Title

Likely change indexes improve estimates of individual change on patient-reported outcomes

Journal

Quality of Life Research, 32(5)

ISSN

0962-9343

Authors

Peipert, John Devin
Hays, Ron D
Cella, David

Publication Date

2023-05-01

DOI

10.1007/s11136-022-03200-4

Peer reviewed

Quality of Life Research
https://doi.org/10.1007/s11136-022-03200-4

SPECIAL SECTION: METHODOLOGIES FOR MEANINGFUL CHANGE

Likely change indexes improve estimates of individual change on patient-reported outcomes

John Devin Peipert[1] · Ron D. Hays[2] · David Cella[1]

Accepted: 7 July 2022

Abstract

Purpose Individual change on a patient-reported outcome (PRO) measure can be assessed by statistical significance and meaningfulness to patients. We explored the relationship between these two criteria by varying the confidence levels of the coefficient of repeatability (CR) on the Patient-Reported Outcomes Measurement Information System (R) Physical Function (PF) 10a (PF10a) measure.

Methods In a sample of 1129 adult cancer patients, we estimated individual-change thresholds on the PF10a from baseline to 6 weeks later with the CR at 50%, 68%, and 95% confidence. We also assessed agreement with group- and individual-level thresholds from anchor-based methods [mean change and receiver operating characteristic (ROC) curve] using a PF-specific patient global impression of change (PGIC).

Results CRs at 50%, 68%, and 95% confidence were 3, 4, and 7 raw score points, respectively. The ROC- and mean-change-based thresholds for deterioration were −4 and −6; for improvement they were both 2. Kappas for agreement between anchor-based thresholds and CRs for deterioration ranged between $\kappa = 0.65$ and 1.00, while for improvement, they ranged between 0.35 and 0.83. Agreement between the PGIC and all CRs always fell below "good" ($\kappa < 0.40$) for deterioration (0.30–0.33) and were lower for improvement (0.16–0.28).

Conclusions In comparison to the CR at 95% confidence, CRs at 50% and 68% confidence (considered likely change indexes) have the advantage of maximizing the proportion of patients appropriately classified as changed according to statistical significance and meaningfulness.

Keywords Individual change · Patient-reported outcomes · Meaningful change · Cancer

Introduction

Estimating change on patient-reported outcomes (PROs) for individual patients can be informative in clinical trials and clinical monitoring of individual patients. To be confident that change is real, it needs to be differentiable from error. In addition, change should be meaningful to patients. Psychometric tradition has focused on differentiating true change from error using 90% or 95% confidence intervals [1–3]. It is also important to know whether a patient feels they have changed or not. For this reason, the current regulatory guidance from the United States Food and Drug Administration (US FDA) emphasizes meaningful within-patient change to define clinical benefit on a PRO [4].

There are notable challenges around selecting thresholds for whether individual patients, instead of groups of patients, have changed on a PRO measure. The amount of improvement or deterioration considered meaningful varies from one person to the next [5–7]. While this inter-individual variation can be captured through single-item approaches like retrospective change ratings [e.g., patient global impression of change (PGIC)] [8], this cannot be done directly for multi-item scales. While applying group-level meaningful change estimates to classify individuals as changed has significant problems [9; 10] due to error of individual-change estimates, especially in raw change scores [11], given the lack of actual individual estimates of meaningful change, we

✉ John Devin Peipert
john.peipert@northwestern.edu

[1] Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, 625 Michigan Ave, 21st Floor, Chicago, IL 60611, USA

[2] Division of General Internal Medicine and Health Services Research, University of California Los Angeles, Department of Medicine, Los Angeles, CA, USA

Published online: 03 August 2022

might consider how group-level information can enrich or support our understanding of individual change [12].

To protect against measurement error associated with individual-change estimates, we may require that change is statistically significant. For example, the reliable change index (RCI)[13], or its mathematical transformation to the coefficient of repeatability (CR), provides a way of estimating statistically significant within-patient change based on a pre-specified significance or confidence level and the standard error of measurement (SEM), or the expected measurement error around an individual's score. Using 95% confidence (p-value of < 0.05), the CR tends to generate large estimates of the amount of change needed to be statistically significant. Such estimates are likely to exceed change that has meaning to the individual [14]. Some have suggested relaxing the confidence level required to denote individual change, recognizing that 95% certainty may be overly restrictive and may lead us to misclassify a sizable number of patients who feel they have changed as unchanged [14, 15]. That is, for some applications, it may be acceptable to be less certain that a change threshold is differentiable from measurement error if doing so classifies as changed more patients who feel they have changed. In addition, a better understanding of the trade-offs between decreased statistical confidence and increased likelihood of meaningfulness is needed. In this paper, we examine how CRs calculated at varying levels of confidence classified individual patients as having changed or not changed relate to the patient's perception of meaningful change. Our goal is to explore a *likely change index (LCI)* with a confidence level $< 95\%$ that improves upon use of either group-level meaningful change estimates or standard reliable change analyses by considering information across these approaches. We focused on the cancer setting, examining change in terms of deterioration and improvement over time.

Methods

Data set and participants

We used data from an observational, longitudinal study with cancer patients from five comprehensive cancer centers across the United States: Mayo Clinic (MN), Northwestern University (IL), University of North Carolina (NC), Memorial Sloan-Kettering Cancer Center (NY), and MD Anderson Cancer Center (TX) (PI: Sloan; R01-CA154537) [16]. IRB approval was obtained at each site. Patients were eligible if they were ≥ 18 years of age, had a cancer diagnosis and initiated treatment in the following 7 days or underwent cancer-related surgery within the past 14 days, had an Eastern Cooperative Oncology Group (ECOG) performance status rating of 0–4, and were able to provide informed consent and participate in the study. The study collected data on the Patient-Reported Outcomes Measurement Information System (R) (PROMIS®) PF10a, at enrollment (baseline) and 6 weeks later. A follow-up of 6 weeks was selected to provide sufficient time for change in patients' physical function from baseline. The dataset contained 1829 patients. The analytic sample of 1129 consisted of patients with data on the PGIC anchor and all PROMIS PF10a items at baseline and 6 weeks.

Measures

PROMIS PF10a

The PF10a was created as part of an effort to generate cancer-targeted PROMIS measures. PROMIS items relevant to cancer were identified with focus groups, content experts in cancer, and field-testing with cancer patients. This short form has been found to be highly correlated with a legacy physical function measure, with fewer ceiling effects, and is also correlated with global HRQOL, and has demonstrated known-groups validity and responsiveness to changes in clinical measures of health [17, 18]. To isolate the effect of different individual-change thresholds of interest, we examined simple summated PF10a scores from patients with non-missing data. Each PF10a item has response options numbered from 1 to 5, with higher responses indicated better physical function. Summing these items resulted in a possible score range of 10–50. We elected to use raw summed scores instead of the PROMIS item-response theory (IRT)-estimated T scores because IRT scores approach error differently than summed scores, and we sought to illustrate the comparison of statistical significance and meaningfulness of change with a simple example.

Other measures

A PGIC item was included at the 6-week assessment, asking patients how much their physical function changed since the baseline assessment, with responses options of "A lot better" (1), "A little better" (2), "About the same" (3), "A little worse" (4), and "A lot worse" (5). To facilitate analyses, we dichotomized these categories to derive two additional variables, one representing deterioration [1 = deteriorated ("A lot worse"/"A little worse"), 0 = stayed the same/improved ("About the same"/"A lot better"/"A little better")] and one representing improvement [1 = improved ("A lot better"/"A little better"), 0 = stayed the same/deteriorated ("About the same"/"A lot worse"/"A little worse")]. At baseline, patient performance status was assessed by self-report, as follows: "Please indicate which statement best describes your CURRENT activity level?": 0 (normal activities, without symptoms); 1 (some symptoms, but do not require bed rest during

waking day); 2 (require bed rest for less than 50% of waking day); 3 (require bed rest for more than 50% of waking day); 4 (unable to get out of bed) [19].

Statistical methods

The RCI formula used for this study was $(X_2 - X_1)/\sqrt{2}SEM$, where X_1 and X_2 are the individual patient's value of the PRO at baseline and 6 weeks, the SEM (standard error of measurement) is defined as $SD_1\sqrt{1 - reliability}$ (SD_1 = standard deviation of the PF10a and the reliability is Cronbach's coefficient alpha at baseline). We calculated the RCI at three levels of statistical significance: 95%, 68%, and 50%. We selected 68% since it represents observations within one standard deviation from the mean on a standard normal distribution and has been suggested as a potential threshold value to indicate likely change [15]. The RCI categorizes the patients as significantly changed (deteriorated or improved), if its value exceeds the critical value of the standard normal distribution: 95% = 1.96; 68% = 0.994; 50% = 0.674. To calculate the score threshold on the PF10a, we used the CR, a transformation of the RCI calculated as $criticalvalue * \sqrt{2}SEM$. The CR is also known as the smallest detectable change, minimally detectable change, or smallest real difference. We refer to the CR when discussing uses of this transformation. We also refer to CRs at 68% and 50% confidence as LCIs. Despite using between-subjects information about measurement error (i.e., the sample SD within the SEM), such statistics still yield a valid individual-level interpretation. The numerator of these statistics contains the raw PRO change score for specific individual patients and puts the measurement error inherent in that raw change score into context by dividing by an estimate of measurement error. Given the high intensity of data collection needed to make a purely intra-individual statistic possible, the RCI and LCIs are useful in their practicality, requiring only two measurement occasions [11, 20].

We took two approaches to estimating thresholds for meaningful change at the group level. First, we used receiver operating characteristic (ROC) curve analysis. The ROC approach was conducted by first fitting a logistic model regressing the dichotomized PGIC variables (deteriorated or improved) on the PF10a change score to estimate the area under the curve (AUC) and predicted probability of having changed for each observed change score. We then used the predicted probabilities, number of true and false positives and negatives, sensitivity, and specificity output from logistic model run as input to the %rocplot macro in SAS to generate ROC plots [21]. The threshold for important change from this analysis is defined as the change score suggested by Youden's index, or the sum of sensitivity and specificity – 1 [22]. The mean-change approach was conducted by estimating the mean PF10a change score for patients classified as deteriorated and stayed the same on the PGIC anchor separately. We used a linear model with least squares (LS) means to estimate change scores for this procedure. The threshold for meaningful change from this analysis was defined as the change score estimated for the deteriorator group or improver group, as appropriate. We conducted both the ROC and mean-change analyses for both deterioration (using version of PGIC 1 = deteriorated and 0 = stayed the same/improved) and improvement (using version of PGIC 1 = improved and 0 = stayed the same/ deteriorated). Before selecting final thresholds, we adjusted the threshold suggested by Youden's index for unequal proportions of patients changed vs. not changed [6]. For the ROC and mean-change analyses, thresholds were set by rounding up to the nearest whole number from the relevant estimate so that the assigned threshold exceeds the meaningful change estimate. We further note that these thresholds do not represent the minimally important difference or change because they include patients who have changed a lot as well as a little in the anchor.

We then estimated the agreement between thresholds suggested by CR at 95%, 68%, and 50% significance level critical values with the group-level, anchor-based thresholds suggested by the ROC and mean-change approaches, which constituted a comparison of individual-change thresholds (CR) to group-change thresholds (anchors). We also examined the agreement between thresholds suggested by the CRs with whether patients were categorized as deteriorated or improved using the physical function PGIC anchor, as appropriate. This comparison is advantageous because it examines the agreement between two individual-level variables. For each agreement comparison, we calculated kappa statistics with 95% confidence intervals [23]. The following standards were used to guide interpretation of the kappa: < 0.40 = poor; 0.40- < 0.75 = good; ≥ 0.75 = excellent [24]. Next, we calculated the sensitivity, specificity, positive predictive value, and negative predictive value of correspondence between the methods of classifying patients as having deteriorated or improved, as appropriate.

Results

Patient characteristics are detailed in Table 1. The majority or highest proportions of patients were female (63%), aged 46–65 years (55%), had a diagnosis of breast cancer (27%) or lymphoma/myeloma (22%), were at Stage 4 in their cancer (35%), and had a ECOG performance status rating of 1 (74%).

The means of the PF10a raw scores at baseline and 6 weeks were 40.52 and 39.38, respectively. (Table 1.) The distribution of PROMIS PF10a change scores from baseline to 6 weeks is shown in the histogram in Fig. 1. The

Table 1 Patient characteristics ($N = 1129$)

Female sex, % (*n*)	63% (716)
Age, % (*n*)	
18–45 years	17% (192)
46–65 years	55% (613)
≥65 years	28% (307)
Cancer Site, % (*n*)	
Breast	27% (294)
Cervix	0% (0)
Colorectal	10% (107)
Lung	7% (78)
Lymphoma/Myeloma	22% (244)
Prostate/Bladder	1% (14)
Uterine	0% (0)
Head/Neck/Gastroesophageal	8% (86)
Other	25% (276)
Cancer Stage, % (*n*)	
Stage 1	13% (135)
Stage 2	23% (243)
Stage 3	30% (329)
Stage 4	35% (372)
ECOG Performance Status Rating, % (*n*)	
0	0% (0)
1	74% (824)
2	26% (294)
PROMIS Physical Function 10a, mean (SD, min, max)	
Baseline	40.52 (7.24, 14.00–50.00)
6 weeks	39.38 (7.7, 12. 00–50.00)
Change	−1.24 (6.19, −30,00, 21.00)
Patient Global Impression of Change at 6 weeks, % (*n*)	
A lot better	12% (135)
A little better	19% (219)
About the same	43% (490)
A little worse	21% (233)
A lot worse	5% (52)

mean change score was −1.24 (SD = 6.20), with a minimum of −30.00, median of −1.00, and maximum of 21.00. Of the total analysis sample, 25% ($n = 285$) had deteriorated, reporting being either "A little worse" ($n = 233$) or "A lot worse" ($n = 52$); 31% ($n = 354$) had improved, reporting being either "A little better" ($n = 219$) or "A lot better" ($n = 135$); and 43% ($n = 490$) reported being "About the same." (Table 1.) The Spearman correlation between the PGIC (uncollapsed) and change on the PF10a was −0.42. The point biserial correlation (equivalent to Pearson's correlation) between change on the PF10a and dichotomized deterioration (deteriorated vs. stayed the same/improved) PGIC was −0.36, and the point biserial correlation between change on the PF10a and improvement (improved vs. stayed the same/deteriorated) PGIC was 0.31.

Coefficient alpha for the PROMIS PF10a at baseline was 0.90. The SEM at baseline was 2.29. The CRs were as follows: 95% confidence = 6.35, implying a threshold of 7 points; 68% confidence = 3.22, implying a threshold of 4 points; 50% confidence = 2.18, implying a threshold of 3 points. We note the similarity of the SEM and the CR at 50% confidence in this case (difference = 0.11). Considering the broader relationship between the SEM and CR, the Supplementary Materials feature a table demonstrating that absolute differences between the SEM and CR are smallest for the CR at 50% confidence and largest for the CR at 95% confidence. In addition, within each confidence level, the difference increases as the SEM increases.

Regarding deterioration, on the anchor-based analysis of important group change, the adjusted ROC analysis suggested a threshold of −3 points using Youden's index

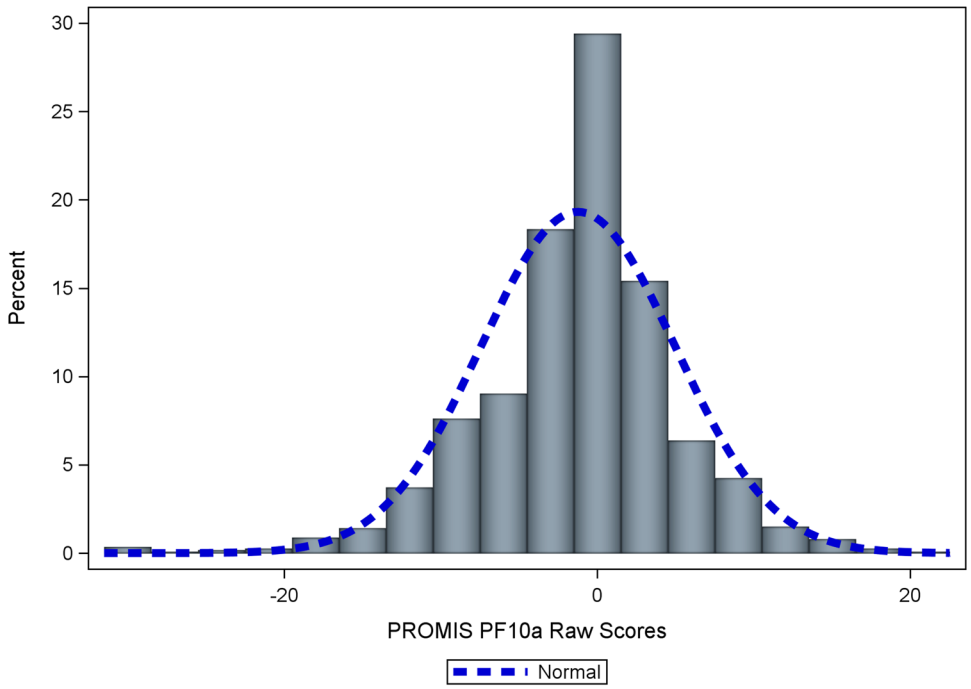

Fig. 1 Distribution of PROMIS PF10a raw change scores

(area under the curve = 0.73, sensitivity = 0.61, specificity = 0.75). (Fig. 2a.) After adjustment for unequal proportions of changed vs. unchanged patients, this threshold was −4. Using the mean-change approach, patients categorized as stayed the same or improved on the PGIC anchor had a mean PF10a change score of 0.05, and those categorized as deteriorated had a mean PF10a change score of −5.04, implying a threshold of −6 points. (Fig. 3a.)

Regarding improvement, on the anchor-based analysis of important group change, the ROC analysis suggested a threshold of 1 point using Youden's index (area under the curve = 0.71, sensitivity = 0.57, specificity = 0.73). (Fig. 2b.) After adjustment for unequal proportions of changed vs. unchanged patients, this threshold was 2. Using the mean-change approach, patients categorized as stayed the same or deteriorated on the PGIC anchor had a mean PF10a change score of −2.54, and those categorized as improved had a mean PF10a change score of 1.62, implying a threshold of 2 points. (Fig. 3b.)

Results of tests of agreement between categorizing patients as deteriorated on the PROMIS PF10a using change thresholds generated from the CRs and categorizing patients as deteriorated and improved using anchor-suggested thresholds on the PROMIS PF10a are shown in Tables 2 and 3. The kappa for agreement between the ROC-generated thresholds and CRs for deterioration indicated good agreement for CR at 95% confidence (κ = 0.69), perfect agreement for the CR at 68% due to the estimate being the same (κ = 1.00), and excellent agreement for the CR at 50% (κ = 0.85). The sensitivity and negative predictive value (NPV) were 1.00 for CRs at 95% and 68% confidence (no false negatives, i.e., categorized as changed on CR but not changed on the ROC-based threshold), and these values were slightly lower for the CR at 50% due to a small number of false negatives. The specificity and positive predictive value (PPV) for the CR at 95% were relatively high at 0.87 and 0.60, respectively, due to some expected false positives (i.e., categorized as not changed on the CR but changed on the ROC-based threshold). These values were perfect for the CR at 68% and 50%, with no false positives at these levels of confidence. For improvement, kappa values for agreement between CRs and the ROC-generated threshold were lower than for deterioration, though agreement was good for CR at the 68% (κ = 0.68) and excellent for the 50% confidence level (κ = 0.83). Sensitivity and NPV for all confidence levels were 1.00 with no false negatives observed. Specificity and PPV increased as the CR confidence level decreased and false positives decreased.

Regarding CRs and the threshold suggested by the mean-change analysis, for deterioration, the best agreement was observed with the CR at 95% (κ = 0.88), with sensitivity and NPV perfect (no false negatives), a specificity of 0.96, and PPV of 0.83. Excellent agreement was also observed for the deterioration CR at 68% confidence (κ = 0.80), and good agreement was observed for the 50% level of confidence (κ = 0.65). For improvement, the best agreement was observed with the CR at 50% confidence (κ = 0.83), which met the standard for excellent agreement. (Table 3.) There were no false negatives for improvement at any level of confidence resulting in perfect sensitivity and

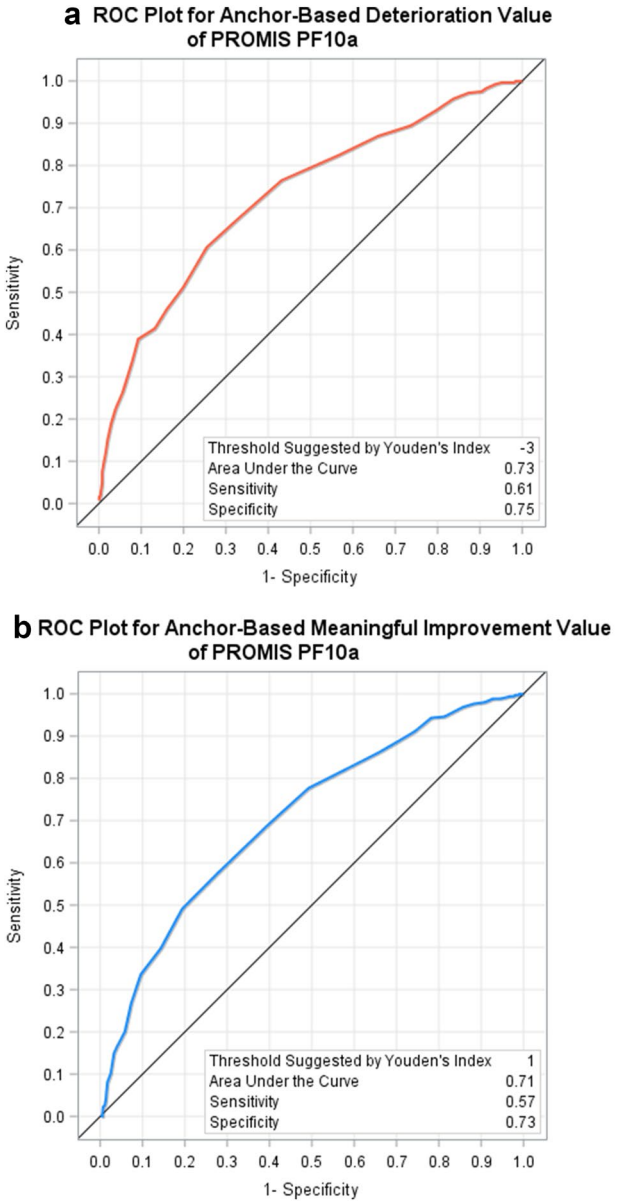

Fig. 2 **a** ROC plot for anchor-based deterioration value of PROMIS PF10a. **b** ROC plot for anchor-based improvement value of PROMIS PF10a

NPV. The proportion of false positives declined with level of confidence.

Regarding agreement between categorizing patients as deteriorated on the PROMIS PF10a using change thresholds generated from the CRs and categorizing patients as deteriorated using the PGIC anchor directly (i.e., not using the PGIC anchor to generate thresholds on the PROMIS PF10a), the contingency tables and kappa statistics are given in Table 4. Kappas always fell below the cut-offs for poor agreement (<0.40). The proportion of false positives (categorized as not changed on the CRs but changed on the PGIC) decreased with the level of confidence, while the proportion of true positives increased. Likewise, sensitivity decreased substantially under these conditions, but specificity increased slightly. Agreement between CR thresholds for improvement and improvement indicated on PGIC was also always poor, with kappas ranging between 0.16 and 0.27. The proportion of false positives decreased with the confidence level. At the same time, the proportion of false negatives increased modestly as confidence level decreased. While the proportion of true negatives decreased slightly, it was always around 60%.

Discussion

In comparison to the CR at 95% confidence, relaxing the confidence level to 50% or 68% to generate LCIs tended to agree better with anchor-based estimates of meaningful change. Even with precise measures such as the PROMIS PF10a, there is non-trivial error in individual measurement, especially when one considers change. Requiring 90–95% confidence that a score has changed over time therefore risks classifying people who have experienced meaningful change as unchanged because larger estimates of change were needed to reach that level of certainty. Assigning thresholds for individual change should include an understanding of the likelihood that the change is due to chance, at least at a conceptual level, and whether it reflects the level of change a patient would find meaningful but not exclude too many patients who feel they have changed. [25; 26]. The LCI statistics presented here accomplish these goals for the PROMIS PF10a.

Our goal in this paper is not to relax the confidence level on the RCI or CR to the point that it matches thresholds from group-level meaningful change and claim to have identified an all-encompassing statistic that captures both statistical significance and meaningfulness. Instead, we highlight the value of LCIs to use two valuable sources of information (significance and meaningfulness) to classify individuals most appropriately as either having changed or not in a given research or clinical application. For some applications, lower levels of confidence that the individual has changed may be appropriate. For example, identification of patients whose current magnitude of change is smaller but who may experience larger change in the future may warrant lower levels of confidence. An additional example of when lower levels of confidence may be useful is studies or clinical applications where the goal is to screen for deterioration on the PRO of interest and there is preference for specificity over sensitivity.

We also ask the reader to consider that many studies will classify individual patients as having changed based on anchor-based criteria appropriate for group-level

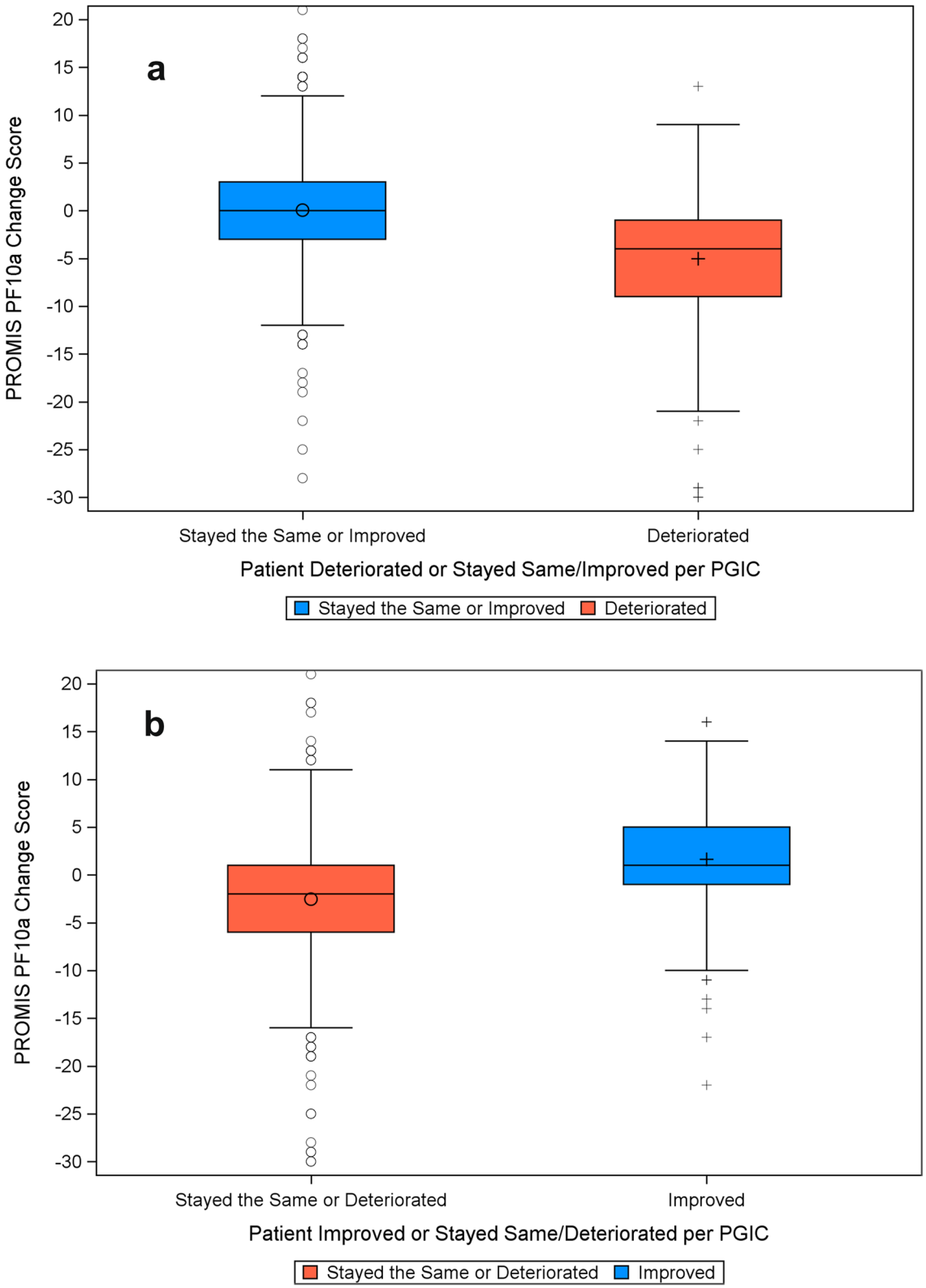

Fig. 3 **a** PROMIS PF10a change scores by PGIC-indicated physical function deterioration. **b** PROMIS PF10a change scores by PGIC-indicated physical function improvement

analyses, such as the mean change and ROC approaches. Such approaches are recommended in the U.S. FDA's guidance [27; 28] and have been championed in the PRO change literature [29]. Regardless of whether the intent of these methods is to reveal meaningful change, it remains a fact that each meaningful change threshold has some level of detectability given a particular sample of patients as a function of the PRO's measurement error. Given that our analyses showed that anchor-based thresholds are closer to the LCIs at 68% and 50% for the PROMIS PF10a, we may consider that many meaningful change thresholds would tend to have about that level of detectability; i.e., classifying individuals with anchor-based, meaningful change thresholds would result in a 32% probability (where it is close to the LCI at 68% confidence) or 50% (where it is close to the LCI at 50% confidence) that the patient would be classified as having changed even if they have not. These scenarios may or may not be acceptable to users of PRO measures; that depends on the specific researcher and specific application. The important point for the present study is to recommend researchers take this information into account when deciding what thresholds to apply.

Put another way, even if our goal is to capture meaningful change, we must acknowledge that the statistical properties

Table 2 Agreement between deterioration and improvement classification thresholds: coefficient of repeatability (CR) thresholds on PROMIS PF 10a change vs. ROC-based threshold on PROMIS PF 10a change

	Deterioration per ROC-Based Threshold (4 points)		Statistic	Value
Deterioration per CR 95% (7 points)	Yes, n (%)	No, n (%)	Kappa (95% CI)	0.69 (0.64, 0.74)
Yes, n (%)	189 (17%)	0 (0%)	Sensitivity	1.00
No, n (%)	125 (11%)	815 (72%)	Specificity	0.87
			Positive Predictive Value	0.60
			Negative Predictive Value	1.00
	Improvement per ROC-Based Threshold (2 points)		Statistic	Value
Improvement per CR 95% (7 points)	Yes, n (%)	No, n (%)	Kappa (95% CI)	0.35 (0.29, 0.40)
Yes, n (%)	88 (8%)	0 (0%)	Sensitivity	1.00
No, n (%)	236 (21%)	805 (71%)	Specificity	0.77
			Positive Predictive Value	0.27
			Negative Predictive Value	1.00
	Deterioration per ROC-Based Threshold (4 points)		Statistic	Value
Deterioration per CR 68% (4 points)	Yes, n (%)	No, n (%)	Kappa (95% CI)	1.00 (1.00, 1.00)
Yes, n (%)	314 (28%)	0 (0%)	Sensitivity	1.00
No, n (%)	0 (0%)	815 (72%)	Specificity	1.00
			Positive Predictive Value	1.00
			Negative Predictive Value	1.00
	Improvement per ROC-Based Threshold (2 points)		Statistic	Value
Improvement per CR 68% (4 points)	Yes, n (%)	No, n (%)	Kappa (95% CI)	0.68 (0.63, 0.73)
Yes, n (%)	193 (17%)	0 (0%)	Sensitivity	1.00
No, n (%)	131 (12%)	805 (71%)	Specificity	0.86
			Positive Predictive Value	0.60
			Negative Predictive Value	1.00
	Deterioration per ROC-Based Threshold (4 points)		Statistic	Value
Deterioration per CR 50% (3 points)	Yes, n (%)	No, n (%)	Kappa (95% CI)	0.85 (0.82, 0.88)
Yes, n (%)	314 (27%)	74 (7%)	Sensitivity	0.81
No, n (%)	0 (0%)	741 (66%)	Specificity	1.00
			Positive Predictive Value	1.00
			Negative Predictive Value	0.91
	Improvement per ROC-Based Threshold (2 points)		Statistic	Value
Improvement per CR 50% (3 points)	Yes, n (%)	No, n (%)	Kappa (95% CI)	0.83 (0.80, 0.87)
Yes, n (%)	252 (22%)	0 (0%)	Sensitivity	1.00
No, n (%)	72 (7%)	805 (71%)	Specificity	0.92
			Positive Predictive Value	0.78
			Negative Predictive Value	1.00

Table 3 Agreement between deterioration and improvement classification thresholds: coefficient of repeatability (CR) thresholds on PROMIS PF 10a change vs. mean-change-based threshold on PROMIS PF 10a change

	Deterioration per Mean-Change-Based Threshold (6 points)		Statistic	Value
Deterioration per CR 95% (7 points)	Yes, n (%)	No, n (%)	Kappa (95% CI)	0.88 (0.85, 0.92)
Yes, n (%)	189 (17%)	0 (0%)	Sensitivity	1.00
No, n (%)	40 (3%)	900 (80%)	Specificity	0.96
			Positive Predictive Value	0.83
			Negative Predictive Value	1.00
	Improvement per Mean-Change-Based Threshold (2 points)		Statistic	Value
Improvement per CR 95% (7 points)	Yes, n (%)	No, n (%)	Kappa (95% CI)	0.35 (0.29, 0.40)
Yes, n (%)	88 (8%)	0 (0%)	Sensitivity	1.00
No, n (%)	236 (21%)	805 (71%)	Specificity	0.77
			Positive Predictive Value	0.27
			Negative Predictive Value	1.00
	Deterioration per Mean-Change-Based Threshold (6 points)		Statistic	Value
Deterioration per CR 68% (4 points)	Yes, n (%)	No, n (%)	Kappa (95% CI)	0.80 (0.75, 0.84)
Yes, n (%)	229 (20%)	85 (8%)	Sensitivity	0.73
No, n (%)	0 (0%)	815 (72%)	Specificity	1.00
			Positive Predictive Value	1.00
			Negative Predictive Value	0.91
	Improvement per Mean-Change-Based Threshold (2 points)		Statistic	Value
Improvement per CR 68% (4 points)	Yes, n (%)	No, n (%)	Kappa (95% CI)	0.68 (0.63, 0.73)
Yes, n (%)	193 (17%)	0 (0%)	Sensitivity	1.00
No, n (%)	131 (12%)	805 (71%)	Specificity	0.86
			Positive Predictive Value	0.60
			Negative Predictive Value	1.00
	Deterioration per Mean-Change-Based Threshold (6 points)		Statistic	Value
Deterioration per CR 50% (3 points)	Yes, n (%)	No, n (%)	Kappa (95% CI)	0.65 (0.61, 0.70)
Yes, n (%)	229 (20%)	159 (14%)	Sensitivity	0.59
No, n (%)	0 (0%)	741 (66%)	Specificity	1.00
			Positive Predictive Value	1.00
			Negative Predictive Value	0.82
	Improvement per Mean-Change-Based Threshold (2 points)		Statistic	Value
Improvement per RCI 50% (3 points)	Yes, n (%)	No, n (%)	Kappa (95% CI)	0.83 (0.80, 0.87)
Yes, n (%)	252 (22%)	0 (0%)	Sensitivity	1.00
No, n (%)	72 (7%)	805 (71%)	Specificity	0.92
			Positive Predictive Value	0.78
			Negative Predictive Value	1.00

of PROs, such as reliability and error, hinder our ability to hear the patient's voice clearly as captured on PROs. For this reason, reliability, and its implications for how certain we can be that PRO scores reflect signal and not noise, is an indispensable element of estimating individual change. A common rule of thumb suggests that measures should demonstrate at least 90% reliability for applications with individual patients [30]. The analyses conducted in this study reflect this high level of reliability, but the results would have differed if reliability were lower in that larger thresholds would be required for statistically significant change. While it is difficult to increase the reliability of a PRO without adding additional items, and therefore additional administration burden, the benefits of higher reliability may influence

measure selection for applications with individual patients. For example, owing to their reliance on IRT, PROMIS measures tend to be more reliable at the score level than many classical test theory (CTT)-based PROs, especially when administered as computer adaptive tests [31; 32].

There are some limitations to consider. First, as we note above, the level of confidence used with the LCI should reflect the needs of the application. We caution potential users that as the confidence level approaches 50%, the likelihood of that change being due to chance increases; on the other hand, the likelihood of classification (changed versus not changed) based on average perceived level of meaningfulness may also increase. Indeed, the LCI at 50% reflects equal likelihood that the change is or is not due to chance. A researcher using the LCI at lower confidence levels must accept a level of uncertainty around whether the change is due to measurement error alone (not true change). Related to this point, we note that the LCI at 50% will always be slightly lower than the SEM for any application, regardless of the measure used and the sample

Table 4 Agreement between Deterioration and Improvement Classification: Coefficient of Repeatability (CR) Thresholds on PROMIS PF 10a Change vs. Physical Function Patient Global Impression of Change (PGIC)

Deterioration per CR 95%	Deterioration per PGIC		Statistic	Value
	Yes, n (%)	No, n (%)	Kappa (95% CI)	0.33 (0.27, 0.40)
Yes, n (%)	111 (10%)	78 (7%)	Sensitivity	0.59
No, n (%)	174 (15%)	766 (68%)	Specificity	0.81
			Positive Predictive Value	0.39
			Negative Predictive Value	0.91
	Improvement per PGIC		Statistic	Value
Improvement per CR 95%	Yes, n (%)	No, n (%)	Kappa (95% CI)	0.16 (0.11, 0.21)
Yes, n (%)	58 (5%)	30 (3%)	Sensitivity	0.66
No, n (%)	296 (26%)	745 (66%)	Specificity	0.72
			Positive Predictive Value	0.16
			Negative Predictive Value	0.96
	Deterioration per PGIC		Statistic	Value
Deterioration per CR 68%	Yes, n (%)	No, n (%)	Kappa (95% CI)	0.30 (0.24, 0.36)
Yes, n (%)	146 (13%)	168 (15%)	Sensitivity	0.46
No, n (%)	139 (12%)	676 (60%)	Specificity	0.83
			Positive Predictive Value	0.51
			Negative Predictive Value	0.80
	Improvement per PGIC		Statistic	Value
Improvement per CR 68%	Yes, n (%)	No, n (%)	Kappa (95% CI)	0.27 (0.22, 0.33)
Yes, n (%)	119 (10%)	74 (7%)	Sensitivity	0.62
No, n (%)	235 (21%)	701 (62%)	Specificity	0.75
			Positive Predictive Value	0.34
			Negative Predictive Value	0.90
	Deterioration per PGIC		Statistic	Value
Deterioration per CR 50%	Yes, n (%)	No, n (%)	Kappa (95% CI)	0.31 (0.26, 0.37)
Yes, n (%)	173 (15%)	215 (19%)	Sensitivity	0.45
No, n (%)	112 (10%)	629 (56%)	Specificity	0.85
			Positive Predictive Value	0.61
			Negative Predictive Value	0.75
	Improvement per PGIC		Statistic	Value
Improvement per CR 68%	Yes, n (%)	No, n (%)	Kappa (95% CI)	0.28 (0.22, 0.34)
Yes, n (%)	141 (12%)	111 (10%)	Sensitivity	0.56
No, n (%)	213 (19%)	664 (59%)	Specificity	0.76
			Positive Predictive Value	0.40
			Negative Predictive Value	0.86

This contingency table is between patients classified as deteriorated or improved on the PROMIS PF10a using thresholds from the CR classifications of deterioration or improvement and on the PGIC anchor directly (not based on a classification of PROMIS PF10a scores)

studied. Two terms from the LCI's equation, the critical value and $\sqrt{2}$, are sample and measure independent. The LCI at 50%'s critical value is 0.674, and when multiplied by $\sqrt{2}$ yields 0.953. This value multiplied by the SEM, which yields the LCI at 50%, will always be 95% of the SEM. Second, our findings are from a single sample of patients. Though this sample boasts some attractive qualities in terms of its diversity and having been drawn from multiple study sites, our results may be sample-specific. Third, while we relied upon current standard approaches to estimate meaningful change with anchors at the group level, recent research has pointed to flaws in these methods [7; 33], including both the ROC method [6] and the mean-change method [29; 34]. Use of newer methods may have improved our estimates of group-level meaningful change, but our objective here was to compare CR-based thresholds with those from the most common approaches. It is important to consider the lack of agreement between the individual-level thresholds for meaningful change based on the PGIC directly with the CR-based thresholds. It is not difficult to understand why the individual-level indicator of meaningful change would agree less often with the CR since individual variation is high, and this variation is averaged-out at the group level. This has ramifications for the LCIs suggested here; we note that they can be said to agree with change that patients, on average, find to be meaningful but not necessarily with what the individual the LCI is calculated for finds meaningful. Yet, we also acknowledge the potential that some known biases in the PGIC (e.g., recall bias) may have affected these results [8]. Regarding calculation of the RCI and LCI, some may consider that the SD of change is more appropriate to use than the SD of the baseline. Future research should address this topic. Finally, we elected to define thresholds from meaningful and significant change estimates by rounding up to the nearest integer. This approach has the benefit of ensuring that all patients counted as having changed have exceeded the change estimate. On the other hand, when the estimate is closer to integer below it than the one above it, our approach may be conservative when classifying patients as having changed.

In conclusion, LCIs using 68% or 50% confidence may be a good way to balance multiple, potentially competing needs for estimates of individual patient change on PROs. While relaxing the confidence level for the RCI entails a trade-off in terms of the amount of certainty we have that a patient has changed, it is more often aligned with the amount of change that, on average, patients find meaningful. To maximize the proportion of appropriately classified individuals across improved, unchanged, and deteriorated categories, researchers should consider use of the LCI.

Acknowledgements We express our sincere appreciation to the investigators and patients who contributed data from: "Assessing PROMIS and other simple patient-reported measures for cancer research" (J. Sloan, PI, R01-CA154537)

Author contributions JDP wrote the first draft and performed data analysis. RDH and DC provided critical edits and data interpretation.

Funding Dr. Hays was supported in part by the UCLA Resource Center for Minority Aging Research/Center for Health Improvement of Minority Elderly (RCMAR/CHIME) funded by National Institutes of Health (NIH), National Institute on Aging (NIA) P30-AG021684. Dr. Cella was supported by the 2UG1CA189828 subaward to Northwestern University from ECOG-ACRIN Cooperative Group. Dr. Peipert was supported in part by a grant from the Peter G. Peterson Foundation (#19041; PI, Cella) and a grant from the National Cancer Institute (U01CA233169; mPIs, Gray and Wagner).

Data availability The data used in this study are not available upon request.

Code availability N/A.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval The study was reviewed by the IRB of each of the participating sites, and all patients provided consent to enter the study.

Consent to participate N/A.

References

1. McNemar, Q. (1958). On growth measurement. *Educational and Psychological Measurement, 18*(1), 47–55.
2. Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement, 16*(4), 421–437.
3. Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin, 74*(1), 68–80.
4. US Food and Drug Administration. (2019). *Discussion document for patient-focused drug development public workshop on guidance 4: Incorporating clinical outcome assessments into endpoints for regulatory decision-making*. Silver Spring, MD: United States Department of Health and Human Services.
5. Terwee, C. B., Peipert, J. D., Chapman, R., Lai, J. S., Terluin, B., Cella, D., Griffith, P., & Mokkink, L. B. (2021). Minimal important change (MIC): A conceptual clarification and systematic review of MIC estimates of PROMIS measures. *Quality of Life Research, 30*(10), 2729–2754.
6. 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.
7. Terluin, B., Eekhout, I., Terwee, C. B., & de Vet, H. C. (2015). Minimal important change (MIC) based on a predictive modeling approach was more precise than MIC based on ROC analysis. *Journal of Clinical Epidemiology, 68*(12), 1388–1396.

8. 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.
9. Hays, R. D., & Peipert, J. D. (2018). Minimally important differences do not identify responders to treatment. JOJ Sciences, 1(1).
10. Hays, R. D., Brodsky, M., Johnston, M. F., Spritzer, K. L., & Hui, K. K. (2005). Evaluating the statistical significance of health-related quality-of-life change in individual patients. *Evaluation and the Health Professions, 28*(2), 160–171.
11. Moinpour, C. M., Donaldson, G. W., Davis, K. M., Potosky, A. L., Jensen, R. E., Gralow, J. R., Back, A. L., Hwang, J. J., Yoon, J., Bernard, D. L., Loeffler, D. R., Rothrock, N. E., Hays, R. D., Reeve, B. B., Smith, A. W., Hahn, E. A., & Cella, D. (2017). The challenge of measuring intra-individual change in fatigue during cancer treatment. *Quality of Life Research, 26*(2), 259–271.
12. King, M. T., Dueck, A. C., & Revicki, D. A. (2019). Can methods developed for interpreting group-level patient-reported outcome data be applied to individual patient management? *Medical Care, 57*, S38–S45.
13. Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12–19.
14. Cella, D., Bullinger, M., Scott, C., & Barofsky, I. (2002). Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life. *Mayo Clinic Proceedings, 77*(4), 384–392.
15. Donaldson, G. (2008). Patient-reported outcomes and the mandate of measurement. *Quality of Life Research, 17*(10), 1303–1313.
16. Lee, M. K., Schalet, B. D., Cella, D., Yost, K. J., Dueck, A. C., Novotny, P. J., & Sloan, J. A. (2020). Establishing a common metric for patient-reported outcomes in cancer patients: Linking patient reported outcomes measurement information system (PROMIS), numerical rating scale, and patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE). *J Patient Rep Outcomes, 4*(1), 106.
17. Jensen, R. E., Potosky, A. L., Reeve, B. B., Hahn, E., Cella, D., Fries, J., Smith, A. W., Keegan, T. H. M., Wu, X.-C., Paddock, L., & Moinpour, C. M. (2015). Validation of the PROMIS physical function measures in a diverse US population-based cohort of cancer patients. *Quality of Life Research, 24*(10), 2333–2344.
18. Wahl, E., Gross, A., Chernitskiy, V., Trupin, L., Gensler, L., Chaganti, K., Michaud, K., Katz, P., & Yazdany, J. (2017). Validity and responsiveness of a 10-item patient-reported measure of physical function in a rheumatoid arthritis clinic population. *Arthritis Care & Research, 69*(3), 338–346.
19. Oken, M. M., Creech, R. H., Tormey, D. C., Horton, J., Davis, T. E., McFadden, E. T., & Carbone, P. P. (1982). Toxicity and response criteria of the Eastern Cooperative Oncology Group. *American Journal of Clinical Oncology, 5*(6), 649–655.
20. Hays, R. D., & Peipert, J. D. (2021). Between-group minimally important change versus individual treatment responders. *Quality of Life Research, 30*(10), 2765–2772.
21. SAS Institute Inc. (2021). Plot ROC curve with cutpoint labeling and optimal cutpoint analysis. Retrieved September 29, 2021, from https://support.sas.com/kb/25/018.html
22. Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*(1), 32–35.
23. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.
24. Fleiss, J. L., Levin, B., & Paik, M. C. (2004). The measurement of interrater agreement. In Statistical methods for rates and proportions (pp. 598–626). John Wiley & Sons, Inc.
25. Terwee, C. B., Terluin, B., Knol, D. L., & de Vet, H. C. W. (2011). Combining clinical relevance and statistical significance for evaluating quality of life changes in the individual patient. *Journal of Clinical Epidemiology, 64*(12), 1465–1467.
26. Terwee, C. B., Roorda, L. D., Knol, D. L., De Boer, M. R., & De Vet, H. C. W. (2009). Linking measurement error to minimal important change of patient-reported outcomes. *Journal of Clinical Epidemiology, 62*(10), 1062–1067.
27. US Food and Drug Administration. (2009). *Guidance for industry patient-reported outcome measures: Use in medical product development to support labeling claims*. Rockville, MD: US Department of Health and Human Services.
28. US Food and Drug Administration. (2018). *Discussion document for patient-focused drug development public workshop on guidance 3: Select, develop or modify fit-for-purpose clinical outcome assessments*. Silver Spring, MD: United States Department of Health and Human Services.
29. Coon, C. D., & Cook, K. F. (2018). Moving from significance to real-world meaning: Methods for interpreting change in clinical outcome assessment scores. *Quality of Life Research, 27*(1), 33–40.
30. Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.
31. Segawa, E., Schalet, B., & Cella, D. (2020). A comparison of computer adaptive tests (CATs) and short forms in terms of accuracy and number of items administrated using PROMIS profile. *Quality of Life Research, 29*(1), 213–221.
32. Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., Devellis, R., DeWalt, D., Fries, J. F., Gershon, R., Hahn, E. A., Lai, J. S., Pilkonis, P., Revicki, D., & Goup, P. C. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *J Clin Epidemiol, 63*(11), 1179–1194.
33. Terluin, B., Griffiths, P., van der Wouden, J. C., Ingelsrud, L. H., & Terwee, C. B. (2020). Unlike ROC analysis, a new IRT method identified clinical thresholds unbiased by disease prevalence. *Journal of Clinical Epidemiology, 124*, 118–125.
34. 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.