on developing COA instruments that meet regulatory requirements for labeling; however, the role of score interpretation in regulatory review merits special consideration.

The 2009 FDA PRO Guidance recommends examining changes in individual patient scores as part of the sponsor's evaluation of a drug's efficacy, safety, and/or tolerability [1]. The responder definition for the COA is determined a priori and is used to construct an efficacy endpoint in a clinical study. Responder definitions allow evaluation of whether a large majority of those on treatment experienced a treatment effect. Though interpretation of group-level differences in score change has intuitive and practical appeal (such as in power analyses when planning clinical trials and in reimbursement decisions), it is not specifically mentioned in the 2009 FDA PRO Guidance and, thus, is not discussed in this manuscript.

The term "responder definition" invokes the image of a single line drawn in the sand, but in practice, identifying a *range* of scores can be more appropriate than estimating a single value. When there is uncertainty regarding where along the score continuum a treatment benefit becomes meaningful, it can be appropriate to report a *responder range*—i.e., the range of scores within which it is reasonable to define an individual as a "responder." A responder range is useful for gauging the potential meaning of score changes and reflects the fact that there is no single, "true" responder threshold.

In the regulatory context, the preference for statistical analysis based on the responder definition (i.e., responder analysis) depends on the review division within FDA. Some divisions prefer a formal responder analysis in which significance testing is based on the null hypothesis that the proportion of subjects in the treatment group(s) who reported a meaningful benefit is equal to the proportion that did not. For example, a *Chi* square test could be used to compare responder rates. Other divisions may prefer analysis of continuous group-level differences in score changes. In this case, the responder definition can be used as supplementary and interpretative, describing the responder rate in each study arm. Of course, analyzing continuous data (i.e., continuous group-level differences in COA score changes) yields more statistical power than dichotomizing the COA data into groups of responders and non-responders. In the PhRMA position paper on responder analyses, retention of the continuous-level data for efficacy analyses was recommended [13]. The onus is on individual sponsors to communicate with the appropriate FDA review division(s) early and often in the drug development process to ensure that the statistical analysis plan meets the division's expectations. Regardless of how the responder definition is utilized in FDA review, sponsors are urged to establish a responder definition prior to entering Phase 3 so that it can be incorporated into the clinical trial analysis plan. Use of Phase 3 data for establishing responder thresholds is discouraged, as the thresholds must be generalizable and unbiased by the efficacy data.

Putting this in practice

As has been conveyed in this paper, establishing responder definitions for interpreting scores on COAs is an important but complex step in instrument development. Methods for establishing thresholds should be considered early in the study design to ensure that the resulting data will be appropriate for anchor-based analyses. Applying multiple methods for estimating thresholds yields results that can be triangulated to gain greater confidence in a threshold score or range of scores. Relying on a single method to establish a responder definition is a simple, but also simplistic approach. Evaluating thresholds based on multiple anchors and analytic methods can be triangulated, yielding confidence in responder definitions. The goal of triangulation is to hone in on a defensible threshold value or range of values using multiple sources of information. Though multiple methods may provide a range, it may not necessarily be a definitive one. There will be false positives and false negatives no matter where the threshold is set. Therefore, reconciliation of locations suggested by different methods should focus on the implications of incorrect classification. Responder definitions require consideration of the therapeutic area and benefit–risk profile of the treatment being considered. Researchers need to carefully consider the data and determine what is most defensible for their context of use.

While a single responder threshold is easiest to apply, it may not be an adequate representation, and the data may not support a single value. In such a case, flexibility is needed until further data can be collected or the threshold can be evaluated in practice. What is meaningful to patients also may depend on the specific patient. Some patients may see benefit in a small improvement, while others may demand full resolution of symptoms or return of normal functioning to say that a treatment works. By providing a range, the interpretation can be left to the reviewer, be it the regulatory agency, payer, clinician, or patients themselves.

An example from a product label

Recently, Xermelo^{TM} was approved by the FDA for the treatment of carcinoid syndrome diarrhea based on a primary efficacy endpoint of change in number of daily bowel movements (BMs) over the treatment period [14]. While the continuous-level COA analysis produced statistically significant results, the FDA sought to ensure clinical meaningfulness of the results at the patient level [15]. In response, the sponsor proposed a responder definition based on anchor-based analyses using six COAs. The FDA excluded results from two of the COAs that were judged to be more difficult to interpret than the BM measure itself. For the other four COAs, the FDA reviewed means, effect sizes, ECDFs, and EPDFs for change in BM frequency for each of the anchor categories [16]. The correlations between the BM measure and the anchors were small to moderate ($r = -0.23$ to -0.57), and the sample size modest ($n = 9$–19 in the target anchor categories), but the median change in BM frequency

Table 1 Good practices for interpreting change on COAs

Key topic	Suggested practice
Study design	Ensure that the COA and anchor are administered in a temporally appropriate manner. For example, if the PGIS is the anchor measure, then the COA and the PGIS ought to be administered concurrently at the beginning of the study before treatment begins and at end of the treatment (at a minimum)
	Ensure that the study design is appropriate for determining a responder threshold (e.g., longitudinal study with observed change)
Anchor selection	Select at least one anchor (ideally, a PRO instrument) that measures the same concept as the COA. Generally speaking, anchors that measure a patient's *current* condition or disease state (i.e., with no recall required) are preferred to ones that ask the reporter to think about how the patient's condition has changed over time
Psychometric methods	Report the correlation between the anchor scores and the COA score changes, being sure to select a correlation measure appropriate for use with the datatypes generated by the anchor scores on the one hand and the COA score changes on the other
	Consider using multiple anchors as well as multiple anchor-based methods to gain confidence in the responder definition
Regulatory planning	When the COA is being used to construct a primary, co-primary, or secondary efficacy endpoint in a clinical trial:
	-Determine the responder threshold prior to beginning the Phase 3 trial (if possible)
	-Speak with the regulatory agency about how the responder definition should be applied to the interpretation of clinical trial results

was consistently 2 BMs/day across the patient-reported anchors. Thus, though there were data inadequacies, triangulation across multiple anchors created confidence in the responder definition. As a result, the FDA allowed a drug label that included a CDF showing the proportion of patients in treatment and placebo groups who experienced different reductions in BM frequency [14]. The CDF includes a vertical line highlighting the proportion of patients who achieved an average reduction of 2 BMs/day. The proportions in each treatment arm also are reported in text. FDA communication with the sponsor stated that the CDF was included “to facilitate health care provider’s interpretation of the population mean change in bowel movements reported in the label” [15]. Indeed, the inclusion of this material supports patients’ and clinicians’ understanding of research on the drug’s impact—namely, that Xermelo™ significantly reduced BM frequency as compared to placebo, and 33% of patients randomized to Xermelo experienced at least a 2 BM/day reduction. Though this particular drug review did not require a responder analysis as part of the primary endpoint, descriptive reporting of the responder rate was deemed advantageous for helping clinicians interpret the primary endpoint results.

Summary

There are numerous approaches for establishing responder definitions for interpreting change on COA scores, and researchers should be discouraged from applying a one-size-fits-all approach. Study designs and methods should be selected that make the most sense in the context of use, and they should focus on obtaining information that crafts a credible story for interpreting the data. Using good practices for interpreting change on COA (Table 1), researchers can be confident that their efforts will provide the information needed to transform COA scores from abstract numbers to a meaningful metric.

Acknowledgements This manuscript is based on work presented at the DIA 2016 Annual Meeting, Philadelphia, PA, June 26–30, 2016 moderated by Ms. Marian Strazzeri with Drs. Scott Komo and Cheryl Coon as panelists. The authors thank Ms. Strazzeri and Dr. Komo, as well as Drs. Laura Lee Johnson and Wen-Hung Chen, for their contribution and insight during the development of the forum session and for their feedback on this manuscript.

Funding The authors received no funding for this work.

Compliance with ethical standards

Conflict of interest Dr. Coon declares that she has no conflict of interest. Dr. Cook declares that she has no conflict of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

References

1. US Food and Drug Administration. (2009). Guidance for industry on patient-reported outcome measures: Use in medical product development to support labeling claims. *Federal Register, 74*(235), 65132–65133.
2. Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology, 61*(2), 102–109.
3. Hays, R. D., Farivar, S. S., & Liu, H. (2005). Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *COPD, 2*(1), 63–67.
4. Coon, C. D., & Cappelleri, J. C. (2016). Interpreting change in scores on patient-reported outcome instruments. *Therapeutic Innovation & Regulatory Science, 50*(1), 22–29.
5. Coon, C. D. (2016). Telling the interpretation story: the case for strong anchors and multiple methods. Plenary presentation at the 23rd annual conference of the International Society of Quality for Life Research; October 2016. Copenhagen, Denmark.
6. Norman, G. R., Stratford, P., & Regehr, G. (1997). Methodological problems in the retrospective computation of responsiveness to change: The lesson of Cronbach. *Journal of Clinical Epidemiology, 50*(8), 869–879.
7. Wyrwich, K. W., Norquist, J. M., Lenderking, W. R., Acaster, S., & The Industry Advisory Committee of International Society for Quality of Life Research (ISOQOL). (2013). Methods for interpreting change over time in patient-reported outcome measures. *Quality of Life Research, 22*(3), 475–483.
8. Nixon, A., Doll, H., Kerr, C., Burge, R., & Naegeli, A. N. (2016). Interpreting change from patient reported outcome (PRO) endpoints: patient global ratings of concept versus patient global ratings of change, a case study among osteoporosis patients. *Health and Quality Life Outcomes, 14,* 25.
9. Fayers, P. M., & Hays, R. D. (2014). Don’t middle your MIDs: Regression to the mean shrinks estimates of minimally important differences. *Quality of Life Research, 23*(1), 1–4.
10. Gerlinger, C., Schumacher, U., Faustmann, T., Colligs, A., Schmitz, H., & Seitz, C. (2010). Defining a minimal clinically important difference for endometriosis-associated pelvic pain measured on a visual analog scale: analyses of two placebo-controlled, randomized trials. *Health and Quality Life Outcomes, 8*(1), 138.
11. Gerlinger, C., & Schmelter, T. (2011). Determining the non-inferiority margin for patient reported outcomes. *Pharmaceutical Statistics, 10*(5), 410–413.
12. Wyrwich, K. W., Bullinger, M., Aaronson, N., Hays, R. D., Patrick, D. L., & Symonds, T. (2005). Estimating clinically significant differences in quality of life outcomes. *Quality of Life Research, 14*(2), 285–295.
13. Uryniak, T., Chan, I. S. F., Fedorov, V. V., et al. (2011). Responder analyses—a PhRMA position paper. *Statistics in Biopharmaceutical Research, 3*(3), 476–487.
14. Xermelo [package insert]. (2017). *The Woodlands*. Texas: Lexicon Pharmaceuticals, Inc.
15. US Food and Drug Administration, Center for Drug Evaluation and Research. Xermelo NDA 208794 summary review, February 28, 2017. Retrieved May 8, 2017, from https://www.accessdata.fda.gov/drugsatfda_docs/nda/2017/208794Orig1s000SumR.pdf.
16. US Food and Drug Administration, Center for Drug Evaluation and Research. Xermelo NDA 208794 statistical review and evaluation, clinical outcome assessment, November 29, 2016. Retrieved May 8, 2017, from https://www.accessdata.fda.gov/drugsatfda_docs/nda/2017/208794Orig1s000StatR.pdf.