Published in final edited form as:

Expert Rev Pharmacoecon Outcomes Res. 2011 April ; 11(2): 163–169. doi:10.1586/erp.11.12.

Interpreting patient-reported outcome results: US FDA guidance and emerging methods

Lori D McLeod^{†,1}, Cheryl D Coon¹, Susan A Martin¹, Sheri E Fehnel¹, and Ron D Hays²

¹RTI Health Solutions, 3040 Cornwallis Road, PO Box 12194, Research Triangle Park, NC 27709-2194, USA

²UCLA Department of Medicine, Los Angeles, CA, USA

Abstract

In recent years, the US FDA has become more critical of instruments used to measure patient-reported outcomes (PROs) in clinical trials. To facilitate decisions related to the approval of drugs, labels and promotional claims based on PROs, the FDA created the Study Endpoints and Label Development (SEALD) group. SEALD has developed a PRO guidance related to the use of PRO measures used to support drug approvals and label claims, including recommendations for establishing thresholds for meaningful change at the individual level (i.e., defining a responder). This article examines in detail the FDA-recommended methodology for defining a responder and analyzing responder-based PRO measure results. We also present other responder analysis approaches for consideration in furthering the precision and interpretation of this methodology.

Keywords

individual change; minimal important differences; reliable change index; responder

In recent years, the US FDA has become more critical of instruments used to measure patient-reported outcomes (PROs) in clinical trials. Specifically, the agency now requires a higher degree of scientific rigor in the development and psychometric evaluation of these measures, particularly those intended to support drug approval or labeling claims.

To facilitate decisions related to the approval of drugs, labels and promotional claims based on PROs, the FDA created the Study Endpoints and Label Development (SEALD) group. Members of SEALD are often included in pivotal meetings with study sponsors to provide feedback regarding proposed PRO endpoints. SEALD also serves as an advisory group to all reviewing divisions of the FDA and has developed both a draft and final version of a guidance related to the use of PRO measures used to support drug approvals and label claims.

The draft version of the Guidance for Industry Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims, released in February 2006, focused on describing both the recommended development of a PRO measure and the psychometric properties for which evidence must be presented if the measure is to support

regulatory approval or promotional claims [101]. Both the draft version of the PRO guidance and its successor clearly stipulate that any PRO measure used to support drug approval or label claims must be developed with extensive input from patients and thoroughly evaluated in the population involved in the clinical trials.

The final version of the PRO guidance, released in December 2009, focuses less on the development of PRO measures and describes in detail how the FDA reviews and evaluates existing, modified or newly developed PRO instruments to support drug approvals and product labeling [102]. Additional key modifications include an increased emphasis on documentation supporting the use and content validity of PRO measures, definitions and examples of key components of this documentation, and more explicit guidance related to the evidence that must be submitted to facilitate the interpretation of results pertaining to PROs.

Relevant to the focus of this article, the most significant modification between the two versions was the elimination of the term ‘minimal important difference’ (MID) from the final PRO guidance. Specifically, in the draft PRO guidance, MID was defined as ‘the amount of difference or change observed in a PRO measure between treatment groups in a clinical trial that will be interpreted as a treatment benefit’ [101]. The draft PRO guidance suggested “it would be logical to establish the null hypothesis to rule out a difference less than or equal to the MID” but also acknowledged that in practice this was rarely carried out [102]. In the final PRO guidance, the elimination of MID is accompanied by an emphasis on establishing meaningful change in PRO measures at the individual level (i.e., defining a responder) versus at the treatment group level. This new focus is demonstrated by the definition provided for a responder in the final PRO guidance as being “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” [102]. This article examines in detail the FDA-recommended methodology for defining a responder and analyzing responder-based PRO measure results. We also present other responder analysis approaches for consideration in order to further the precision and interpretation of this methodology in the context of interpreting the results of clinical trials and supporting eventual label claims.

Interpretation of PRO results: FDA guidance

In the final PRO guidance, the FDA recommends *a priori* responder definitions. While not labeled as such, the recommendations are consistent with approaches used to estimate MIDs. The primary approach recommended in the final guidance is an empirical anchor-based approach. The guidance recommends distribution-based methods as supportive or secondary approaches to provide additional evidence for the resulting responder definitions [102].

Anchor-based method

Researchers at McMaster University (Hamilton, Ontario, Canada) pioneered the use of self-reported retrospective measures of change as ‘external’ anchors for the estimation of meaningful change [1]. Using this method, mean changes in a PRO measure over time are compared with patient- or clinician-reported global ratings of overall change in the same construct using a balanced ordinal change scale. Specifically, in a McMaster study of asthma, the anchor contained seven categories for positive change or ‘improvement’, one category for no change, and seven categories for negative change or ‘deterioration’. To identify the responder definition, the mean change for the domains on the target PRO measure (e.g., an asthma-specific quality-of-life questionnaire) were calculated for all the patient responses within each category of change on the anchor. The mean change for the patients reporting either ‘a little improvement’ or ‘a little deterioration’ was the threshold for

important change (i.e., a responder estimate). This method is described as an anchor-based or mean change method because the PRO measure results are ‘anchored’ or defined by differences in ratings of change external to the instrument (either self- or clinician-reported). Using the responder estimate, patients obtaining change scores less than the threshold are designated as nonresponders and patients obtaining at least the threshold are designated as responders.

As recommended in the final PRO guidance, to be useful as an anchor, the measure used to interpret the PRO must be both a valid measure of change and easier to interpret than the PRO measure. In addition to global ratings, changes in related measures that have established consensus-based thresholds can be used to define the responder threshold for the PRO measure. For example, for a PRO measuring annoyance of incontinence, the PRO guidance presents a 50% reduction in incontinence episodes as measured in a daily incontinence diary as a proposed viable anchor in estimating meaningful change.

The anchor measure used to interpret the PRO should be easy to interpret and also without recall bias. In 2005, Hays and colleagues warned of possible recall problems with self-reported retrospective questions [2]. When retrospective questions are used as anchors, researchers should determine whether they reflect the baseline and also present status equally. In theory, retrospective change items should correlate positively with the post-test and have a negative correlation of equal magnitude with the pretest. In reality, retrospective self-reports tend to correlate more strongly with the post-test than they do with the pretest because current status unduly influences the retrospective perception of change. For example, Walters and Brazier found moderate correlations (mean: 0.45; range: 0.18–0.57) between responses to a retrospective measure of global change and the Short Form (SF)-6D at follow-up across nine studies [3]. Correlations with the initial assessment were systematically lower (mean: 0.22; range: 0.01–0.41). Thus, the relationships between the anchor questions and the PRO measure of interest should be evaluated to help with interpretations of change.

Distribution-based method

The final PRO guidance describes the anchor-based method as the primary method for establishing responder thresholds, but also mentions distribution-based methods as providing support for defining responder thresholds. Distribution-based methods identify the raw score change on a PRO measure that will produce a prespecified effect size. The most common effect size specified is the half standard deviation (0.5 SD). Another distribution-based approach is the standard error of measurement (SEM):

$$SEM = SD \sqrt{1 - \alpha}$$

where SD is the SD of the domain score at baseline and α is the reliability estimate of the scale of interest at baseline [4]. Others have argued, and we agree, that distribution-based methods provide supportive, rather than primary, evidence of meaningful change [2].

Interpretation of PRO results: emerging quantitative methods

Cumulative distribution functions

New in the final version of the PRO guidance is the concept of presenting ‘the entire distribution of responses for treatment and control group’ using a cumulative distribution function (CDF). A basic concept in statistics, the CDF shows a continuous plot of the proportion of patients at each point along the scale score continuum who experience change at that level or lower. Such an approach amounts to calculating the percentage of responders

at each value of the PRO change score, if that value were considered the responder threshold. This technique has been used periodically in clinical research but has been met with increasing favor since it was proposed in its application of comparing treatment groups at any appropriate responder level [5]. This approach offers the benefit of visually comparing the separation between groups across all levels of the PRO change score so that a variety of responder definitions can be considered simultaneously. Support for clinical efficacy is implied by greater separation of the CDFs in the hypothesized direction.

The consideration of CDFs is consistent with the lack of consensus on the approach for establishing a responder threshold. Different methods (i.e., anchor-based vs distribution-based) may suggest different responder definitions, and CDFs allow all proposed responder definitions to be evaluated simultaneously. Because the CDF communicates the proportion of responders at every value along the PRO change score scale, this approach allows the user to select a responder threshold appropriate for the particular application, thus ‘avoiding the need to pick a responder definition’ [102].

While CDFs are typically presented as the cumulative proportion of patients who achieve a particular value (i.e., the proportion of patients who experience a score change of X points or lower), they are most easily interpreted in the context of responder assessment when the y-axis represents the proportion of patients who are considered responders at a particular threshold value (i.e., the proportion of patients who experience a score change of X points or above). In this case, when higher change scores indicate improvement, the curves will most often be shown as decreasing from left to right because as change scores increase, fewer patients achieve improvement at that value or higher. This format, an example of which is shown in Figure 1, would favor treatments where the curves are shifted to the right of the control group curve.

Cumulative distribution functions are often presented where observed points are connected with a straight line, essentially interpolating the probability of response when values are unobserved. A more conservative approach would be to connect the points as a step function, so that the probability of response does not increase until the next higher value is observed.

While the CDF is a useful method for simultaneously evaluating a range of responder thresholds using visual evidence, the method also lends itself to statistical testing of the difference between treatment and control groups. Methods such as the chi-square test simply assess if there are significantly more responders in the treatment group than the control group at a particular level of the clinical response. This approach is consistent with the recent application of responder methods where the treatment group is considered efficacious if it has a greater percentage of responders than the control group. Such an approach would require selecting one or more responder thresholds and conducting a separate test at each level.

Other methods, such as the Kolmogorov–Smirnov test [6] or the tests of the area under the curve (AUC) [7], determine if the curves are significantly different. The Kolmogorov–Smirnov test is a nonparametric test of equality between two distributions, which evaluates if the treatment group and the control group came from the same distribution. While this test can determine if the treatment group curve is distinct from the control group curve, it evaluates only the point at which the two curves are furthest apart, so the test may not correspond to the responder threshold of interest. Alternatively, the AUC assesses whether the two curves differ using either parametric (e.g., maximum likelihood smoothing) or nonparametric (e.g., adding trapezoids) methods. However, the AUC test does not

correspond to the responder threshold of interest because it evaluates the entire distribution simultaneously.

The addition of confidence bands to the curves combines both of these approaches to determine the particular region in which the curves differ [8]. This approach is commonly used by the US Environmental Protection Agency to compare groups, although it has been suggested in the context of clinical response by Gagnon [9]. With the addition of confidence bands, a range of preliminary responder thresholds can be evaluated simultaneously. While this approach has the advantage of simultaneous assessment in the specific range of interest, it must be carefully implemented so that it results in true confidence bands rather than pointwise confidence intervals. Figure 2 shows an example CDF with pointwise 95% confidence interval (CI) bands calculated based on the survival density function from PROC LIFETEST in SAS® 9.2.

Evaluating thresholds

Another approach for assessing the appropriateness of responder thresholds combines an anchor-based approach and receiver operating characteristic (ROC) curves to identify the responder thresholds that best predict classification based on an external criterion. First, an external anchor (or another external criterion that is appropriate to identify responders) is used to classify all participants into responder or nonresponder groups. Then, to check for appropriateness, the relationship between the external criterion and the PRO measure is examined through correlation analyses. Finally, the predictive accuracy of how well the specific PRO change scores relate to the classifications is evaluated using logistic regression analyses and graphically displayed as an ROC curve. Each point on the curve provides the sensitivity (true positives) versus one minus the specificity (false positives) trade-off for identifying responders for a specific unit of change on the PRO measure. The entire range of change scores and the likelihood classifications are provided within the one curve. Figure 3 provides an example of an ROC curve with the change at zero points and at five points illustrated. At zero points, the PRO measure achieves sensitivity of 65% and specificity of 77%. At five points, the sensitivity is 83% and specificity is 56%, for this example. A diagonal line is included in the figure as a reference. Evaluations that produce curves close to the diagonal line should be viewed with caution because the diagonal line represents chance classification.

Farrar and colleagues provide an example in chronic pain [10]. In their study, data were evaluated from ten completed clinical trials investigating change in pain intensity. For each of the trials, pain intensity was measured on an 11-point pain intensity numerical rating scale (PI-NRS) ranging from 0 = no pain to 10 = worst possible pain. In addition, an external criterion based on a seven-point patient global impression of change (PGIC) in pain intensity was available. For each patient within each trial, changes in the PI-NRS from baseline to the endpoint were compared with the PGIC and those patients who indicated 'much improved' or 'very much improved' were defined as responders. Those patients indicating 'minimally improved', 'no change', 'minimally worse', 'much worse' or 'very much worse' were defined as nonresponders. The relationship between the PI-NRS and the PGIC was evaluated and found to be appropriate for the meaningful change evaluation and consistent across the ten trials, regardless of disease type, age, sex, treatment group or trial result. Using an ROC curve, Farrar and colleagues defined the responder threshold for the PI-NRS based on the point at the intersection of a 45° line with the ROC curve. This point mathematically defines the change where the trade-off (importance) of sensitivity and specificity is equal and where the total amount of misclassification (false positives and false negatives) is at its minimum. The results in the study using this rule supported a responder recommendation of a two point or 30% reduction in the PI-NRS, whereas prior to these analyses, a 50% reduction had been the common definition of a pain responder.

Statistical significance of change

As noted earlier in this article, anchor-based methods were originally used to define MIDs in PRO measures. In essence, the MID is the average change in the domain of interest on the target measure among the subgroup of people deemed to change by a minimal (but important) amount according to an 'anchor'. Hence, the MID estimate is the expected group mean change for people who have improved by enough (but not too much) according to the external standard.

If one uses the expected mean change among those who were deemed to have minimally important change to define responders, then approximately half the people in the group that changed a minimally important amount on the anchor will be classified as responders, assuming a normal distribution. (If the median is used to define the MID, then half the people in the subgroup will be classified as responders). That is, this approach leads to about half the people for whom the anchor indicates a minimally important amount of change to be classified as responders and half as nonresponders.

The problem with this approach is that the group average is not an appropriate threshold for individual change. We agree that using the mean change from the group as the best estimate of the amount of individual change is appropriate if there is no other information available; however, we would argue that the mean may not be significant at the individual level. Group change and individual change have different standard errors, and thus group-level estimates should not be used to define responders. A minimum criterion for a responder is that the individual improved significantly (i.e., individual change is greater than the measurement error associated with the PRO measure). There are a variety of related ways to estimate the significance of individual change and one or more of these should be used to determine if individual change is statistically significant or not. The main options for evaluating significance of individual change include the CI around the SEM and the reliable change index (RCI).

The 95% CI around the observed or estimated true score at baseline can be obtained by observed or estimated true baseline score $\pm (1.96) \times SD \times (1 - \text{reliability})^{1/2}$. If the measure is scored so that a higher score is better, then a score at follow-up that is greater than the upper end of the confidence interval represents significant individual change that represents improvement or a responder.

The RCI represents a t-test for individual change with change in the numerator and an estimate of noise in the denominator ($1.414 \times SEM$). If the t-statistic exceeds the standard cutoff for significance and change is positive, then the individual is a responder.

Hays and colleagues illustrated how the amount of change needed to be statistically significant is substantial even for highly reliable measures [11]. For example, even for a measure with reliability of 0.94, the amount of change required to reach statistical significance is about a 0.5 SD (medium effect size) for the SEM and about 0.70 SD for the RCI. Because changes need to be substantial for individual-level change to be beyond the error of measurement, using MID thresholds alone to define responders is inappropriate.

We recognize and can appreciate that we are approaching this issue somewhat differently than the current standard. The MID is not estimated for individuals. A minimum standard for saying an individual has responded (improved) should include that the change in score is statistically significant. A measure with less reliability does require a bigger change in score to yield a significant individual change because the standard error is larger and the CI around the score is wider. The average change on a measure for a group that was classified

as improved on an external anchor does not necessarily equate to change that is sufficiently large to yield confidence that individual change has occurred.

Finally, we would also like to highlight that reliability and standard error estimates tailored to where the individual is on the underlying continuum (e.g., those obtained using item response theory estimates) are also preferable to those based on group data.

Interpretation of PRO results: emerging qualitative methods

It is important to note that the final PRO guidance admits that judgment is still required when evaluating ‘whether the individual responses are meaningful.’ Evidence must be provided for the use of a particular responder threshold as an indicator of meaningful change from the patient perspective. Along with employing one or more of the quantitative methods described previously, collecting input directly from patients or clinical experts qualitatively may add to interpretation efforts. Given that patients are the ultimate consumers of new medications, in the next section we review potential approaches for directly eliciting their feedback on meaningful changes.

Patient-based methods

Eliciting input directly from patients regarding the amount of change on a PRO measure that they would consider to be meaningful can provide information to support a responder threshold value. However, qualitative approaches are not without challenges as it can be difficult for patients to describe what they would consider to be a meaningful change in relation to quantitative values, such as changes in a PRO measure score even at the item level. Another challenge arises with patients who have not experienced any change in their health status for a prolonged period of time. For these patients, it may be difficult for them to imagine a change in their condition that would be meaningful, beyond the often reported response of ‘completely cured.’ Owing to these challenges, PRO measures that may be best suited to a qualitative responder assessment are likely to be those with which patients are more familiar or comfortable with at a quantitative level. For example, patients with asthma have been able to provide changes that they would consider meaningful when presented with the concept of ‘symptom-free days’ [12]. In this study, patients were readily able to quantify the change in the number of days with no asthma symptoms that would represent a treatment success for them [12]. This same approach would probably not be as successful if patients were asked to provide numerical changes on a total PRO measure scale that ranged from 0 to 100, for instance. In addition, identifying patients for qualitative assessment that not only match the expected clinical trial population on the key inclusion and exclusion criteria but that have also experienced a change in status or have initiated a new medication for their condition in the previous 3–6 months can assist in the ability of the patients to quantify meaningful change. While it is unlikely that qualitative methods would ever be the sole basis for defining a responder, these approaches can provide a framework from which the responder definitions arrived at using quantitative methodology can be evaluated from the patient’s perspective.

Consensus-based methods

Eliciting input directly from patients or clinicians via consensus-based methods may also provide evidence to support the amount of change on a PRO measure that could be considered clinically meaningful. Standard setting techniques, such as the modified Angoff method used to define performance standards and cutoffs in large-scale educational and certification testing, can be further developed to define responder thresholds using the consensus of a patient or clinical panel. Brandon provides a general overview of the modified Angoff standard setting method where judges independently assign a 1 to items

that a borderline candidate would endorse and a 0 to items that a borderline candidate would not. The cutoff is then based on the average of the summed scores [13].

Expert commentary & five-year view

With the increased critical review of PRO measures, there is greater demand to better understand and document the value and magnitude of score changes. This article has described a full spectrum of guidelines and methods to support the interpretation of PRO results, including guidance from the FDA and emerging quantitative and qualitative methods. The choice of which approach to use to interpret change for a particular application should be based on the objectives of the study and the specific implementation of the PRO measure within the study. If the PRO measure is intended to support drug approval or a label claim, the methods outlined in the FDA final PRO guidance should be included in the evaluation of change. In addition, the emerging methods described in this article hold great promise for more robust exploration of change and should be considered in an overall plan to better evaluate and document the changes in a PRO measure that are likely to be meaningful to patients.

Acknowledgments

Financial & competing interests disclosure Ron Hays was supported by NIH/NIA Grants P30-AG028748 and P30-AG021684, and NCMHD Grant 2P20MD000182.

References

Papers of special note have been highlighted as:

- of interest
 - of considerable interest
1. Juniper ER, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific quality of life questionnaire. *J. Clin. Epidemiol.* 1994; 47(1):81–87. [PubMed: 8283197] •• Provides a description of the anchor-based approach to identifying a minimal important change for a patient-reported outcome (PRO) measure.
 2. Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *COPD: J. Chron. Obstruct. Pulmon. Dis.* 2005; 2:63–67.
 3. Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. *Health Qual. Life Outcomes.* 2003; 1:4. [PubMed: 12737635]
 4. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J. Clin. Epidemiol.* 1999; 52(9):861–873. [PubMed: 10529027] •• Provides a description of the the standard error of measurement-based approach to identifying a minimal important change for a PRO measure.
 5. Farrar JT, Dworkin RH, Max MB. Use of the cumulative proportion of responders analysis graph to present pain data over a range of cutoff points: making clinical trial data more understandable. *J. Pain Symptom Manage.* 2006; 31(4):369–377. [PubMed: 16632085] •• Provides a working example of how a cumulative distribution function can be used to evaluate a range of responder cutoff points for an 11-point numerical pain rating scale.
 6. Riffenburgh, RH. *Statistics in Medicine.* Elsevier Academic Press; NY, USA: 1999.
 7. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology.* 1983; 148(3):839–843. [PubMed: 6878708]
 8. Diaz-Ramos, S.; Stevens, DL., Jr; Olsen, AR. *EMAP statistical methods manual.* US Environmental Protection Agency Office of Research and Development, National Health Effects and

Environmental Research Laboratory, Western Ecology Division; Corvallis, OR, USA: 1996. EPA/620/R-96/002

9. Gagnon, DD. Some logical problems with tests for clinically important changes in HRQoL. Presented at: 8th Annual Conference of the International Society for Quality of Life Research; Amsterdam, The Netherlands. 7–10 November 2001;
10. Farrar JT, Young JP, LaMoreaux L, et al. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain*. 2001; 94:149–158. [PubMed: 11690728] • Provides a working example of an anchor-based approach to identifying a responder threshold for an 11-point numerical pain rating scale using a mean change method and receiver operating characteristic analyses.
11. Hays RD, Brodsky M, Johnston MF, Spritzer KL, Hui K. Evaluating the statistical significance of health-related quality of life change in individual patients. *Eval. Health Prof.* 2005; 28:160–171. [PubMed: 15851771] • Provides an overview of the evaluation of statistically significant change in health-related quality of life for individual patients.
12. Martin, S.; Stanford, R.; Dale, P.; Fehnel, S. What represents a meaningful improvement in SFD and RFD? The patient's perspective. Presented at: The 2010 European Respiratory Society Meeting; Barcelona, Spain. 18–22 September 2010;
13. Brandon P. Conclusions about frequently studied modified Angoff standard-setting topics. *Appl. Meas. Educ.* 2004; 17(1):59–88.
14. McLeod LD, Fehnel SE, Brandman J, Symonds T. Evaluating minimal clinically important differences (MCID) for the acne-specific quality of life questionnaire (Acne-QoL). *Pharmacoeconomics*. 2003; 21(15):1069–1079. [PubMed: 14596626]

Websites

101. US Department of Health and Human Services (USDHHS). Draft guidance for industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. February. 2006 www.ispor.org/workpaper/FDAPROGuidance2006.pdf
102. US Department of Health and Human Services (USDHHS). Accessed 1 December 2010 Guidance for industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. December. 2009 www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf • US FDA guidance for industry related to the development and review of PRO measures.

Key issues

- The US FDA has become more critical of instruments used to measure patient-reported outcomes (PROs) in clinical trials.
- The FDA has provided guidance on the use of PROs, including recommendations for establishing thresholds for meaningful changes.
- This article examines the methods recommended by the FDA's PRO guidance, including anchor-based approaches as primary methods for establishing responder thresholds and distribution-based approaches as secondary or supportive methods.
- In addition, this article discusses emerging methods that warrant further evaluation.

Figure 1. Sample cumulative distribution function

The data used to draw this figure were generated for illustration purposes.

Figure 2. Sample cumulative distribution function with confidence bands

The data used to draw this figure were generated for illustration purposes. CI: Confidence interval.

Figure 3. Sample receiver operating characteristic curve
Data from [14].