Comparison of Anchor- and Distribution-Based Methods for Estimating Thresholds of Meaningful Within-Patient Change Using Simulated PROMIS PF 20a Data Under Various Joint Distribution Characteristic Conditions

Shanshan Qin (sqin@rti.org)
RTI Health Solutions Research Triangle Park https://orcid.org/0000-0002-8574-5508
Lauren Nelson
RTI Health Solutions Research Triangle Park
Nicole Williams
RTI Health Solutions Research Triangle Park
Valerie Williams
RTI Health Solutions Research Triangle Park
Randall Bender
RTI Health Solutions Research Triangle Park
Lori McLeod
RTI Health Solutions Research Triangle Park

Research Article

Keywords: Patient-reported outcome, distribution-based method, anchor-based method, responder definition, meaningful change, minimal important difference, MID, minimal clinically important difference, MCID

Posted Date: May 27th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-381557/v1

Abstract

Purpose

To compare the performance of anchor-based and distribution-based methods for estimating thresholds of meaningful within-patient change of clinical outcome assessments in conditions reflecting data characteristics of small- to medium-sized clinical trials.

Methods

Data sets were generated from the joint distributions of the PROMIS PF 20a T-score changes and a seven-point global change anchor measure. The 108 simulation conditions (1,000 replications per condition) included combinations of three marginal distributions of T-score changes, three improvement percentages in the anchor measure, four levels of responsiveness correlations, and three sample sizes. Threshold estimation methods included mean change, median change, ROC curve, predictive modeling, half SD, and SEM. Relative bias, precision, accuracy, and measurement significance of the estimates were evaluated based on comparison with true thresholds and IRT-based individual reliable changes of PROMIS scores. Quantile regression models were applied to select and interpret effects of simulation conditions on estimation bias.

Results

When PROMIS T-score changes were distributed normally, the predictive modeling method performed best with 50% or more responders identified by the anchor; the mean and median methods were preferred with 30% responders. For skewed distributions, the median method and ROC method gained more advantages. Among the evaluated study conditions, the improvement percentage condition had the most obvious effects on estimation bias.

Conclusion

To establish accurate and precise thresholds, clinical researchers are recommended to prioritize study designs with at least 50% anchor-defined responders and strongly responsive target endpoints with highly reliable scoring calibration and to select optimal anchor-based methods given the data characteristics.

Introduction

Statistically significant mean change on clinical outcome assessment (COA) scores does not necessarily correspond to meaningful changes from the perspective of patients, observers, or clinicians [1]. To guide interpretation of COA score changes for regulatory review, thresholds characterizing a clinically meaningful within-patient change (or responder thresholds) are commonly requested by the United States Food and Drug Administration (FDA) in submissions for COA labeling claims [1, 2]. Patients can be classified as either “responders” or “nonresponders” based on the threshold to felicitate later treatment efficiency evaluation.

Given the high stakes of COAs used as primary or key secondary endpoints, it is important to ensure that thresholds of meaningful within-patient changes are developed using thoughtful study design and rigorous methods. Of the available estimation methods, the current FDA COA guidance documents recommend anchor-based methods as the primary approach, with distribution-based methods being supportive [1, 2]. Anchor-based methods rely on an external criterion to characterize meaningful improvement, such as change based on a relevant global rating or change in a well-established outcome (e.g., biomarker or COA) with an accepted threshold identifying clinical improvement. Commonly used anchor-based methods include anchor-based mean, anchor-based median, receiver operating characteristic (ROC) curve, and predictive modeling methods [3–5]. Distribution-based methods use the target COA’s statistical or measurement attributes (e.g., the half standard deviation [SD] of the baseline COA score or the standard error of measurement [SEM] [7–9]) as a guide for interpreting change.

To date, few studies have systematically evaluated the performance of different estimation methods for meaningful within-patient change thresholds under various combinations of data conditions [3, 9]. This study attempts to provide information to address this concern by using a simulation study for the performance comparison under conditions designed to reflect realistic characteristics of clinical trial or observational data. Although the current study focuses on anchor-based methods, two distribution-based methods were also incorporated to provide further context and supportive boundaries. Estimation performance was evaluated by comparing the sample estimates of each method to the true thresholds under the corresponding simulation conditions and to the respective individual measurement precision. Additionally, the impacts of simulation conditions on the biases in anchor-based estimates were evaluated to facilitate researchers’ planning on key study design and post-hoc data selection for the application of these methods.

The current simulation considered the systematic impacts of four data conditions. The first condition was the distribution of target COA changes, characterized by differences in variance and skewness. The second condition was the proportion of patients reporting improvement on a global anchor measure, which was used to identify the population-level true threshold of meaningful within-patient change in target COA. The third condition was the relationship strength (i.e., correlation) between target COA change score and anchor measure. Revicki et al. [11] suggested a correlation greater than 0.30; Hays et al. [12] suggested at least 0.371; and de Vet et al. [13] recommended a higher bar of 0.50. The fourth condition was overall sample size. Although it is essential to ensure that the estimation methods provide consistent thresholds with large sample sizes, it is also important to understand how the methods behave with small sample sizes, as the recruitment of large sample sizes is often not feasible in rare disease clinical trials.

Methods

Simulation

The Patient-Reported Outcomes Measurement Information System (PROMIS)® Item Bank v2.0 Physical Function Short Form 20a (PF SF 20a) was used as the target COA. The PF SF 20a contains 20 items, each with a response scale of 1 (unable to do) to 5 (without any difficulty). Individual item attributes (e.g., discrimination, difficulty) contribute to the overall PROMIS T-score in a scaled range of − 11.7 to 62.7 points (which corresponds to a raw sum score of 20 to 99), where higher values indicate better physical function.

The advantage of using the PROMIS T-score as the target COA score is the availability of individual standard errors (SEs) based on item response theory (IRT) and the existing large-scale calibration, which can be used to compute individual reliable changes (RCs) to evaluate threshold estimates in terms of within-person statistical significance conditioned on the available COA measurement precision [11]. However, a challenge in defining SEs for simulated PROMIS T-scores is that different item response patterns can result in the same T-score but have somewhat varying SEs. To streamline the simulation and provide a look-up reference table for interested researchers, the 2,436,200 possible response patterns of the 20 PROMIS items were generated, and the 180,979 patterns were scorable using the HealthMeasures Scoring Service website [14]. This exercise yielded a distribution of IRT-based SEs for each T-score (Supplemental Table 1). The 99 th percentile of the SEs for each T-score was used as a conservative referencing SE to compute the individual RCs for later performance comparison. Mean (SD) baseline T-scores (35 [5]) were generated from a normal distribution, which represents a patient population with moderate to severe physical function impairment according to the general guidelines for interpreting PROMIS scores in the 2000 US normative sample [14].

Table 1 provides an overview of the simulation study design conditions.

1. PROMIS T-score change distribution: Changes were sampled from three marginal distributions. The first two distributions were normal with the same mean but different SD, and the third was a negatively skewed distribution with the same mean and an approximate SD to match the second distribution. The population mean of the T-score change was fixed to 7 such that at least 50% of subjects could achieve change above an extreme RC of 6.9 (computed from an extreme SE = 2.5 for the T-scores). This is intended to represent a trial or other longitudinal study designed with an effect size of overall change that PROMIS PF SF 20a can detect for at least 50% of the subjects.

2. Anchor measure distribution: The anchor measure was a hypothetical seven-category Patient Global Impression of Change (PGIC) generated separately from three types of marginal distributions (Table 1) characterized by levels of meaningful improvement or responder percentages: 30%, 50%, and 70%.

3. Strength of correlation: To create four weak to strong Spearman correlations (p = 0.1, 0.3, 0.5, and 0.7) between the T-score change and PGIC, the Iman-Conover method was implemented through a series of matrix factorization, multiplication, and pairing rearrangement [15, 17]. The method generated monotonic relationships between the two variables, without specifying a precise linear or nonlinear model or distributional assumptions (unlike for the polyserial correlation), to extend the generalizability of the current results.

4. Sample size: Three sample sizes (n = 50, 100, and 300) were simulated. The first sample size represented a reasonable scenario for rare disease. The second and third reflected typically sized clinical trials evaluating COA measures in practice [17].

When combined, the design conditions yielded 108 settings (3×3×4×3). For each setting, 1,000 datasets were generated via repeated sampling. All data generation and analyses were performed using SAS version 9.4 or higher for Windows statistical software [19].

Methods for estimating thresholds of within-patient meaningful change

The anchor-based mean method, anchor-based median method, ROC curve analysis (using logistic regression and optimizing sensitivity and specificity), and a predictive modeling method (using logistic regression) (Table 2) were applied to each simulated dataset [3, 3, 5, 12, 13, 20]. Two distribution-based methods were applied to provide supportive estimates: half SD and SEM at baseline. Finally, individual RCs served as reference values to evaluate the estimates, given that these IRT-based values are constant for specific item response patterns.

Performance evaluation criteria

The 1,000 threshold estimates per method in every simulation setting were compared with two types of reference values: the population-level true thresholds and the individual RCs.

Comparison with true thresholds

For each simulation setting, the true threshold was defined as the quantile of T-score change corresponding to the target population-level anchor-based percentage of improvement. For example, for the 30%-improvement condition, the true threshold corresponded to the 0.7 quantile of the normal distribution of T-score change and the 0.30 quantile of Gamma (shape = 1.5, scale = 6) which was then subtracted from 16 (Table 1).

Three performance statistics were computed: relative bias (RB) as symmetric difference from the true value, coefficient of variation (CV) as random error around the average of the estimates, and relative root mean squared error (rRMSE) to quantify how accuracy was impacted by both bias and precision. These statistics were computed as percentages to facilitate comparison across settings:

$$RB = 100 \times \frac{E(T)-\tau}{\tau}, CV = 100 \times \frac{\sqrt{Var(T)}}{E(T)}, rRMSE = 100 \times \frac{\sqrt{(E(T)-\tau)^2+Var(T)}}{\tau},$$

where T was one estimate from one sample; $E(T)$ and $Var(T)$ were the τ mean and variance of the 1,000 estimates per method × simulation setting, respectively; and τ was the corresponding true threshold. Relative bias closest to 0% (no bias) is preferred, with the positive or negative direction indicating the

risk of misclassifying responders or increasing false responders, respectively. Smaller CV and rRMSE values indicate better estimation with higher precision and higher accuracy.

Comparison with individual reliable change

For each subject in each simulated sample, the individual RC was computed:

$$RC = 1.96\sqrt{SE_{BL}^2 + SE_{FU}^2},$$

where SE_{BL} and SE_{FU} were reference SE of T-score at baseline and follow-up, respectively (Supplemental Table 1). Two performance statistics were computed: percentage of subjects whose individual RCs were not greater than an estimate, and the median positive difference of the threshold estimate minus RC within every simulated sample. Higher values of both indicate that the threshold estimate was more likely to exceed individual measurement errors, one necessary aspect for an appropriate within-person threshold. The distributions of the first statistic were tabulated. The values of the second statistic were displayed in probability density function (PDF) plots.

Impact and interaction of study data characteristics within each anchor-based method

The predictors (Table 1) of estimation bias (i.e., (T- τ) for each anchor-based method were screened through a quantile regression selection process using an adaptive Lasso method [21]. Candidate predictors were treated as categorical, as were all higher-order interactions. Due to nonconvergence of the full model using all candidate predictors, separate quantile regression selection processes were implemented and stratified by distribution type of T-score change, which, however, limited the later evaluation of interactions with distribution type. Based on results from these stratified models, a subset of predictors was selected. These selected predictors, their lower-order terms, and the distribution type of T-score change were used to predict estimation bias in a final quantile regression model for each anchor-based method. The effects of significant predictors on the 0.25, 0.50, and 0.70 quantiles of estimation bias were plotted.

Results

Performance evaluation

Table 3 presents results for the comparison between estimated and true thresholds across the 108 simulation settings. No one method was the best performer overall (Supplemental Table 2). The strengths and weaknesses of the methods varied with settings and performance criteria. The most consistent observation was that the SEM method had the lowest CV values (smallest variability around the mean of every 1,000 estimates) across all settings—an inherent benefit of the large-scale IRT calibration of the PROMIS items. The second consistent observation was that, among all anchor-based methods, the predictive modeling method produced the lowest CV values across all settings. The third consistent observation was that the predictive modeling method was usually the best performer on all criteria, with 50% and 70% population-level improvement or subjects being classified as responders by the anchor with generally normal distributions of T-score change.

The remaining findings in Table 3 were more setting and/or criterion specific. To facilitate the interpretation, the between-method difference in the magnitude of a performance statistic was considered to be similar if it was < 1% and otherwise comparable if < 5%, or superior/inferior if ≥ 5%. When about 30% of subjects were responders with normally distributed T-score change, the RBs of the mean and median methods were similar and produced less bias than the two logistic methods. The mean method had the smallest rRMSE (highest accuracy) in these settings. When the T-score changes were negatively skewed, the median method was clearly superior with 30% of responders, yielding the smallest RBs and rRMSE; the ROC curve method performed best with 50% of responders, and the predictive modeling method showed advantages again with 70% of responders. Additionally, the RBs and rRMSE of the half-SD method were the smallest in magnitude across all methods under the normal distribution (7.0, 7.0) of T-score change and 70% of responders—this is likely due to the coincidence that the true threshold of 3.3 under these conditions was near the half SD of 2.5 for the simulated population (Supplemental Table 2).

Table 4 presents the minimum, median, and maximum values of the percentages of subjects with estimates greater than individual RCs for the 1,000 replications (see Supplemental Table 3 for complete lists). (The comparison with half SD and SEM was not included because, notably, the two distribution-based estimates were always smaller than individual RCs.) As Table 4 shows, the predictive modeling method tended to provide the most consistent protection against individual measurement errors for which the minimum percentages of subjects with individual RCs not greater than the estimates were almost always highest across the simulation settings. However, based on the median value of those percentages (≥ 95%), the threshold estimates of all four anchor-based methods were greater than RCs most of the time.

Figures 1a-b show the PDF plots of the within-sample median positive differences between the estimate and individual RCs (see Supplemental Fig. 1 for complete results). Similar to the CV performance results, the variability (spread) of the median positive difference values for the predictive modeling method (red curve) was the smallest. Furthermore, the modes and right tails of the predictive modeling method curves usually fall to the left of the other curves, indicating that when the estimates of the other methods (especially the median method [orange curves]) exceeded individual measurement errors, their positive differences tended to be larger (higher thresholds) than the difference from the predictive modeling method.

Significant impact and interaction of clinical data characteristics

The predictor selection process retained only first-order predictors and select two-way interactions. The reference predictor classes were designated as normal distribution (7.0, 3.5), 50% improvement, ρ = 0.70, and n = 300 because the overall rRMSE (systematic and random difference) tended to be minimal within those reference classes, despite a few local reverse trends (Table 3).

The selected predictors were very similar across the models for the different estimation methods. All estimated effects of significant predictors by the anchor-based methods are plotted in Fig. 2, except for the interaction between ρ and n, which yielded very small effect sizes (− 0.00 to 0.40) and almost undifferentiated lines if plotted (Supplemental Table 4). Across the anchor-based methods, the most prominent and consistent predictors were improvement percentage and its interaction with correlation strength (ρ)—departure from the reference classes (50% improvement and $\rho = 0.70$) generally increased estimation bias. For example, for the predictive modeling method (the anchor-based method most sensitive to varying improvement percentage), 70% improvement had a main effect of 1.0-point positive increase on the 0.50 quantile of estimation bias compared with 50% improvement; at $\rho = 0.30$, an additional 0.87-point positive increase was introduced due to the improvement percentage × ρ interaction.

The next most prominent and consistent predictors were ρ and n, which impacted the mean and median methods more than the other two methods (Fig. 2). Lower correlation generally produced a negative increase in the bias; for example, a main effect of 1.23-point negative increase was shown on the 0.50 quantile of estimation bias by the mean method when ρ was reduced from 0.70 to 0.10. Smaller sample sizes generally increased the bias positively or negatively depending on the quantile location.

With normal distributions, larger variance (i.e., population SD = 7.0) tended to increase bias positively for the mean and median methods. A comparison of skewed and normal distributions (both with SD ~ 7) indicated an ~ 1.0-point negative increase in bias with skewed distribution when using the mean and predictive modeling methods on the 0.50 quantile of estimation bias.

Discussion And Conclusion

Although no single recommended method exists for estimating thresholds of meaningful within-patient change, in practice researchers tend to use the anchor-based mean approach as the primary method and distribution-based approaches as supportive. Alternatively, researchers tend to prefer the median anchor-based method whenever the COA change scores or anchor-measure distributions are skewed [e.g., 22, 23]. Using data generated for changes in PROMIS PF SF 20a T-scores, our simulation study compared four widely recognized anchor-based and two distribution-based methods for estimating thresholds of meaningful within-patient change under conditions designed to mimic realistic clinical and observational studies.

As expected, among the anchor-based methods, the optimal choice depended on the clinical data characteristics. Although the results supported the common application of mean or median anchor-based methods, the results identified scenarios where the other methods should be strongly considered. Specifically, when ≥ 50% of participants were true responders and PROMIS change scores generally formed a normal distribution, the predictive modeling method performed best overall on controlling bias, increasing precision and accuracy, and exceeding individual measurement errors. Although this method did not always yield the smallest bias on average, its variability around mean estimates was almost the smallest among the anchor-based methods. This high precision was consistent with the simulation finding by Terluin et al. [6] that the 95% CI for the ROC curve was wider in length than that obtained by the predictive modeling method in the setting of 50% improvement prevalence and normal distribution of target COA change. The likely reason for this finding is that both logistic regression methods use the entire sample to locate the threshold estimate based on sensitivity, specificity, or odds, whereas the mean and median methods focus on the group at one anchor level (e.g.," minimally improved"). Therefore, higher precision (low CV)—especially for larger sample sizes for the two logistic methods—was not surprising.

With < 50% (e.g., 30%) of responders under normal distributions of T-score change, method preferences trended toward mean and median anchor-based methods for the smallest of RBs and satisfactory protection against measurement error most of the time. One major reason for this preference, as shown in Table 3 and Fig. 2, is that the mean and median methods had smaller increases in bias than the two logistic methods for the 30%-improvement group when the 50%-improvement group was used as the reference. At first glance, this finding seemed in conflict with the simulation findings by Terluin et al. [6], that changing the "prevalence of improvement" alone did not affect the estimates of the two logistic-based methods. However, the current study and Terluin et al. [6] applied different simulation conditions. The population percentage of improvement simulated for the anchor-based methods impacted the true threshold or responder definition in the current study, while the "prevalence of improvement" in Terluin et al. [6] may not have matched the underlying responder percentage. In Terluin et al. [6], the true threshold was fixed to 3.5 when the prevalence changed from 50–70%, but in the current study, the true thresholds varied with the population improvement percentage of the PGIC.

For skewed T-score change distributions, the median method and ROC curve method performed best at the conditions of 30% and 50% improvement, respectively. As shown in Table 3 and Fig. 2, this finding was likely related to the smaller effects of positive increases in bias due to skewed distributions and the countereffect of a negative increase on bias due to 30% improvement for the two methods, in contrast to the larger positive effects of both predictors on the mean method and predictive modeling method. In the 70%-improvement condition, the countereffects were observed in the predictive modeling and mean methods, while the combined positive increases further inflated the bias resulting from the other two methods.

Among the conditions investigated, the most suitable for minimizing rRMSE (hence reducing bias and increasing precision overall) was the setting related to a normal distribution (7.0, 3.5), 50% improvement, $\rho = 0.70$, and n = 300. As a result of the PROMIS IRT-based calibration, the SEM method consistently demonstrated much smaller CV values than the anchor-based methods and the half-SD method; the median within-sample percentages of subjects with individual RCs not greater than the anchor-based estimated thresholds was at least 95%. These findings highlight the importance of selecting a reliable (small random variance in measurement) and valid (adequate relationship with anchor measure) COA in addition to identifying a robust data source (where both responders and nonresponders are well represented) when conducting analyses to identify a meaningful within-patient change threshold. For example, if researchers intend to use interim data cuts of ongoing trials to establish the meaningful within-person threshold, it is sensible to wait until ~ 50% of the subjects can be considered responders, based on multiple anchor measures or external gold standards (where bias tends to be minimal and precision and accuracy tend to be maximized across methods), if feasible for related therapeutic areas. For literature reviews or meta-analyses of meaningful change, greater weight can be placed on thresholds estimated when approximately 50% of the participants were responders. Not surprising, this study's results further

emphasize the need for a strong responsiveness correlation—however, this does not imply that the correlation must be perfect, because the unique value of the target COA (in addition to the anchor measures) is established in theory and qualitatively. To maximize estimation precision, wise decisions must be made with respect to item selection, calibration, and scoring rule (i.e., valid, reliable, discriminative, highly intercorrelated items; raw versus pattern scoring; weekly versus monthly scores; and missing-data rule). As always, a larger sample and normal distribution of target COA change are desirable.

Finally, the half-SD and SEM methods generally underestimated the thresholds in most settings specified. This finding confirmed their roles as supportive estimates, in addition to the RC value, in identifying the minimal value when reporting a range of thresholds.

Limitations And Future Research

Although this study was designed to generalize to typical applications, there are limitations. This research focused on thresholds for detecting *improvement* in a COA; therefore, the results cannot be easily applied to COA thresholds for use in clinical trials or observational studies aimed at mitigating the progression (worsening) of a condition.

In addition, the correlation between PROMIS change and PGIC was simulated as a Spearman correlation to free the assumptions regarding linear relationship or normal distribution of the target COA change. Readers should be cautious if directly applying these findings to situations with other correlation types (e.g., Pearson).

Another important consideration is that the simulation used a retrospective anchor measure with minimal measurement error (only from random sampling). In practice, retrospective anchors could be subject to additional measurement error due to response-shift bias or recall bias [24]. Fayers and Hays [24] recommend inclusion of both retrospective and concurrent anchors (e.g., global ratings of current severity) in clinical trial designs. Our simulated PGIC values could be considered as change between two administrations of Patient Global Impression of Severity (PGIS) rating scales. However, PGIS change likely would have provided more levels than our simulated PGIC, resulting in use of a different type of correlation. Similar caution would be required in settings using a continuous anchor measure but only two response classes: “responder” versus “nonresponse” (e.g., a biomarker with only one reference cutoff or change in the 22-item Sinonasal Outcome Test using the recommended cutoff of − 8.9 [25]), which would result in more flexibility in correlation computation.

Regardless of anchor measure type (retrospective or concurrent), more measurement error is still possible in practice. This would not only undermine the responder classification but also attenuate the responsiveness correlation [25]. Hence, a correlation corrected for measurement error [25] and sensitivity analyses on the responder classification at different confidence limits of the anchor should be considered in these situations.

Finally, due to computational limitations, the current study did not model the relationship between the baseline score and follow-up change in the target COA and did not allow for varying true thresholds or responder percentages conditioned on baseline scores. These knowledge gaps can be addressed by future studies to facilitate discussions about how to thoughtfully estimate responder thresholds under different clinical data characteristics.

Declarations

Funding: The study was supported by RTI Health Solutions.

Conflicts of interest/Competing interests: Shanshan Qin, Lauren Nelson, Nicole Williams, Valerie Williams, Randall Bender, and Lori McLeod are researchers employed by RTI Health Solutions, which provides clinical outcome assessment development and psychometric evaluation support for pharmaceutical companies.

Availability of data and material: Due the extremely large sizes, the simulated datasets are available upon request.

Ethical Approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Acknowledgments:

The authors sincerely thank Theresa Coles of Duke University’s Department of Population Health Sciences in the early simulation design, and Caitlyn Matuska and John Forbes of RTI Health Solutions for their editorial reviews on this paper.

References

1. S. Department of Health and Human Services Food and Drug Administration (2009). *Guidance for industry: Patient-reported outcome measures: use in medical product development to support labeling claims.* Retrieved March 21, 2021, from https://www.fda.gov/media/77832/download
2. Food and Drug Administration. (2018). Patient-focused drug development guidance public workshop: methods to identify what is important to patients & select, develop or modify fit-for-purpose clinical outcomes assessments. Guidance 3 discussion document. Retrieved March 21, 2021, from https://www.fda.gov/media/116277/download
3. McLeod, L. D., Coon, C. D., Martin, S. A., Fehnel, S. E., & Hays R. D. (2011). Interpreting patient-reported outcome results: US FDA guidance and emerging methods. *Expert Review of Pharmacoeconomics & Outcomes Research,* 11(2), 163–169. doi: 10.1586/erp.11.12.
4. Coon, C. D. (2016). Telling the interpretation story: The case for strong anchors and multiple methods. Plenary presentation at the ISOQOL 23rd Annual Conference; October 19, 2016. Copenhagen, Denmark.

5. 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. doi: 10.1007/s11136-017-1616-3. Epub 2017 Jun 15.
6. Terluin, B., Eekhout, I., & Terwee, C. (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. doi: 10.1016/j.jclinepi.2016.12.015.
7. Norman, G. R., Sloan, J.A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality-of-life: The remarkable universality of half a standard deviation. *Medical Care*, 4, 582–592. doi: 10.1097/01.MLR.0000062554.74615.4C.
8. Wyrwich, K. W., Tierney, W. M., & Wolinsky, F.D. (1999). Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *Journal of Clinical Epidemiology*, 52, 861–873. doi: 10.1016/s0895-4356(99)00071-2.
9. 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. doi: 10.1177/0163278705275339.
10. Crosby, R.D., Kolotkin, R. L., & Williams, G. R. (2003). Defining clinically meaningful change in health-related quality of life. *Journal of Clinical Epidemiology,* 56(5), 395–407. doi: 10.1016/s0895-4356(03)00044-1.
11. 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. doi: 10.1016/j.jclinepi.2007.03.012.
12. Hays, R.D., Farivar, S.S., & Liu, H. (2005). Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 2, 63–67. doi: 10.1081/copd-200050663.
13. de Vet, H.C., Terluin, B., Knol, D. L., Roorda, L.D., Mokkink, L. B., Ostelo, R. W., et al. (2010). Three ways to quantify uncertainty in individually applied "minimally important change" values. *Journal of Clinical Epidemiology,* 63(1), 37–45. doi: 10.1016/j.jclinepi.2009.03.011.
14. Gershon, R., Cella, D., Rothrock, N., Hanrahan, R.T., & Bass, M. (2010). The use of PROMIS and assessment center to deliver patient-reported outcome measures in clinical research. *Journal of Applied Measurement*, 11(3), 304314.
15. Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., et al. (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. *Journal of Clinical Epidemiology,* 63(11):1179–1194. doi: 10.1016/j.jclinepi.2010.04.011.
16. Iman, R. L., & Conover W.-J. (1982). A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics – Simulation and Computation,* 11(3), 311–334. doi: 10.1080/03610918208812265
17. Wicklin, R. (2013). *Simulating Data with SAS®*. SAS Institute, Inc.
18. Coles, T.M., Chen, W., Nelson, L. M., Williams, V. S., Williams, N.J., & McLeod L.D. Current sample size practices in the psychometric evaluation of patient-reported outcome measures for use in clinical trials. Poster presented at the 2014 ISPOR 17th Annual European Congress; November 2014. Amsterdam.
19. SAS Institute, Inc. (2012). *SAS proprietary software, version 9.4.* SAS Institute, Inc.
20. Froud, R., & Abel, G. (2014). Using ROC curves to choose minimally important change thresholds when sensitivity and specificity are valued equally: the forgotten lesson of Pythagoras. Theoretical considerations and an example application of change in health status. *PLoS One*, 9(12), e114468. doi: 10.1371/journal.pone.0114468.
21. Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.
22. Mehta, L., McNeill, M., Hobart, J., Wyrwich, K. W., Poon, J. L., Auguste, P., et al. (2015). Identifying an important change estimate for the Multiple Sclerosis Walking Scale-12 (MSWS-12v1) for interpreting clinical trial results. *Multiple sclerosis journal - Experimental, Translational, and Clinical,* 1, 2055217315596993
23. Symonds, T., Spino, C., Sisson, M., Soni, P., Martin, M., Gunter, L., et al. (2007) Methods to determine the minimum important difference for a sexual event diary used by postmenopausal women with hypoactive sexual desire disorder. *Journal of Sexual Medicine*, 4(5), 1328–1335.
24. 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.
25. Hopkins C, Gillett S, Slack R, Lund VJ, & Browne JP. (2009). Psychometric validity of the 22-item sinonasal outcome test. *Clinical Otolaryngology*, 34, 447–454.
26. Spearman, C. (1904) The proof and measurement of association between two things. *American Journal of Psychology*, 15 (1), 72–101.

Tables

Table 1. Simulation parameters of clinical trial characteristics for outcome response generation

Simulation Conditions	Parameters
PROMIS T-score change	
Population marginal distribution of the PROMIS PF SF 20a T-score change Change range: -51 to 51 (the T-score in a scale of 11.7 to 62.7)	1. Restricted Normal (7, 3.5) 2. Restricted Normal (7, 7) 3. A skewed distribution from 16 − Gamma (shape = 1.5, scale = 6), with mean of 7, SD of 7.3, skewness = − 1.6
Anchor measure distribution	
Population marginal distribution of the PGIC 1 = very much worse 2 = much worse 3 = minimally worse 4 = no change 5 = minimally improved (Anchor) 6 = much improved 7 = very much improved	Multinomial probabilities: 1. 30% improvement (Very much worse = 0.1, much worse = 0.15, minimally worse = 0.2, no change = 0.25, minimally improved = 0.2, much improved = 0.1, very much improved = 0)[a] 2. 50% improvement (Very much worse = 0.05, much worse = 0.1, minimally worse = 0.15, no change = 0.2, minimally improved = 0.2, much improved = 0.15, very much improved = 0.15) [a] 3. 70% improvement (Very much worse = 0.05, much worse = 0.05, minimally worse = 0.1, no change = 0.1, minimally improved = 0.2, much improved = 0.3, very much improved = 0.2) [a]
Strength of correlation	
Population responsiveness correlation between T-score change and PGIC	1. ρ = 0.1 2. ρ = 0.3 3. ρ = 0.5 4. ρ = 0.7
Sample size	
Study sample size	1. n = 50 2. n = 100 3. n = 300
Total settings	3 × 3 × 4 × 3 = 108 combinations/cases

PF physical function, *PGIC* patient global impression of change, *PROMIS* Patient-Reported Outcomes Measurement Information System, *SF* short form

Information System; SD, standard deviation; SF, short form.

Note: Except sample size, the sample statistics of population parameters were subject to random sampling fluctuation. To maintain the different levels of percentage improvement (based on PGIC) and responsiveness correlation, each simulated data set retained for analysis was required to have absolute difference of sample – population responsiveness correlation < 0.065, and an absolute difference of sample – population improvement ≤ 10%.

[a] Bold text indicates improvement.

Table 2. Anchor- and distribution-based estimation methods

Method	Estimation
Anchor-based	
Mean method	Arithmetic mean of the changes in the PROMIS T-scores for the subjects with “minimally improved” PGIC values
Median method	Median value of the changes in the PROMIS T-scores for the subjects with “minimally improved” PGIC values
ROC method	Observed PROMIS T-score change that minimized the sum of $(1 - \text{sensitivity})^2$ and $(1 - \text{specificity})^2$ in the receiver operator characteristic curve of the logistic regression to predict PGIC responder classification of at least minimally improved
Predictive model-based	Estimated PROMIS T-score change associated with the post-odds equal to the pre-odds of the logistic regression to predict PGIC responder classification of at least minimally improved
Distribution-based	
Half SD	One-half of the standard deviation of the baseline PROMIS T-scores
SEM	Mean of the IRT-based standard errors of the baseline PROMIS T-scores

PGIC patient global impression of change, PROMIS Patient-Reported Outcomes Measurement Information System, ROC receiver operator characteristic, SD standard deviation, SEM standard error of measurement

Table 3. Performance of threshold estimation methods over 1,000 replications in comparison with true thresholds

			Anchor-Based														Distribution-Base
			Mean Method			Median Method		ROC Curve			Predictive Model			Half SD			S
Distribution	ρ	n	RB	CV	rRMSE	RB	CV	rRMSE	RB	CV	rRMSE	RB	CV	rRMSE	RB	CV	rF
Normal (7, 3.5)																	
30%	0.10	50	-16.6	14.0	20.3	-16.4	16.8	21.5	-17.3	16.5	22.0	-19.0	7.0	19.9	-71.7	10.1	7
		100	-17.0	10.2	19.0	-17.1	12.5	20.0	-17.6	12.8	20.6	-19.1	5.1	19.5	-71.6	7.3	7
		300	-16.9	5.6	17.6	-16.8	7.0	17.8	-18.1	8.7	19.4	-19.1	2.9	19.2	-71.5	4.1	7
	0.30	50	-9.0	12.5	14.5	-8.4	14.6	15.8	-13.4	13.7	17.9	-16.2	6.8	17.1	-71.7	10.1	7
		100	-9.7	9.4	12.9	-9.7	11.4	14.2	-14.9	10.9	17.5	-16.4	5.0	16.9	-71.7	7.3	7
		300	-9.9	5.1	10.9	-9.9	6.3	11.4	-15.5	7.5	16.8	-16.4	2.8	16.6	-71.5	4.0	7
	0.50	50	-1.7	11.1	11.0	-1.8	12.9	12.8	-10.7	12.2	15.3	-13.4	6.8	14.6	-71.7	10.0	7
		100	-2.6	8.4	8.6	-3.0	10.0	10.2	-11.9	9.4	14.5	-13.6	5.0	14.2	-71.6	7.3	7
		300	-2.9	4.6	5.3	-3.0	5.6	6.2	-12.8	6.1	13.8	-13.6	2.8	13.8	-71.5	4.1	7
	0.70	50	5.2	9.8	11.5	4.5	11.4	12.8	-8.0	11.1	13.0	-10.2	6.9	11.9	-71.7	10.1	7
		100	4.5	7.4	9.0	4.1	8.6	9.8	-9.0	8.3	11.8	-10.4	5.0	11.3	-71.6	7.4	7
		300	4.0	4.1	5.8	3.7	4.8	6.3	-9.7	5.3	10.8	-10.5	2.9	10.8	-71.5	4.0	7
50%	0.10	50	1.3	16.6	16.8	0.9	19.7	19.9	0.3	15.5	15.6	0.0	6.9	6.9	-64.5	10.0	6
		100	1.7	10.9	11.2	1.2	13.9	14.1	0.0	13.2	13.2	0.1	4.9	4.9	-64.3	7.3	6
		300	1.3	6.4	6.7	1.3	8.0	8.2	0.6	8.4	8.5	0.1	2.8	2.8	-64.2	4.1	6
	0.30	50	4.3	15.4	16.6	3.9	18.1	19.2	1.6	13.6	13.9	0.1	6.7	6.7	-64.5	10.0	6
		100	4.5	10.2	11.5	4.1	12.7	13.9	1.3	10.7	10.9	0.1	4.8	4.8	-64.4	7.3	6
		300	3.9	6.1	7.4	3.8	7.5	8.6	0.8	7.1	7.2	0.1	2.8	2.8	-64.2	4.1	6
	0.50	50	7.1	14.0	16.6	6.7	16.1	18.5	1.3	11.9	12.1	-0.1	6.7	6.7	-64.5	10.1	6
		100	7.1	9.3	12.2	6.6	11.5	13.9	0.7	9.4	9.5	-0.1	4.8	4.8	-64.3	7.4	6
		300	6.6	5.5	8.8	6.5	6.7	9.6	0.8	6.2	6.3	0.2	2.8	2.8	-64.2	4.1	6
	0.70	50	9.5	12.3	16.4	9.3	13.7	17.6	1.3	11.2	11.4	-0.1	7.0	7.0	-64.5	10.2	6
		100	9.4	8.0	12.8	9.1	9.5	13.8	0.8	8.7	8.8	-0.1	5.0	5.0	-64.3	7.3	6
		300	9.2	4.8	10.6	9.0	5.6	10.9	0.8	5.7	5.8	0.2	2.9	2.9	-64.2	4.1	6
70%	0.10	50	31.8	16.8	38.8	32.0	20.0	41.5	32.3	16.7	39.1	32.7	7.3	34.1	-52.3	9.9	5
		100	32.4	11.3	35.7	32.1	14.7	37.5	32.2	13.2	36.6	32.8	5.2	33.6	-52.1	7.4	5
		300	32.9	6.6	34.0	32.7	8.2	34.4	32.4	9.2	34.6	32.7	3.0	33.0	-51.8	4.1	5
	0.30	50	28.4	16.5	35.4	28.6	19.2	37.8	28.0	14.8	33.8	28.4	7.5	30.0	-52.3	9.9	5
		100	28.6	11.5	32.2	28.2	14.8	33.9	27.5	11.8	31.3	28.2	5.4	29.1	-52.1	7.4	5
		300	29.3	6.5	30.5	29.3	8.0	31.1	28.3	8.0	30.2	28.2	3.2	28.5	-51.8	4.1	5
	0.50	50	24.2	16.1	31.5	24.7	18.1	33.5	23.1	14.0	28.8	23.2	8.0	25.3	-52.2	10.0	5
		100	24.4	11.2	28.1	24.4	13.8	29.9	23.5	11.0	27.1	23.3	5.8	24.4	-52.0	7.4	5
		300	25.9	6.3	27.0	26.1	7.6	27.8	24.3	7.3	26.0	23.5	3.5	23.9	-51.8	4.1	5
	0.70	50	20.4	14.6	26.9	21.0	16.5	29.0	18.5	13.6	24.5	18.0	9.0	20.9	-52.3	9.9	5
		100	21.1	10.1	24.4	21.3	11.9	25.7	18.2	10.5	22.1	18.1	6.4	19.6	-52.0	7.4	5
		300	22.4	5.7	23.4	22.6	6.8	24.1	18.6	6.9	20.4	18.3	3.9	18.9	-51.8	4.1	5

CV coefficient of variation, *RB* relative bias, *ROC* receiver operator characteristic, *rRMSE* relative root mean squared error, *SD* standard deviation, *SEM* standard

Note: The lowest absolute (best) values of the three performance statistics are presented in bold font and shaded in each row (simulation setting) within the the lowest (best) absolute value across all six methods is based on a distribution-based method, the value is presented in bold font without shading.

			Anchor-Based														Distribution-Base
Normal (7, 7.0)																	
30%	0.10	50	-28.5	26.6	34.3	-28.0	31.9	36.2	-29.1	30.9	36.4	-32.4	13.8	33.7	-76.7	10.1	7
		100	-29.2	19.4	32.3	-29.1	24.2	33.8	-29.9	23.8	34.3	-32.5	9.9	33.2	-76.7	7.3	7
		300	-29.0	10.8	30.0	-28.6	13.4	30.2	-31.3	16.8	33.3	-32.5	5.7	32.7	-76.6	4.1	7
	0.30	50	-16.2	22.3	24.7	-14.9	26.1	26.8	-23.6	25.8	30.7	-27.7	12.8	29.3	-76.7	10.0	7
		100	-17.3	16.6	22.1	-17.1	20.3	24.0	-25.8	20.1	29.8	-28.1	9.5	28.9	-76.7	7.3	7
		300	-17.5	9.0	19.0	-17.2	11.2	19.5	-26.5	14.4	28.5	-28.1	5.3	28.3	-76.6	4.1	7
	0.50	50	-4.2	18.6	18.3	-4.1	21.7	21.2	-19.2	22.5	26.4	-23.1	12.5	25.0	-76.7	10.0	7
		100	-5.5	14.1	14.4	-5.7	16.9	16.9	-20.9	17.0	24.8	-23.4	9.2	24.4	-76.7	7.3	7
		300	-6.0	7.6	9.4	-5.9	9.4	10.6	-22.1	11.3	23.8	-23.5	5.2	23.8	-76.6	4.1	7
	0.70	50	7.3	15.7	18.3	6.5	18.6	20.8	-14.0	19.6	21.9	-17.8	12.5	20.5	-76.7	10.1	7
		100	6.1	11.9	14.0	5.8	13.8	15.7	-16.2	14.7	20.4	-18.2	9.0	19.6	-76.7	7.3	7
		300	5.4	6.5	8.7	5.2	7.7	9.7	-17.2	9.5	18.9	-18.3	5.2	18.8	-76.6	4.1	7
50%	0.10	50	2.0	32.8	33.5	1.9	39.0	39.8	1.0	30.3	30.6	-0.1	13.7	13.7	-64.5	10.1	6
		100	2.8	21.4	22.2	2.3	27.4	28.2	0.4	26.1	26.3	-0.1	9.7	9.7	-64.3	7.3	6
		300	2.3	12.7	13.2	2.6	15.7	16.3	0.8	16.6	16.8	-0.0	5.6	5.6	-64.2	4.1	6
	0.30	50	7.7	29.8	33.1	7.5	35.6	39.0	2.7	26.8	27.6	0.0	13.3	13.3	-64.5	10.1	6
		100	8.6	19.4	22.8	8.4	24.4	27.8	1.9	21.2	21.7	-0.1	9.6	9.6	-64.4	7.3	6
		300	7.5	11.7	14.6	7.6	14.4	17.3	0.8	14.2	14.4	0.0	5.5	5.5	-64.2	4.1	6
	0.50	50	14.1	26.1	32.9	13.5	30.6	37.3	1.6	23.9	24.3	-0.3	13.3	13.3	-64.5	10.2	6
		100	13.9	17.4	24.2	13.2	21.7	27.9	1.2	19.2	19.5	-0.3	9.7	9.6	-64.3	7.4	6
		300	12.9	10.3	17.4	13.0	12.5	19.2	1.4	12.3	12.6	0.1	5.6	5.6	-64.2	4.1	6
	0.70	50	18.7	22.5	32.6	18.5	24.8	34.8	2.5	21.9	22.6	-0.2	14.1	14.1	-64.5	10.1	6
		100	18.7	14.6	25.4	18.2	17.4	27.5	0.7	17.0	17.1	-0.3	10.0	10.0	-64.3	7.3	6
		300	18.3	8.8	21.0	18.2	10.3	21.8	1.3	11.1	11.3	0.2	5.8	5.8	-64.2	4.1	6
70%	0.10	50	103.4	34.1	124.5	104.5	40.2	133.0	102.2	33.8	122.9	105.3	14.8	109.6	-24.9	10.0	2
		100	104.9	23.1	115.0	104.5	30.1	121.4	105.2	26.4	118.3	105.9	10.5	108.1	-24.4	7.3	2
		300	105.9	13.4	109.4	105.9	16.8	111.4	103.6	18.8	110.4	105.6	6.2	106.3	-24.1	4.1	2
	0.30	50	92.1	35.0	114.0	93.6	40.4	122.0	90.6	31.5	108.7	92.3	15.6	97.1	-24.9	9.9	2
		100	92.9	23.8	103.6	92.6	30.9	110.1	88.6	24.9	100.3	91.4	11.4	93.9	-24.6	7.5	2
		300	95.1	13.6	98.7	95.3	16.9	100.8	90.5	17.1	96.2	91.5	6.7	92.4	-24.1	4.1	2
	0.50	50	79.2	35.1	101.1	80.8	39.5	107.8	74.3	31.4	92.3	76.1	17.7	82.2	-24.7	10.0	2
		100	80.2	24.3	91.4	80.5	30.2	97.2	75.7	24.4	87.0	76.0	12.7	79.2	-24.4	7.5	2
		300	84.2	13.6	87.8	85.2	16.4	90.4	77.9	15.8	82.8	76.9	7.6	78.1	-24.1	4.1	2
	0.70	50	67.8	33.0	87.6	70.0	36.8	93.9	60.2	31.3	78.4	59.7	21.0	68.4	-24.8	9.9	2
		100	70.0	22.7	79.9	70.7	26.8	84.2	60.6	24.4	72.1	60.0	15.0	64.6	-24.4	7.4	2
		300	73.6	12.8	76.9	74.5	15.2	79.1	60.9	16.2	66.2	60.7	9.1	62.5	-24.1	4.1	2
Negative Gamma																	
30%	0.10	50	-33.7	27.6	38.3	-20.5	24.4	28.2	-23.2	24.2	29.7	-36.8	14.8	38.0	-78.8	10.1	7
		100	-34.8	19.5	37.1	-20.3	17.1	24.4	-23.0	19.4	27.5	-37.7	10.7	38.2	-78.7	7.2	7

CV coefficient of variation, *RB* relative bias, *ROC* receiver operator characteristic, *rRMSE* relative root mean squared error, *SD* standard deviation, *SEM* standard

Note: The lowest absolute (best) values of the three performance statistics are presented in bold font and shaded in each row (simulation setting) within the the lowest (best) absolute value across all six methods is based on a distribution-based method, the value is presented in bold font without shading.

			Anchor-Based												Distribution-Base		
		300	-35.3	11.3	36.1	-19.9	10.1	21.5	-21.9	12.5	23.9	-37.9	6.3	38.1	-78.6	4.2	7
	0.30	50	-22.7	19.9	27.4	-11.3	18.4	19.8	-17.0	17.9	22.6	-31.6	13.2	32.9	-78.7	10.5	7
		100	-24.4	14.7	26.8	-11.4	13.2	16.3	-17.9	14.1	21.3	-32.6	9.8	33.2	-78.7	7.2	7
		300	-25.0	8.6	25.8	-11.4	7.8	13.3	-18.2	9.5	19.8	-32.8	5.9	33.1	-78.6	4.2	7
	0.50	50	-12.0	13.8	17.1	-4.0	14.4	14.4	-13.6	14.6	18.5	-25.1	11.9	26.6	-78.7	10.4	7
		100	-13.5	10.6	16.3	-3.8	10.4	10.7	-14.6	12.1	17.8	-26.0	8.9	26.8	-78.7	7.1	7
		300	-14.4	6.4	15.4	-3.7	5.9	6.8	-14.5	8.0	16.0	-26.4	5.4	26.7	-78.6	4.2	7
	0.70	50	-2.4	10.1	10.1	2.7	10.8	11.4	-9.2	12.5	14.7	-17.5	10.6	19.6	-78.7	10.4	7
		100	-3.4	7.5	8.0	2.9	7.9	8.6	-10.1	10.4	13.7	-18.3	8.1	19.4	-78.7	7.1	7
		300	-4.1	4.6	6.0	2.9	4.5	5.5	-11.0	6.6	12.5	-18.8	4.8	19.2	-78.6	4.2	7
50%	0.10	50	-18.6	33.2	32.8	-1.1	28.8	28.5	-3.2	24.9	24.3	-19.8	14.5	23.0	-72.1	10.4	7
		100	-18.6	22.7	26.2	0.6	19.9	20.0	-2.1	19.2	18.9	-20.4	10.5	22.1	-72.0	7.2	7
		300	-18.7	12.8	21.4	1.8	10.9	11.2	-0.1	12.0	12.0	-20.7	6.2	21.3	-71.9	4.2	7
	0.30	50	-13.8	30.1	29.4	3.1	25.7	26.7	0.3	20.2	20.3	-18.1	14.2	21.5	-72.0	10.4	7
		100	-13.4	20.4	22.2	4.1	18.0	19.2	-0.1	15.5	15.5	-18.8	10.6	20.7	-72.0	7.2	7
		300	-13.4	11.5	16.7	5.4	10.0	11.8	-0.0	10.6	10.6	-19.0	6.3	19.7	-71.9	4.2	7
	0.50	50	-6.6	24.4	23.8	7.1	22.0	24.6	0.4	19.0	19.1	-15.0	14.5	19.4	-72.1	10.3	7
		100	-6.8	17.2	17.4	7.8	15.7	18.6	-0.2	14.7	14.7	-15.6	10.7	18.0	-71.9	7.2	7
		300	-6.6	9.7	11.2	8.9	8.7	13.0	0.3	9.6	9.6	-15.7	6.4	16.6	-71.9	4.2	7
	0.70	50	2.2	18.5	19.0	11.3	17.8	22.8	1.2	16.9	17.1	-10.2	14.3	16.4	-72.1	10.3	7
		100	1.5	13.6	13.9	11.2	13.1	18.4	0.5	13.4	13.4	-10.6	10.7	14.3	-72.0	7.2	7
		300	1.6	7.7	8.0	12.2	7.4	14.8	0.5	8.9	8.9	-10.8	6.4	12.2	-71.9	4.2	7
70%	0.10	50	37.0	36.8	62.5	69.5	32.4	88.6	64.8	26.0	77.7	38.4	15.5	44.0	-50.4	10.2	5
		100	37.9	25.3	51.6	72.3	21.7	81.4	67.1	21.6	76.2	36.9	11.4	40.1	-50.1	7.1	5
		300	37.9	13.7	42.4	74.5	12.3	77.5	71.3	13.9	75.2	36.8	6.7	37.9	-50.0	4.2	5
	0.30	50	30.1	38.8	58.8	60.6	34.1	81.7	61.0	24.2	72.4	31.8	16.8	38.7	-50.2	10.4	5
		100	31.8	27.0	47.7	63.9	23.3	74.4	61.9	19.5	69.5	30.6	12.5	34.8	-50.1	7.2	5
		300	33.3	14.3	38.4	67.8	12.8	71.1	63.4	13.2	67.0	30.5	7.4	32.0	-50.0	4.2	5
	0.50	50	26.9	38.8	56.1	54.0	33.8	75.0	54.1	23.6	65.2	27.5	18.6	36.3	-50.3	10.2	5
		100	28.4	26.9	44.7	56.7	23.8	67.9	53.8	19.3	61.5	26.1	14.0	31.5	-50.1	7.1	5
		300	31.5	14.2	36.6	61.0	13.0	64.5	55.6	13.0	59.1	26.2	8.4	28.2	-50.0	4.2	5
	0.70	50	29.2	35.2	54.1	49.1	32.0	68.5	44.0	25.6	57.4	24.4	21.2	36.0	-50.3	10.4	5
		100	31.3	24.0	44.4	51.7	21.9	61.4	43.9	19.7	52.3	23.7	15.6	30.5	-50.1	7.2	5
		300	32.9	13.0	37.1	53.7	12.4	57.0	43.6	13.2	47.5	23.6	9.4	26.3	-50.0	4.2	5

CV coefficient of variation, *RB* relative bias, *ROC* receiver operator characteristic, *rRMSE* relative root mean squared error, *SD* standard deviation, *SEM* standard

Note: The lowest absolute (best) values of the three performance statistics are presented in bold font and shaded in each row (simulation setting) within the the lowest (best) absolute value across all six methods is based on a distribution-based method, the value is presented in bold font without shading.

Table 4. Distributions of within-sample percentages of subjects with individual reliable changes ≤ estimated or true thresholds

			Minimum, Median, Maximum of Within-sample Percentages of Subjects With Individual RCs ≤ Estimated or True Thresholds				
Distribution	ρ	n	Mean Method	Median Method	ROC Curve	Predictive Model	True Threshold
Normal (7, 3.5)							
30%	0.10	50	0.0, 100, 100	0.0, 100, 100	0.0, 100, 100	48.0, 100, 100	98.0, 100, 100
		100	1.0, 99, 100	0.0, 99, 100	0.0, 99, 100	92.0, 99, 100	98.0, 100, 100
		300	95.0, 99, 100	10.3, 99, 100	6.7, 99, 100	97.0, 99, 100	99.0, 100, 100
	0.30	50	0.0, 100, 100	0.0, 100, 100	0.0, 100, 100	92.0, 100, 100	98.0, 100, 100
		100	1.0, 100, 100	1.0, 100, 100	3.0, 100, 100	95.0, 100, 100	98.0, 100, 100
		300	98.0, 100, 100	97.0, 100, 100	95.0, 99, 100	97.7, 99, 100	99.0, 100, 100
70%	0.10	50	0.0, 98, 100	0.0, 98, 100	0.0, 98, 100	12.0, 100, 100	0.0, 0, 8
		100	0.0, 99, 100	0.0, 99, 100	0.0, 99, 100	87.0, 99, 100	0.0, 1, 6
		300	80.0, 99, 100	1.3, 99, 100	0.7, 99, 100	95.0, 99, 100	0.0, 1, 4
	0.30	50	0.0, 98, 100	0.0, 98, 100	0.0, 98, 100	0.0, 98, 100	0.0, 0, 8
		100	0.0, 98, 100	0.0, 98, 100	0.0, 98, 100	83.0, 98, 100	0.0, 1, 6
		300	42.7, 98, 100	2.7, 98, 100	0.0, 98, 100	93.0, 98, 100	0.0, 1, 4
Negative Gamma							
30%	0.10	50	0.0, 100, 100	0.0, 100, 100	0.0, 100, 100	0.0, 100, 100	96.0, 100, 100
		100	0.0, 99, 100	0.0, 100, 100	1.0, 100, 100	3.0, 99, 100	98.0, 100, 100
		300	1.3, 99, 100	68.7, 100, 100	10.7, 100, 100	71.7, 98, 100	99.3, 100, 100
	0.30	50	0.0, 100, 100	0.0, 100, 100	0.0, 100, 100	0.0, 100, 100	98.0, 100, 100
		100	1.0, 100, 100	1.0, 100, 100	4.0, 100, 100	16.0, 100, 100	98.0, 100, 100
		300	93.0, 100, 100	97.7, 100, 100	93.0, 100, 100	94.0, 99, 100	99.3, 100, 100
70%	0.10	50	0.0, 98, 100	0.0, 100, 100	0.0, 100, 100	0.0, 98, 100	0.0, 2, 12
		100	0.0, 98, 100	0.0, 100, 100	0.0, 100, 100	1.0, 97, 100	0.0, 3, 10
		300	0.3, 98, 100	0.7, 100, 100	1.0, 100, 100	7.7, 97, 100	0.7, 3, 6
	0.30	50	0.0, 96, 100	0.0, 100, 100	0.0, 100, 100	0.0, 96, 100	0.0, 2, 12
		100	0.0, 97, 100	0.0, 100, 100	0.0, 100, 100	0.0, 96, 100	0.0, 3, 10
		300	0.3, 97, 100	0.3, 100, 100	1.0, 100, 100	0.7, 95, 100	0.7, 3, 6

RC Reliable Change, *ROC* receiver operator characteristic
Note: Results related to the second normal distribution, ρ = 0.5 and 0.70, and 50% improvement were omitted due to the similarity with the first normal distribution and corresponding lower- level conditions.

Figures

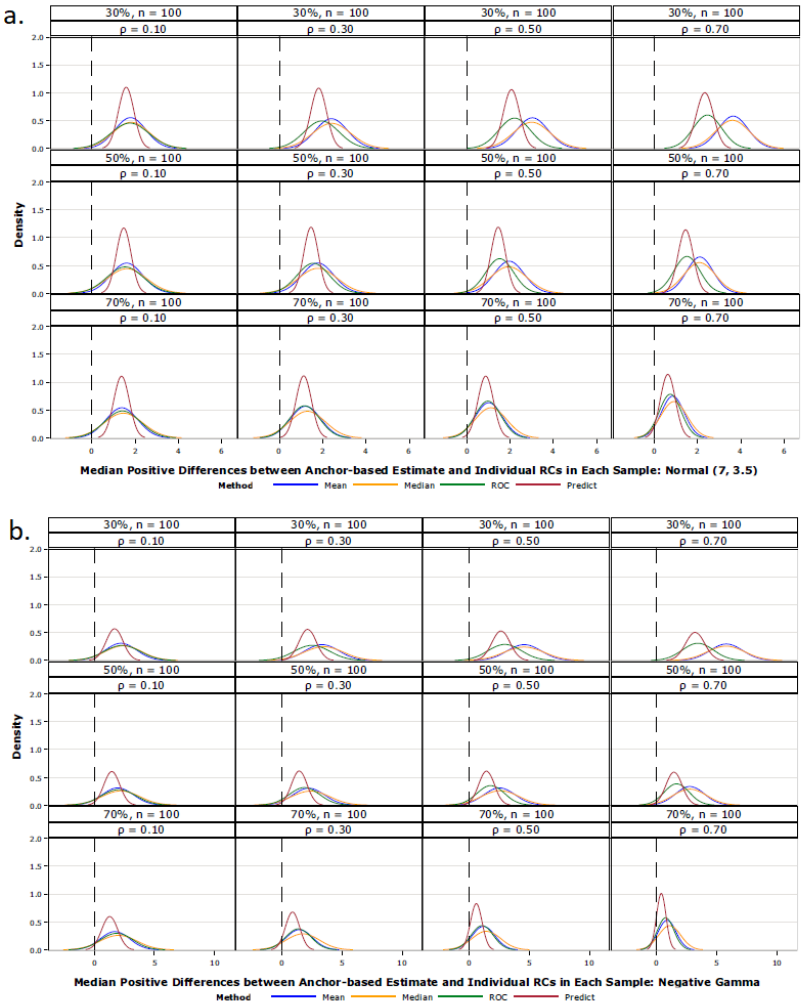

Figure 1

Probability Density Function Plots of Median Positive Differences between Anchor-based Estimate and Individual RCs by Anchor-based Methods a. Normal (7, 3.5), n = 100 b. Negative Gamma, n = 100

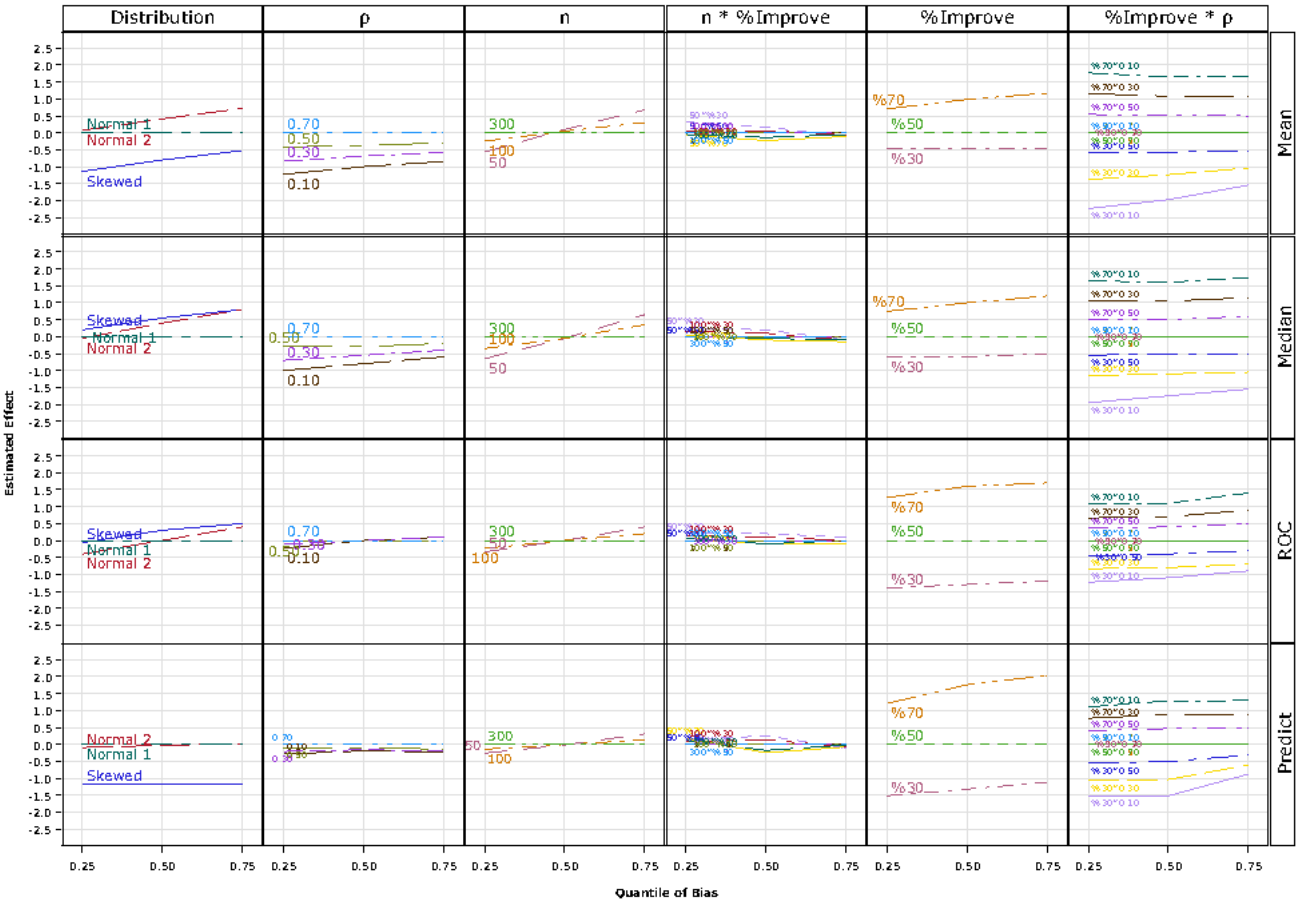

Figure 2

Paneled Effect Plots of Significant Predictors of Estimation Bias by Anchor-Based Methods

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalFig1a1.tiff
- SupplementalFig1a2.tiff
- SupplementalFig1a3.tiff
- SupplementalFig1b1.tiff
- SupplementalFig1b2.tiff
- SupplementalFig1b3.tiff
- SupplementalFig1c1.tiff
- SupplementalFig1c2.tiff
- SupplementalFig1c3.tiff
- SupplementaryTables30March2021.docx