The Challenges Inherent With Anchor-Based Approaches To The Interpretation of Important Change In Clinical Outcome Assessments

Kathleen W. Wyrwich (kathy.wyrwich@gmail.com)
McMaster University https://orcid.org/0000-0002-8851-350X
Geoffrey R. Norman
McMaster University

Research Article

Keywords: clinical outcome assessment, patient-reported outcome, health-related quality of life, minimal important difference, anchor-based, regression, standard setting

Posted Date: May 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-386501/v1

Abstract

Purpose: Anchor-based methods have been used to derive clinical outcome assessment (COA) interpretation thresholds of meaningful change over time for understanding individual (within-patient) as well as within-group change and between-group differences. The methods explore the associations between the targeted concept of the COA measure and the concept measured by the external anchor(s), typically a global rating, chosen as easier to interpret than the COA measure. While they are valued for providing plausible interpretation thresholds, anchor-based methods pose a number of inherent theoretical and methodological conundrums.

Methods: This investigation provides a critical appraisal of anchor-based methods for COA interpretation thresholds and details a key bias in anchor-based methods that directly influences the magnitude of the interpretation threshold.

Results: Five important concerns for the use of anchor-based methods have emerged from the literature: 1) global estimates of change are consistently biased toward the present state; 2) the use of static current state global measures, while not subject to artifacts of recall, may exacerbate the problem of estimating clinically meaningful change; 3) the anchor assessment response(s) that indicates meaningful change usually involves an arbitrary judgment; 4) the calculated interpretation thresholds are sensitive to the proportion of patients who have improved; and 5) examination of anchor-based regression methods reveals that the correlation between the COA change scores and the anchor has a direct linear relationship to the magnitude of the interpretation threshold derived using an anchor-based approach. Stronger correlations yielding larger interpretation thresholds.

Conclusions: While anchor-based methods are recognized for their utility in deriving interpretation thresholds for COAs, the biases associated with estimation of the threshold using these methods may impede progress in the development of standard-setting methodologies for COAs.

Introduction

Clinical outcomes assessments (COAs), a term that encompasses patient-reported outcomes (PROs), clinician-reported outcomes (ClinROs), observer-reported outcomes (ObsROs), performance outcome (PerfOs), as well as certain COAs derived from technologies such as mobile health technologies [1], are crucial to interpretation of clinical studies as they uniquely describe or reflect how patients feel or function. It is essential to understand the COA change over time that is meaningful to patients to appropriately interpret clinical study findings.

Early pioneers in the quest to incorporate the patient voice into clinical studies and the need for meaningful interpretation of COA results used global patient-reported ratings of change as external anchor items to understand patients' overall assessment of change [2, 3]. The anchor-item responses were then used to classify patients' individual PRO domains/instrument change scores to inform the interpretation of clinical study results. Jaeschke et al noted "Despite the absence of a criterion measure,

establishing the meaning of changes in a new measure requires some sort of independent standard. Global ratings represent one credible alternative. ...This information will be useful in interpreting questionnaire scores, both in individuals and in groups of patients participating in controlled trials." [3]

Anchor-based approaches have been central to many empirical methods employed to aid the interpretation of clinically meaningful change over time in COAs. For over a decade, the US Food and Drug Administration (FDA) has promoted the use of these methods to support the interpretation threshold "for an individual patient PRO score change over a predetermined time period that should be interpreted as a treatment benefit" [4]. More recently, the FDA has confirmed this approach, noting "To aid in the interpretation of study results, FDA is interested in what constitutes a meaningful within patient change (i.e., improvement and deterioration from the patients' perspective) in the concepts assessed by COAs" and "recommends the use of anchor-based methods to establish meaningful within-patient changes." [5]. In the same 2019 document, FDA diminished the role of distribution-based methods, stating:

Distribution-based methods (e.g., effect sizes, certain proportions of the standard deviation and/or standard error of measurement) do not directly take into account the patient voice and as such cannot be the primary evidence for within-patient clinical meaningfulness. Distribution-based methods can provide information about measurement variability. [5]

Anchor-based methods explore the associations between the targeted concept of the COA measure and an external criterion measured by the anchor (or multiple anchors) chosen to be easier to interpret than the COA measure [1, 4, 5]. Simply stated, "the anchor measure(s) are used as external criteria to define patients who have experienced a meaningful change in their condition." [1]. As noted above in the early interpretation studies, the anchor is often a global rating [2, 3]. By identifying the patients who experienced meaningful change based on the anchor measure(s), the meaningful change threshold of the COA measure can be derived [1].

While the anchor(s) are valued for providing plausible interpretation criterion measures to aid interpretation, anchor-based methods contain a number of inherent theoretical and methodological problems that have emerged from the health-related quality of life (HRQoL) literature. The intent of this paper is to provide a critical appraisal of the use of the anchor-based methods by detailing five key concerns for the use of these external criterion in the interpretation of: clinically meaningful change at the individual level or clinically significant differences between groups of patients. Finally, we propose strategies that represents alternatives to both distribution- and anchor-based methods to address the underlying challenges.

Methods

This investigation reviewed key publications that contribute to the appraisal of anchor-based methods employed to derive COA interpretation thresholds. The appraisal detailed a key bias in anchor-based methods that directly influences the magnitude of the interpretation threshold.

Results

Five important concerns for the use of anchor-based methods have emerged from the literature: 1) global estimates of change are consistently biased toward the present state; 2) the use of static current state global measures, while not subject to artifacts of recall, may exacerbate the problem of estimating clinically meaningful change; 3) the anchor assessment response(s) that indicates meaningful change usually involves an arbitrary judgment; 4) the calculated interpretation thresholds are sensitive to the proportion of patients who have improved; and 5) for anchor-based regression methods, the correlation between the COA change scores and the anchor has a direct linear relationship to the magnitude of the interpretation threshold derived using an anchor-based approach, with stronger correlations yielding larger interpretation thresholds. Each of these five concerns is discussed below.

1. Global Estimates of Change Are Consistently Biased Toward the Present State

Anchor-based methods based on a patient global estimate of change (e.g., “Please choose the response that best describes the overall change in your since you started taking the study medication: Much better, A little better, No change, A little worse, Much worse” [6]) have consistently demonstrated bias by overweighting the present state and underweighting the initial state. Empirically, if reports of within-patient change are an unbiased estimator of the difference between study baseline and the present condition, they should result in a high positive correlation with present state and a negative correlation of equal magnitude with the baseline scores [7]. However, empirical investigations correlating the assessment of change on the anchor with independent measures of baseline and present state have consistently demonstrated a high positive correlation with the patients’ current status, and a near zero, and occasionally positive, correlation with baseline assessments [7–10].

The fundamental problem with the approach is that remembering and estimating change from a baseline several weeks or months earlier can be an extremely difficult recall task; as a consequence, people devise alternative, albeit unconscious, strategies [7]. One identified strategy is *implicit theory of change* [11]. Using numerous examples from the social science literature, Ross [12] documented how individuals do not directly recall the initial state; instead, they use implicit theories based on their current state to estimate their initial state and then reconstruct the estimate of change over time. As a result, implicit theories of patient-perceived stability and/or change lead to recall bias and overweight current status in the change estimation. Ross’s work provides a framework to understand the empirical evidence [7–10] that retrospective ratings are a reflection of the patient’s perception of the current status rather than an accurate assessment of change over time.

2. The Use of Static Current State Global Measures May Exacerbate the Problem of Estimating Clinically Meaningful Change

The challenge to appropriately identify the patients who have a meaningful change is perhaps more difficult when a static (current state) patient global impression of severity (PGIS) scale is used (e.g., “Please choose the response below that best describes the severity of your illness over the past week:

None, Mild, Moderate, Severe) as recommended by the FDA [6]. This approach does not directly elicit information from patients about the magnitude of meaningful change.

Moreover, these patient assessments of present state may also suffer from a bias analogous to the implicit theory of change or stability. The related bias is called *response shift* – as a patient's health state changes, their expectation of ideal health may change with it. Patients with chronic or degenerative diseases may acclimatize to their health state, so report good or excellent health despite obvious infirmities. As a consequence, "HRQoL scores can be stable despite changes in HRQoL" [14]. That is, the static PGIS response given at baseline may not reflect the health framework used by the patient at later PGIS assessments.

3. The Anchor Assessment Response(s) That Indicates Meaningful Change Usually Involves an Arbitrary Judgment

Using the patient global estimate of change (e.g., "Please choose the response that best describes the overall change in your since you started taking the study medication: Much better, A little better, No change, A little worse, Much worse" [6]) to understand and interpret meaningful change requires the selection of a specific global response(s) to anchor the change score analyses. What then constitutes a meaningful change? Does a patient need to be *Much better* or perhaps simply *A little better* for the COA change/improvement to be meaningful? What if the disease or condition is known for rapid patient deterioration on the COA's concept of interest, should a patient change response of *No change* indicate a meaningful change or improvement over time given the historically-known downward disease trajectory? Moreover, the selected meaningful change level(s) need to identify the patients who have changed, while at the same time, have not changed too much. That is, it is important to identify the subset of patient who have experienced a meaningful change, but at the same time, to not overestimate the important change threshold through the inclusion of patients with large changes in the meaningful change estimation process [13].

The situation is exacerbated when static state measures are used because relevant change levels are determined by computing the difference between the two states, yet meaningful change is not directly estimated by patients. Recognizing this, the FDA asks sponsors to specify and justify "the anchor response category that represents a clinically meaningful change to patients on the anchor scale, e.g., a 2-category decrease on a 5-category patient global impression of severity scale." [1]. However, when the criterion judgment of meaningful change over time on this static scale is left in the hands of the investigating team or an expert panel, this undermines the process of identifying meaningful patient-informed change using the static anchor.

It is suggested [1, 5] that interpretation of meaningful change may be assisted by graphic display of the empirical cumulative density function (eCDF) at each global change value. While the eCDF curves from each change level provides descriptive information on the relationship of the COA to the anchor's change score, these graphic displays do not directly inform the anchor's meaningful change threshold. That is, by noting that "The meaningful within-patient threshold of the target COA should be explored by the eCDF of

the anchor category where the patients are defined and judged (by the anchor measure) as having experienced meaningful change in their condition" [1] assumes that meaningful change level for the anchor is known or has been established.

Indeed, without an adequate qualitative investigation [15] of the global item's response options to understand what patients within the target population consider a meaningful change in how they feel or function on the global item's scale, the selected relevant level(s) for the global assessment response(s) that indicates meaningful change may involve an arbitrary judgment by the investigating team, and that judgment can differ over time [3, 16]. The use of a static anchor (e.g., PGIS) and eCDF displays does not address the crucial issue of *how* the anchor's meaningful change level is established.

Finally, the reliability of anchor ratings is generally unknown, with limited evidence supporting test–retest reliability of anchor item(s). The paucity of evidence of reliability for the anchor assessments was first noted in 1997 by Norman, Stratford and Regehr [7], and as described by Lavigne in 2016, if the anchor item(s) used to assess the meaningful change is not reliable, the resulting change threshold for meaningful improvement or decline may not be reliable [17].

4. The Calculated Interpretation Threshold Is Sensitive to the Proportion of Patients Who Have Improved

Terluin et al [18] examined the impact of the proportion of improved patients on minimally important change (MIC) thresholds in: 1) multiple simulations of patient samples from anchor-based MIC studies, and 2) in a clinical study dataset. A group MIC was compared to the average of all individual patient reported MIC levels if the patient reported an important change/improvement using a global change anchor, and the group MIC was calculated using two methods, receiver-operator characteristic (ROC) curves and predictive modeling [18]. Not surprisingly, when the proportion of improved patients was less than 50%, the group MIC underestimated the average of individual MICs because proportionately more observations came from the unchanged group. Conversely when more than half the patients had an important change/improvement, the group MIC overestimated the average of individual MICs for the same reason [18].

The FDA has discouraged the use of ROC analysis as the primary method for understanding the meaningful within-patient change threshold for similar reasons, making note that this method is "partially a distribution-based approach" and "the most sensitive threshold identified by ROC [analysis] may not actually be the most clinically meaningful threshold to patients" [5].

5. The Strength of the Relationship Between Changes in the Anchor and Changes in the COA has a Direct Impact on the Magnitude of the Meaningful Change Threshold

An important and often overlooked source of bias in anchor-based methods that directly influences the magnitude of the meaningful change threshold is the correlation between change assessed by the anchor and the COA change scores [13]. It is self evident that there should be some relationship between change in the COA and change assessed by the anchor scale [1, 5]. This can be visualized by considering extreme

cases. If there is a no relationship between change on the COA and change on the anchor (r_{xy} = 0.0), then no amount of change in the anchor will lead to a non-zero predicted change in the COA. The two measures are independent. Conversely, if there is a perfect linear relationship (r_{xy} = 1.0), then any change in the anchor will lead to an equivalent 1 SD change in the COA. And presumably intermediate change correlations/relationships between the two measures must result in correspondingly intermediate values on the meaningful change on the COA.

Without providing specific thresholds, FDA notes that an anchor should be "sufficiently correlated to the targeted COA." [1, 5]. Hays, Farivar, & Liu reported in 2005 that a correlation coefficient of $r \geq 0.371$ (equivalent to an effect size of 0.80) defines "a noteworthy (large effect) association" between change on the anchor and change on target COA measure [19]. Other authors have recommended a range of 0.30–0.70 for the magnitude of this change scores correlation [20, 21]. Leaving aside the broad nature of these recommendations, what is not recognized is that selecting a correlation range does not nullify the impact of the correlation on the calculation of the meaningful change threshold. Quite the opposite; the magnitude of the association is a direct determinant of the magnitude of this threshold when regression analysis is used to estimate change in the COA that results from meaningful change on the anchor [13]. Expressed in standard deviation (SD) units, the meaningful change threshold is the difference on the anchor scale in SD units corresponding to meaningful change multiplied by the correlation coefficient.

This key issue makes clear that the magnitude of the resulting meaningful change threshold that emerges from the regression analysis will *increase* with the strength of this relationship. [13]. Moreover, the effect is non-trivial – depending on the strength of the correlation, the meaningful change threshold can vary from 0 to 1 SD, and is minimal when the correlation is smallest – the weakest relationship. Setting arbitrary ranges of correlation such as 0.30 to 0.70 reduces the impact of correlation, but it remains a major determinant.

While the magnitude of this bias can be directly estimated, it is unclear how to address the role of this key determinant. Fayers and Hays recommended a strategy called *linking* that equates the standardized change in the COA and the anchor; this strategy is equivalent to assuming a perfect linear relationship (r_{xy} = 1.0) between the change scores [13]. This strategy is, of course, problematic in justifying the use of the anchor if there is a notably weak association between the two change scores. Moreover, the effect of this strategy is not trivial. Using the results from Suner et al [22], Fayers and Hays' report the estimated minimal important difference (MID) on the 25-item National Eye Institute Visual Function Questionnaire (NEI VFQ-25; scored 0 (worst) to 100 (best vision-related function)) using at least a 15-word change in visual acuity as the anchor was 4.3 points using linear regression models ($r < 0.3$) and 21.8 via the linking approach—two widely different thresholds for interpreting change over time for patients with neovascular age-related macular degeneration [13, 22].

This finding is worrisome in that a stronger association between change on the COA and anchor will yield higher values for meaningful change threshold, while a weaker anchor relationship yields a lower threshold for demonstrating a meaningful change using regression analysis.

The importance of this finding may give cause for the reconsideration of all meaningful change thresholds computed to date using an anchor-based approach and regression analysis.

Discussion

Anchors have assumed a central role in the interpretation of health status changes related to therapeutic interventions. This investigation explored five known challenges to an external anchor-based approach to interpretation of changes in the COA of interest, and interpretation of the evidence from anchor-based methods is vulnerable to these biases.

At the heart of this discussion is the utility and interpretation of the external anchor measure. When detailing the considerations for selecting anchors, the FDA states “anchors should be plainly understood in context, easier to interpret than the COA itself, and sufficiently correlated to the targeted COA.” [1]. Naturally, the patient’s global change assessment is often used as such an anchor, allowing patients to directly provide “the standard by which to measure the benefits and harms of their treatments.” [18]. However, global change assessments consistently suffer from recall bias; these and other anchor measures often are subject to arbitrary level-setting for meaningful change, limited and often unknown test-retest reliability, and vulnerability to biases from implicit theory of change and response shift. Empirically, an understanding that the change score correlation between the COA and the anchor is directly associated with the magnitude of the meaningful change threshold derived using regression analysis questions the role of these assessments in the interpretation process.

There is an underlying paradox. Extensive resources are invested in research related to the development of content and psychometrically valid COAs. Yet the interpretation of change over time, and the critical act of determining the magnitude of change on these often multi-item measures that will be regarded as clinically meaningful is accomplished by comparing the domain/concept change scores to a plain and easy anchor measure(s) with often unknown and irreducible biases. In 2016, Coon and Cappelleri described this dilemma, noting:

However, even if a related, interpretable, serial anchor is used for determining how to interpret PRO scores, the development of the anchor is unsettlingly more simple than that of the PRO. In other words, how can we hinge the interpretation of our rigorously developed PRO on an overly simple anchor? [21]

A New Approach: Defining the Meaningful Change Threshold Within the COA

In our view, the external anchor-based method, while providing an early and useful approach to interpretation, may have reached the limit of its development. At a conceptual level, it is simply illogical to interpret a carefully developed COA using a “gold standard” that consists of a single and/or simpler item(s), and is known to be vulnerable to identified biases.

One solution is to refocus efforts to development of methods that derive thresholds of change *within* the COA itself. One very promising development in this regard initiated by the Patient Reported Outcomes

Measurement Information System® (PROMIS® [23]) researchers is the *PRO-Bookmarking* [24] approach, an adaptation of the Bookmark procedure for pass/fail standard-setting used by US state academic achievement assessment systems [25]. Using Item Response Theory (IRT) methods, the PROMIS® was developed calibrated PROs to assess physical, emotional and social health of patients in clinical care, observational studies and clinical trials [23]; in addition, PROMIS® researchers developed PRO-Bookmarking to identifying key change thresholds in PROs using the items *within* the PROMIS® measure [24].

In its original form, a key feature of the Bookmark procedure is the Ordered Item Booklet (OIB), which contains all academic test items (one per page) ordered by empirically-determined difficulty (easiest to hardest). Using the OIB, a panel of subject matter experts (SMEs) determines where to place a bookmark between two items such that the "minimally qualified" student is expected to have mastered the items below the bookmark, with multiple rounds of training and discussion between SME panelists [25, 26].

In the most recently published PROMIS® bookmarking report investigating meaningful change in the concepts of rheumatoid arthritis (RA) pain interference and fatigue, the original Bookmark procedure was modified in a number of ways [27]. The OIB was replaced with a series of patient vignettes (natural language short stories) that describe 4 or 5 key symptoms from PROMIS item banks of the two investigated concepts. Each vignette had a PROMIS® IRT-based score, and these vignettes were presented to the panels in order from lowest to highest scores. The SMEs panels including two key informants, clinician and patients; each panel separately reviewed ordered vignettes to identify (bookmark) transition points from none-mild, mild-moderate and moderate-severe across the severity spectrum for these two concepts. Using vignettes at the severe or mild level as a baseline, each panel was then asked to identify the vignettes that represented a meaningful improvement or worsening, respectively, for each investigated concept, with the PROMIS® change scores then used to identify the magnitude of the meaningful change thresholds reported for each panel [27].

The approach has a critical difference from previous distribution- and external anchor-based methods in that calibration arises *within* the COA, thereby obviating many of the concerns raised earlier about anchor-based methods. Second, it clearly articulates the patient's voice, unlike distribution-based methods, and provides a clear comparison to the clinician perspective, which did not always align with the patient perspective. A third difference is that, while it originated in the educational context of standard setting for academic tests, the PROMIS® bookmarking adaptation for PRO assessments provides a relevant and rich opportunity to learn from key informants how meaningfulness is discerned.

However, the Bookmark procedure is only one strategy used by educators for calibrating criterion referenced cut scores using known information from within the test. Others – including Angoff [28], Hofstee [29], Nedelsky [30] - may be adaptable to COA assessments and deserve investigation [31, 32, 33]. With heightened awareness and a deeper understanding of the biases inherit to external anchor-based methods used to derive meaningful change thresholds to understand both individual and group

change over time, we encourage researchers to further pursue new methods that use the information within a COA based on their usefulness in the education standard-setting arena.

Declarations

No funding was received to assist with the preparation of this manuscript, and the authors have no conflicts of interest to declare that are relevant to the content of this article. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

1. Patient-Focused Drug Development (2018). *Guidance 3 Discussion Document: Select, Develop or Modify Fit-for-Purpose Clinical Outcomes Assessments*. Retrieved March 26, 2021, from https://www.fda.gov/media/116277/download.
2. Deyo, R. A., & Inui, T. S. (1984). Toward clinical applications of health status measures: sensitivity of scales to clinically important changes. *Health Services Research, 19*(3), 275–289.
3. Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled Clinical Trials, 10*(4), 407–415. https://doi.org/10.1016/0197-2456(89)90005-6.
4. U.S. FDA. (2009). *U.S Department of Health and Human Services Food and Drug Administration Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims*. Retrieved March 26, 2021, from https://www.fda.gov/media/77832/download.
5. Patient-Focused Drug Development (2019). *Guidance 4 Discussion Document: Incorporating Clinical Outcome Assessments into Endpoints for Regulatory Decision Making*. Retrieved March 26, 2021, from https://www.fda.gov/media/132505/download.
6. Patient-Focused Drug Development (2018). *Guidance 3 Discussion Document: Select, Develop or Modify Fit-for-Purpose Clinical Outcomes Assessments. Appendices.* Retrieved March 26, 2021, from https://www.fda.gov/media/116281/download.
7. 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. https://doi.org/10.1016/s0895-4356(97)00097-8.
8. Guyatt, G. H., Norman, G. R., Juniper, E. F., & Griffith, L. E. (2002). A critical look at transition ratings. *Journal of Clinical Epidemiology, 55*(9), 900–908. https://doi.org/10.1016/s0895-4356(02)00435-3.
9. Schmitt, J., & Di Fabio, R. P. (2005). The validity of prospective and retrospective global change criterion measures. *Archives of Physical Medicine and Rehabilitation, 86(12):2270–2276.* https://doi.org/10.1016/j.apmr.2005.07.290.

10. Metz, S. M., Wyrwich, K. W., Babu, A. N., Kroenke, K., Tierney, W. M., & Wolinsky, F. D. (2007). Validity of patient-reported health-related quality of life global ratings of change using structural equation modeling. Quality of Life Research, 16(7):1193–1202. https://doi.org/ 10.1007/s11136-007-9225-1.
11. Ward, C. L., & Wilson, A. E. (2015). Implicit Theories of Change and Stability Moderate Effects of Subjective Distance on the Remembered Self. *Personality and Social Psychology Bulletin., 41*(9), 1167–1179. https://doi.org/10.1177/0146167215591571.
12. Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review, 96*(2), 341–357. https://doi.org/10.1037/0033-295X.96.2.341.
13. 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. https://doi.org/10.1007/s11136-013-0443-4.
14. Schwartz, C. E., Andresen, E. M., Nosek, M. A., & Krahn, G. L., RRTC Expert Panel on Health Status Measurement (2007). Response shift theory: important implications for measuring quality of life in people with disability. *Archives of Physical Medicine and Rehabilitation, 88*(4), 529–536. https://doi.org/10.1016/j.apmr.2006.12.032.
15. Turner-Bowker, D. M., Lamoureux, R. E., Stokes, J., Litcher-Kelly, L., Galipeau, N., Yaworsky, A., Solomon, J., & Shields, A. L. (2018). Informing a priori Sample Size Estimation in Qualitative Concept Elicitation Interview Studies for Clinical Outcome Assessment Instrument Development. *Value in Health, 21*(7), 839–842. https://doi.org/10.1016/j.jval.2017.11.014.
16. Juniper, E. F., Guyatt, G. H., Willan, A., & Griffith, L. E. (1994). Determining a minimal important change in a disease-specific quality of life questionnaire. *Journal of Clinical Epidemiology, 47*(1), 81–87. https://doi.org/10.1016/0895-4356(94)90036-1.
17. Lavigne, J. V. (2016). Systematic Review: Issues in Measuring Clinically Meaningful Change in Self-Reported Chronic Pediatric Pain Intensity. *Journal of Pediatric Psychology, 41*(7), 715–734. https://doi.org/10.1093/jpepsy/jsv161.
18. Terluin, B., Eekhout, I., & Terwee, C. B. (2017). The anchor-based minimal important change, based on receiver operating characteristic analysis or predictive modeling, may need to be adjusted for the proportion of improved patients. *Journal of Clinical Epidemiology, 83*, 90–100. https://doi.org/10.1016/j.jclinepi.2016.12.015.
19. Hays, R. D., Farivar, S. S., & Liu, H. (2005). Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *Chronic Obstructive Pulmonary Diseases*, *2*(1), 63 – 7. https://doi.org/10.1081/copd-200050663. PMID: 17136964.
20. 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. https://doi.org/10.1016/j.jclinepi.2007.03.012.
21. Coon, C. D., & Cappelleri, J. C. (2016). Interpreting Change in Scores on Patient-Reported Outcome Instruments. *Therapeutic Innovation & Regulatory Science, 50*(1), 22–29. https://doi.org/10.1177/2168479015622667.

22. Suñer, I. J., Kokame, G. T., Yu, E., Ward, J., Dolan, C., & Bressler, N. M. (2009). Responsiveness of NEI VFQ-25 to changes in visual acuity in neovascular AMD: validation studies from two phase 3 clinical trials. *Investigative Ophthalmology & Visual Science, 50*(8), 3629–3635. https://doi.org/10.1167/iovs.08-3225.
23. Khanna, D., Krishnan, E., Dewitt, E. M., Khanna, P. P., Spiegel, B., & Hays, R. D. (2011). The future of measuring patient-reported outcomes in rheumatology: Patient-Reported Outcomes Measurement Information System (PROMIS). *Arthritis Care Research, 63*, S486–S490. https://doi.org/10.1002/acr.20581.
24. Cook, K. F., Cella, D., & Reeve, B. B. (2019). PRO-Bookmarking to estimate clinical thresholds for patient-reported symptoms and function. *Medical Care, 57*(Suppl 1), S13–S17. https://doi.org/10.1097/mlr.0000000000001087.
25. Lewis, D. M., Mitzel, H. C., Green, D. R., & Patz, R. J. (1999). *The Bookmark standard setting procedure*. McGraw Hill.
26. Karantonis, A., & Sireci, S. G. (2006). The Bookmark Standard-Setting Method: A Literature Review. *Educational Measurement Issues and Practice, 25*(1), 4–12. https://doi.org/10.1111/j.1745-3992.2006.00047.x.
27. Bingham, C. O., Butanis, A. L., Orbai, A. M., Jones, M., Ruffing, V., Lyddiatt, A., Schrandt, M. S., Bykerk, V. P., Cook, K. F., & Bartlett, S. J. (2021). Patients and clinicians define symptom levels and meaningful change for PROMIS pain interference and fatigue in RA using bookmarking. Rheumatology, 20. https://doi.org/10.1093/rheumatology/keab014.
28. Ricker, K. L. (2006). Setting cut-scores: A critical review of the Angoff and modified Angoff methods. *Alberta journal of educational research, 52*(1), 53–64.
29. Schindler, N., Corcoran, J., & DaRosa, D. (2007). Description and impact of using a standard-setting method for determining pass/fail scores in a surgery clerkship. *The American journal of surgery., 193*(2), 252–257. https://doi.org/10.1016/j.amjsurg.2006.07.017.
30. Melican, G. J., Mills, C. N., & Plake, B. S. (1989). Accuracy of item performance predictions based on the Nedelsky standard setting method. *Educational and psychological measurement., 49*(2), 467–478.
31. Norcini, J. J. (2003). Setting standards on educational tests. *Medical Education, 37*(5), 464–469. https://doi.org/10.1046/j.1365-2923.2003.01495.x.
32. Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education., 2*(2), 121–141. https://doi.org/10.1207/s15324818ame0202_3.
33. Park, J., Ahn, D. S., Yim, M. K., & Lee, J. (2018). Comparison of standard-setting methods for the Korean radiological technologist licensing examination: Angoff, Ebel, bookmark, and Hofstee. *Journal of educational evaluation for health professions, 15*, 32. https://doi.org/10.3352/jeehp.2018.15.32.