Minimally Important Differences and Severity Thresholds are Estimated for the PROMIS Depression Scales from Three Randomized Clinical Trials

Kurt Kroenke, MD[a,b], **Timothy E. Stump, MS**[c], **Chen X. Chen, PhD**[d], **Jacob Kean, PhD**[e], **Matthew J. Bair, MD**[a,b,f], **Teresa M. Damush, PhD**[a,b,f], **Erin E. Krebs, MD**[g,h], **Patrick O. Monahan, PhD**[c]

[a]Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, USA

[b]Regenstrief Institute, Inc., Indianapolis, IN, USA

[c]Department of Biostatistics, Indiana University Fairbanks School of Public Health and School of Medicine, Indianapolis, IN, USA

[d]Indiana University School of Nursing, Indianapolis, IN, USA

[e]Department of Population Health Sciences, University of Utah School of Medicine, Salt Lake City, Utah, USA

[f]VA Health Services Research and Development Center for Health Information and Communication, Indianapolis, IN, USA

gCenter for Chronic Disease Outcomes Research, Minneapolis VA Health Care System, Minnesota, USA

hUniversity of Minnesota Medical School, Minneapolis, Minnesota, USA

Abstract

Background—Patient Reported Outcomes Measurement Information Systems (PROMIS) scales are increasingly being used to measure symptoms in research and practice. The purpose of this study was to determine the minimally important difference (MID) and severity thresholds (cut-points) for the four fixed-length PROMIS depression scales.

Methods—The study sample was adult participants in three randomized clinical trials (N=651). MID was estimated using multiple distribution- and anchor-based approaches including assessing correspondence between Patient Health Questionnaire (PHQ-9) and PROMIS depression scores.

Results—The best MID estimate was a PROMIS depression T-score of 3.5 points with most methods producing an MID in the 3 to 4 point range across all three samples. MID estimates were similar for all four PROMIS scales. A PHQ-9 1-point change equated to a PROMIS 1.25-point T-score change. PROMIS T-scores of 55, 60, 65, and 70 appeared to be reasonable thresholds for mild, moderate, moderately severe, and severe depression, respectively.

Limitations—The study sample was predominantly male veterans with either chronic pain (2 trials) or previous stroke (1 trial). The severity of depression was mild to moderate.

Conclusion—A T-score of 3 to 4 points is a reasonable MID for PROMIS depression scales and can be used to assess treatment effects in both practice and research as well to calculate sample sizes for clinical trials. Severity cut-points can help interpret the meaning of scores and action thresholds for treatment decisions.

Keywords

1. INTRODUCTION

Depression is the most common mental health disorder(Kroenke and Unutzer, 2017) and the second leading cause of disability.(Collaborators, 2013; Vos et al., 2012) It causes early mortality through suicide, adversely impacts chronic medical disorders, and drives direct and indirect healthcare costs.(O'Connor et al., 2016) Routine screening for depression is an evidence-based recommendation.(Siu et al., 2016) Moreover, measurement-based care has been shown to improve depression outcomes.(Gaynes et al., 2008; Lewis et al., 2019)

Numerous validated measures are available to assess the presence and severity of depressive symptoms and disorders.(Hirschtritt and Kroenke, 2017; Wahl et al., 2014) The National Institutes of Health has sponsored development of the Patient-Reported Outcome Measurement Information System (PROMIS) scales that assess a number of symptoms including depression.(Cella et al., 2010) The PROMIS depression scales have demonstrated reliability and validity across diverse populations.(Amtmann et al., 2014; Choi et al., 2014;

Freedland et al., 2019; Gibbons et al., 2011; Jakob et al., 2017; Kaat et al., 2017; Levin et al., 2015; Pilkonis et al., 2011; Purvis et al., 2019; Schalet et al., 2016; Sunderland et al., 2018; Tang et al., 2019; Vilagut et al., 2015) However, enhanced interpretability is needed to support their usefulness in clinical practice as well as research. One essential component of interpretability is minimally important difference (MID), defined as "the smallest difference in score in the domain of interest that patients perceived as important, either beneficial or harmful, and that would lead the clinician to consider a change in the patient's management."(Guyatt et al., 2002) Alternative terms are "minimal clinically important difference" or "meaningful" difference. Another component of interpretability is a scale's severity cut-points that can help inform action thresholds for treatment decisions.

In this paper, we estimate MIDs for the four fixed-length PROMIS depression scales by analyzing data from three randomized clinical trials (RCTs). Importantly, we triangulate multiple methods for MID and cut-points estimation. Fixed-length scales were chosen rather than computer-adaptive testing (CAT) because in many clinical and research settings fixed-length scales are more feasible to administer and produce comparable results to CAT (Choi et al., 2010)

2. METHODS

2.1. Design and Participants

Data were analyzed from three RCTs conducted between 2012 and 2017 with 651 patients who had complete psychometric data. Sample 1 included 153 primary care patients participating in an RCT to compare the effectiveness of pharmacological versus cognitive-behavioral treatment for chronic low back pain. Sample 2 included 240 primary care patients participating in a pragmatic RCT comparing opioid therapy versus non-opioid medication therapy for chronic back pain or hip or knee osteoarthritis pain. Sample 3 included 258 stroke survivors participating in an RCT evaluating the efficacy of a stroke-self-management program. Samples 1 and 2 were enrolled from Veterans Administration (VA) primary care clinics, and Sample 3 comprised both Veteran and non-Veteran patients. Data were collected from baseline and follow-up interviews by trained research personnel. Follow-up assessments were conducted 6 months after baseline for Sample 1 and 3 months after baseline for Samples 2 and 3. The studies were approved by the Indiana University Institutional Review Board.

2.2. Measures

2.2.1. PROMIS Depression fixed-length scales—Participants completed the following four fixed-length PROMIS depression scales: the 8-item original short form (8b), and the 4-, 6-, and 8-item scales (4a, 6a, 8a) that are part of the PROMIS adult profile instruments (a collection of short forms containing a fixed number of items from key PROMIS domains).(Cella et al., 2015) The three profile scales are nested in that all items in the 4-item scale (4a) are included in the 6-item scale (6a) which in turn constitutes 6 of the items in the 8-item (8a) scale. The 8-item profile and 8-item short-form contain 7 items in common and 1 unique item each.

For each scale, respondents are asked how often in the past 7 days they have experienced specific depression symptoms, using a 5-point ordinal rating scale of "Never," "Rarely," "Sometimes," "Often," and "Always." Raw score totals are converted to an item response theory (IRT) based T-score for which higher scores represent greater depression severity. A T-score of 50 is the average for the US general population with a standard deviation (SD) of 10. Cronbach's alpha for baseline PROMIS depression raw scores in the three trials ranged from 0.89 to 0.95.

2.2.2.   The Patient Health Questionnaire 9-item Depression Scale (PHQ-9)— The PHQ-9 is among the best-validated and widely used depression scales in both clinical practice and research.(Hirschtritt and Kroenke, 2017; Kroenke et al., 2010; Mitchell et al., 2016) The PHQ-9 includes one item for each of the nine DSM-V criterion symptoms used in diagnosing major depression. Respondents are asked how much in the past 2 weeks they have been bothered by each symptom, with the response options being "Not at all", "Several days", "More than half the days", and "Nearly every day." Scores range from 0 to 27 with higher scores indicating greater depression severity. The Cronbach's alpha for baseline PHQ-9 scores in the three trials ranged from 0.76 to 0.85.

2.2.3.   Disability Days— A single item used in several previous studies(Kroenke et al., 2018; Kroenke et al., 2009) assessed the number of patient-reported disability days due to depression or pain: "During the past 4 weeks, how many days did you cut down on the things you usually do for one-half day or more because of problems with either pain or low mood?" The reason pain and depression were asked about in the same question is: a) two of the trials focused on chronic pain; b) pain and depression frequently co-occur; and c) it would be difficult for patients to recall and isolate the individual effects of pain and depression on disability.

2.2.4.   Cross-sectional Global Ratings of Depression— The cross-sectional global rating of depression assesses patient mood on average in the past 7 days. Following the approach developed by Yost et al.,(Yost et al., 2011) a 5-point ordinal scale ranging from 0 = "not unhappy or down at all" to 4 = "very severely unhappy or down" was used.(Yost et al., 2011)

2.2.5.   Retrospective Global Ratings Change (RGRC)— The RGRC assesses the overall clinical response as judged by the participant. At follow-up, participants rated the change in their depression compared to baseline. Response options ranged from −3 = "very much worse," to +3 = "very much better," with 0 representing no change (7 options in total). The RGRC is widely used as an outcome measure in depression trials (Fischer et al., 1999; Johns et al., 2013) and is commonly used to establish MIDs for patient-reported outcome scales.(Chen et al., 2018; Yost et al., 2011)

2.4.   Data Analysis

Data for each RCT were analyzed separately rather than pooled because the three trials involved different clinical populations, interventions, and follow-up time frames. We estimated the MIDs by triangulating distribution- and anchor-based methods.(Chen et al.,

2018; Guyatt et al., 2002; Yost and Eton, 2005; Yost et al., 2011) Distribution-based methods are based on the statistical distribution of the measures, while anchor-based methods are based on external criteria (anchors) that are clinically meaningful.(Revicki et al., 2008) Data were analyzed using SAS software (version 9.4, SAS Institute, Cary, NC).

2.4.1.   Distribution-based methods—For effect size, we calculated 0.2 SD, 0.35 SD, and 0.5 SD of baseline PROMIS Depression scores. Because 0.2 and 0.5 SD approximate small and moderate effect sizes(Kazis et al., 1989), they can be considered the lower and upper bounds of an MID, and an effect size between those boundaries (e.g., 0.35 SD) can be a good approximation of an MID.(Chen et al., 2018; Eton et al., 2004)

Standardized error of measurement (SEM) was calculated using baseline depression scores. (Wyrwich et al., 1999a; Wyrwich et al., 1999b) The PROMIS IRT calibrations estimate the standard error for each T-score. The SEM for each sample was obtained by averaging the individuals' standard errors across the sample.(Embretson, 1996; Hays et al., 2000) Specifically, the square root of the mean of variance (i.e., standard error squared) for each T-score across persons in the sample was computed to derive the sample SEMs. The literature suggests that 1 SEM corresponds closely with anchor-based MIDs for health-related quality of life measures.(Chen et al., 2018; Wyrwich et al., 1999a; Wyrwich et al., 1999b)

2.4.2.   Anchor-based methods—Anchor-based methods map PROMIS depression scores onto clinically meaningful anchors. To evaluate an anchor, we correlated the anchor score with the PROMIS depression score. A correlation   0.3 is one criterion sometimes used to support the use of an anchor.(Revicki et al., 2008; Yost et al., 2011) We performed both cross-sectional and longitudinal anchor-based analyses.(Chen et al., 2018; Eton et al., 2004; Yost et al., 2011)

2.4.2.1.   <u>Cross-sectional anchor-based analyses:</u> Cross-sectional analyses address minimally important between-individual differences. In these analyses, PROMIS depression scores within each time point were mapped onto two primary anchors and one secondary anchor.

PHQ-9 was used as one primary anchor, given its status as a legacy measure and its well-established MID. Correlations between PROMIS depression and PHQ-9 scores averaged 0.72 (median = .71). A 3-point difference on the PHQ-9 represents a reasonable MID with a 5-point difference being an upper bound.(Kroenke et al., 2016; Lowe et al., 2004) Using linear regression, we regressed the PROMIS depression scores on the PHQ-9 scores. The linearity assumption was confirmed by inspecting scatter plots. We estimated the PROMIS depression score that corresponded to a 3-point and a 5-point PHQ-9 change by multiplying the β coefficient for PHQ-9 from the regression model by 3 and 5, respectively.

Disability days in the past 4 weeks due to depression was used as the other primary anchor. Correlations between PROMIS depression scores and disability days ranged from 0.40 to 0.52. Participants were divided into 4 distinct severity categories: 0–7days, 8–14 days, 15–21 days, and 22–28 days.(Chen et al., 2018; Von Korff et al., 1992) These categories are less well-defined in the literature and were operationally chosen since they represent 0–25%, 26–

50%, 51–75%, and >75% of days in the past 4 weeks as disability days. The difference in PROMIS scores between adjacent categories was calculated.

Global rating of depression was used as a secondary anchor because it is less frequently cited in the MID literature. Participants were divided into 5 categories based on being not, mildly, moderately, severely, or very severely unhappy or down. Correlations between PROMIS depression scores and global rating of depression ranged from 0.58 to 0.79. The difference in PROMIS scores between adjacent global rating categories was calculated.

2.4.2.2.   Longitudinal anchor-based analyses: While cross-sectional analyses address minimally important between-individual differences, longitudinal analyses address minimally important change scores within individuals. In these analyses, changes in the PROMIS depression scores (from baseline to follow-up) were mapped onto global depression changes, which were determined both retrospectively and prospectively.(Chen et al., 2018; Yost et al., 2011)

The RGRC score collected at follow-up was used as the retrospective anchor. Participants were divided into 7 distinct severity categories based on RGRC: "much better," "moderately better," "a little better," "no change," "a little worse," "moderately worse," and "much worse." Correlations between PROMIS depression change scores and the RGRC ranged from −0.37 to −0.26 in Samples 1 and 2, and −0.08 to −0.05 in Sample 3. PROMIS depression change scores corresponding to a major category shift (i.e., same to better or same to worse) was the used as a primary method of estimating MID. A proximal category shift (e.g., between "no change" and "a little better," or between "a little better" and "moderately better") was used as secondary way to estimate MID because the number of patients in some categories were too small to provide reliable estimates.

The prospective change in global rating was calculated by subtracting an individual's follow-up global rating of depression from the baseline global rating.(Yost et al., 2011) Correlations between the PROMIS depression change scores and prospective change in global rating scores ranged from 0.31 to 0.52. Since the cross-sectional global rating of depression is score on a 5-point scale ranging from 0 ("Not unhappy or down at all") to 4 ("Very severely unhappy or down"), change scores had a possible range of −4 to +4, where negative numbers indicated worsening depression and positive numbers improved depression. For example, a patient who reported being "severely unhappy or down" at baseline and "mildly unhappy or down" at follow-up had a +2 change (3 *minus* 1), whereas a patient who reported being "moderately unhappy or down" at baseline and "severely unhappy or down" at follow-up had a −1 change (2 *minus* 3). PROMIS depression change scores corresponding to one major category shift (i.e., same [0] to better [+1 to +4 combined] or same [0] to worse [−1 to −4 combined]) was the used as a primary method of estimating MID. A proximal category shift (e.g., between 0 to +1 or between −1 and −2) was used as a secondary way to estimate MID.

2.4.3.   Estimating Severity Thresholds—PROMIS scale cut-points representing different levels of severity were estimated using two methods in our study (cross-sectional global rating of mood and PHQ-9 ordinal severity categories (Kroenke et al., 2010)) as well

as estimates from prior studies.(Amtmann et al., 2014; Cella et al., 2014; Choi et al., 2014; Gibbons et al., 2011; Pilkonis et al., 2014)

3.   RESULTS

3.1.   Sample Characteristics

For all three samples, participants were mostly male, non-Hispanic, white, married, and had some college education (Table 1). Mean PHQ-9 scores indicated that Sample 1 had moderate and Samples 2 and 3 had mild levels of depressive symptoms. The proportion of patients who met DSM-V criteria for major or minor depression using the PHQ-9 diagnostic algorithm(Kroenke et al., 2010) in the three samples was 58.1%, 24.6%, and 33.7%, respectively. The average proportion of patients across the 3 trials with 0–7, 8–14, 15–21, and 22–28 self-reported disability days was 57.9%, 17.6%, 13.7%, and 10.8%, respectively.

3.2.   Distribution-based Estimates

A 0.35 effect size and one SEM constituted our two primary distribution-based MID estimates. As shown in Table 2, these MID estimates for the PROMIS depression T-score in the three samples averaged 3.44 (range, 3.31 to 3.57) and 3.60 (range, 3.32 to 3.76), respectively.

3.3.   Anchor-based Estimates

3.3.1.   Cross-sectional anchor-based estimates—For the 24 models regressing PROMIS depression scores on PHQ-9 scores (2 scores [baseline and follow-up] for 4 PROMIS scales in 3 studies), each 1-point change in PHQ-9 score was associated with a mean PROMIS T-score change of 1.25 points (median = 1.23; interquartile range, 1.19 to 1.32). As shown in Table 2, a 3-point PHQ-9 difference (i.e. MID for PHQ-9) corresponded to an average difference in the PROMIS depression T-score of 3.76 (range, 3.42 to 4.29). Using a 1-category shift among the four ordinal categories of disability days, the average MID estimate was 3.70 (range, 2.70 to 4.97).

3.3.2.   Longitudinal anchor-based estimates—Prospective and retrospective global change in mood ratings (using the three categories of better, same, or worse) yielded MID estimates of 3.52 and 2.51, respectively (Table 2).

3.4.   Summary of MID Estimates across 6 Methods and 3 Samples

The six primary distribution- and anchor-based MID estimates are plotted in Figure 1. Of the 18 MID estimates from the three samples, 15 were narrowly clustered in the 3 to 4 point range, two were fairly close to this cluster, and only one was a major outlier. Integrating all six methods, a T-score of 3.5 points (with a lower and upper boundary of 3 and 4 points) was considered a reasonable MID estimate

3.5.   Secondary anchor-based estimates—As shown in the Supplementary Table, the MID estimate of 7.86 using cross-sectional global rating of mood was substantially higher than MIDs provided by our six primary methods, whereas MID estimates using finer

gradations of prospective and retrospective longitudinal change were lower (2.00 and 1.57, respectively).

3.6 MID Estimates across Four Fixed-Length PROMIS Depression Scales

Across the four fixed-length PROMIS depression scales, the MID estimates were largely comparable (See Figure 2). We observed no particular pattern for the MID estimates except, as expected, the more items in a scale, the smaller the SEM-based estimate. However, the difference between the lowest and highest SEM-based estimates was 1 point or less (0.96, 1.06, and 0.79 in samples 1, 2, and 3, respectively). Therefore, the MIDs reported in Table 2 and Figure 1 are the averages of the estimates across the four fixed-length scales.

3.7. Severity Thresholds for for PROMIS Depression Scales

Table 3 summarizes estimates of PROMIS T-score cut-points for PROMIS depression scales. Although estimates vary with the type of sample and method of estimation, PROMIS T-scores of 55, 60, 65, and 70 were approximate thresholds for mild, moderate, moderately severe, and severe depression, respectively.

4. DISCUSSION

Our study has several important findings. First, the optimal MID point estimate for the PROMIS depression scales was 3.5, with most estimating methods yielding MIDs in the 3 to 4 point range. Second, MID estimates for the four fixed-length PROMIS depression scales were similar. Third, PROMIS depression T-scores correlated strongly with scores on the PHQ-9 legacy scale, and each 1-point change in the PHQ-9 score was associated with a 1.25-point change in the PROMIS depression T-score. Fourth, PROMIS T-scores of 55, 60, 65, and 70 may represent reasonable cut-points for mild, moderate, moderately severe, and severe depression, respectively.

To our knowledge, this study is the first to provide robust MID estimates for PROMIS depression scales by using three clinical trial samples and multiple estimation methods. The only previous study to suggest a possible MID focused on 194 patients undergoing treatment for depression over 12 weeks and used PROMIS CAT administration and a retrospective global rating of change anchor to provide an MID estimate of 2.5 to 5 points.(Pilkonis et al., 2014)

The strong correlations (mean = 0.72) between PROMIS scales and the PHQ-9 were similar to correlations previously reported that ranged from 0.63 to 0.84.(Amtmann et al., 2014; Choi et al., 2014; Pilkonis et al., 2014; Tang et al., 2019; Vilagut et al., 2015) Second, the correspondence between PROMIS and PHQ-9 scores (1.25 point T-score change for each 1 point change in the PHQ-9) may be useful in interpreting studies that use only one of these measures. Although regression of PROMIS scores on PHQ-9 scores met the linearity assumption in our study, further research is needed to substantiate this conversion ratio. Some studies have used IRT, equipercentile or other linking strategies to cross-walk PROMIS scores to scores on the PHQ-9 and other depression scales.(Choi et al., 2014; Gibbons et al., 2011; Kaat et al., 2017; Kim et al., 2017; Pilkonis et al., 2014)

Ordinal categories of depression severity complement continuous scores in both research analyses as well as clinical treatment decisions. Integrating results using several different methods from both our study and prior research (Amtmann et al., 2014; Cella et al., 2014; Choi et al., 2014; Gibbons et al., 2011; Pilkonis et al., 2014) suggest PROMIS T-score depression scores of 55, 60, 65, and 70 might serve as preliminary thresholds for mild, moderate, moderately severe and severe depression, respectively. However, given the considerable variability in threshold estimates depending upon patient sample and methods used to estimate thresholds, further research is warranted.

Our finding that the four fixed-length PROMIS depression scales had comparable MID estimates is concordant with what has been previously reported for the PROMIS pain scales. (Chen et al., 2018). Additionally, fixed-length scales have been found to yield relatively similar results as those administered by CAT.(Choi et al., 2010) Collectively, these findings suggest flexibility for researchers and clinicians in choosing a PROMIS measure depending upon whether depression is a primary or secondary outcome, respondent burden considerations, and degree of desired precision.

The disability days anchor has a couple limitations. Although used as an anchor to assess MID for PROMIS and other pain measures in a previous report (Chen et al., 2018), it has been less studied than other anchors. Also, because the version of the measure used in our study asked patients to report disability days due to either depression or pain, this might have overestimated the number of disability days due specifically to depression. Nonetheless, MID estimates using the disability days anchor approximated estimates derived from other anchors.

Other limitations of our study should also be acknowledged. First, participants were disproportionately male as the majority of enrollees were recruited from VA primary clinics. Second, depression severity was mild to moderate. Therefore, our findings should be replicated in patients seen in mental health settings or who may have more severe depression. Third, two trials focused on patients with chronic pain and the third on stroke survivors. To this end, the concordance of some of our findings with studies of PROMIS depression scales in other populations is encouraging. Fourth, the follow-up times differed across the three studies (3 months or 6 months). Fifth, it is possible that MID estimates derived from the two global anchors that collapsed several levels of change into more inclusive categories (i.e., "better" and "worse") represent more than minimally important differences. In this regard, it is reassuring that these MID estimates were similar in magnitude to those obtained by other distribution-based and anchor-based methods.

In conclusion, six approaches to estimating MID in three samples all converged on a PROMIS depression T-score in the 3 to 4 point range. Since there is not a consensus on one best method for estimating MID, the relative similarity of MID estimates using multiple approaches provides reasonable support for our triangulated estimate. Future research should focus on additional clinical populations including a broader spectrum of primary care and specialty clinic patients. Also, the ordinal depression severity thresholds represented by PROMIS T-scores of 55, 60, 65, and 70 warrant further study. Meanwhile, our MID

estimates can be used to interpret research data and guide clinical decisions as well as inform power calculations for clinical studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

References

Amtmann D, Kim J, Chung H, Bamer AM, Askew RL, Wu S, Cook KF, Johnson KL, 2014 Comparing CESD-10, PHQ-9, and PROMIS depression instruments in individuals with multiple sclerosis. Rehab. Psychol 59, 220–229.

Cella D, Choi S, Garcia S, Cook KF, Rosenbloom S, Lai JS, Tatum DS, Gershon R, 2014 Setting standards for severity of common symptoms in oncology using the PROMIS item banks and expert judgment. Qual Life Res 23, 2651–2661. [PubMed: 24938431]

Cella D, Gershon R, Bass M, Rothrock N, 2015 Depression: a brief guide to the PROMIS depression instruments.

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 JF, Gershon R, Hahn EA, Lai JS, Pilkonis P, Revicki D, Rose M, Weinfurt K, Hays R, 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, 1179–1194. [PubMed: 20685078]

Chen CX, Kroenke K, Stump T, Kean J, Carpenter JS, Krebs EE, Bair MJ, Damush TM, Monahan PO, 2018 Estimating minimally important differences for the PROMIS pain interference scales: results from three randomized clinical trials. Pain 159, 775–782. [PubMed: 29200181]

Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D, 2010 Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. Qual. Life Res 19, 125–136. [PubMed: 19941077]

Choi SW, Schalet B, Cook KF, Cella D, 2014 Establishing a common metric for depressive symptoms: linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. Psychol. Assess 26, 513–527. [PubMed: 24548149]

Collaborators, U.B.o.D., 2013 The state of US health, 1990–2010: burden of diseases, injuries, and risk factors. JAMA 310, 591–608. [PubMed: 23842577]

Embretson SE, 1996 The new rules of measurement. Psychological Assessment 8, 341–349.

Eton DT, Cella D, Yost KJ, Yount SE, Peterman AH, Neuberg DS, Sledge GW, Wood WC, 2004 A combination of distribution- and anchor-based approaches determined minimally important

differences (MIDs) for four endpoints in a breast cancer scale. J Clin Epidemiol 57, 898–910. [PubMed: 15504633]

Fischer D, Stewart AL, Bloch DA, Lorig K, Laurent D, Holman H, 1999 Capturing the patient's view of change as a clinical outcome measure. JAMA 282, 1157–1162. [PubMed: 10501119]

Freedland KE, Steinmeyer BC, Carney RM, Rubin EH, Rich MW, 2019 Use of the PROMIS(R) Depression scale and the Beck Depression Inventory in patients with heart failure. Health Psychol 38, 369–375. [PubMed: 31045419]

Gaynes BN, Rush AJ, Trivedi MH, Wisniewski SR, Balasubramani GK, McGrath PJ, Thase ME, Klinkman M, Nierenberg AA, Yates WR, Fava M, 2008 Primary versus specialty care outcomes for depressed outpatients managed with measurement-based care: results from STAR*D. J. Gen. Intern. Med 23, 551–560. [PubMed: 18247097]

Gibbons LE, Feldman BJ, Crane HM, Mugavero M, Willig JH, Patrick D, Schumacher J, Saag M, Kitahata MM, Crane PK, 2011 Migrating from a legacy fixed-format measure to CAT administration: calibrating the PHQ-9 to the PROMIS depression measures. Qual Life Res 20, 1349–1357. [PubMed: 21409516]

Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, Clinical Significance Consensus Meeting, G., 2002. Methods to explain the clinical significance of health status measures. Mayo Clin Proc 77, 371–383. [PubMed: 11936935]

Hays RD, Morales LS, Reise SP, 2000 Item response theory and health outcomes measurement in the 21st century. Med Care 38, II28–42. [PubMed: 10982088]

Hirschtritt ME, Kroenke K, 2017 Screening for Depression. JAMA 318, 745–746. [PubMed: 28829850]

Jakob T, Nagl M, Gramm L, Heyduck K, Farin E, Glattacker M, 2017 Psychometric Properties of a German Translation of the PROMIS(R) Depression Item Bank. Eval Health Prof 40, 106–120. [PubMed: 26272632]

Johns SA, Kroenke K, Krebs EE, Theobald DE, Wu J, Tu W, 2013 Longitudinal comparison of three depression measures in adult cancer patients. J. Pain Symptom. Manage 45, 71–82. [PubMed: 22921152]

Kaat AJ, Newcomb ME, Ryan DT, Mustanski B, 2017 Expanding a common metric for depression reporting: linking two scales to PROMIS((R)) depression. Qual Life Res 26, 1119–1128. [PubMed: 27815821]

Kazis LE, Anderson JJ, Meenan RF, 1989 Effect sizes for interpreting changes in health status. Med. Care 27, S178–S189. [PubMed: 2646488]

Kim J, Chung H, Askew RL, Park R, Jones SM, Cook KF, Amtmann D, 2017 Translating CESD-20 and PHQ-9 Scores to PROMIS Depression. Assessment 24, 300–307. [PubMed: 26423348]

Kroenke K, Evans E, Weitlauf S, McCalley S, Porter B, Williams T, Baye F, Lourens SG, Matthias MS, Bair MJ, 2018 Comprehensive vs. Assisted Management of Mood and Pain Symptoms (CAMMPS) trial: Study design and sample characteristics. Contemp Clin Trials 64, 179–187. [PubMed: 29031492]

Kroenke K, Spitzer RL, Williams JB, Lowe B, 2010 The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. Gen. Hosp. Psychiatry 32, 345–359. [PubMed: 20633738]

Kroenke K, Theobald D, Norton K, Sanders R, Schlundt S, McCalley S, Harvey P, Iseminger K, Morrison G, Carpenter JS, Stubbs D, Jacks R, Carney-Doebbeling C, Wu J, Tu W, 2009 Indiana Cancer Pain and Depression (INCPAD) Trial: design of a telecare management intervention for cancer-related symptoms and baseline characteristics of enrolled participants. Gen. Hosp. Psychiatry 31, 240–253. [PubMed: 19410103]

Kroenke K, Unutzer J, 2017 Closing the false divide: sustainable approaches to Integrating mental health services into primary care. J Gen Intern Med 32, 404–410. [PubMed: 28243873]

Kroenke K, Wu J, Yu Z, Bair MJ, Kean J, Stump T, Monahan PO, 2016 Patient Health Questionnaire Anxiety and Depression Scale: Initial validation in three clinical trials. Psychosom Med 78, 716–727. [PubMed: 27187854]

Levin JB, Aebi ME, Smyth KA, Tatsuoka C, Sams J, Scheidemantel T, Sajatovic M, 2015 Comparing Patient-Reported Outcomes Measure Information System depression scale with legacy depression

measures in a community sample of older adults with varying Levels of cognitive functioning. Am J Geriatr Psychiatry 23, 1134–1143. [PubMed: 26320720]

Lewis CC, Boyd M, Puspitasari A, Navarro E, Howard J, Kassab H, Hoffman M, Scott K, Lyon A, Douglas S, Simon G, Kroenke K, 2019 Implementing measurement-based care in behavioral health: a review. JAMA Psychiatry 76, 324–335. [PubMed: 30566197]

Lowe B, Unutzer J, Callahan CM, Perkins AJ, Kroenke K, 2004 Monitoring depression treatment outcomes with the patient health questionnaire-9. Med Care 42, 1194–1201. [PubMed: 15550799]

Mitchell AJ, Yadegarfar M, Gill J, Stubbs B, 2016 Case finding and screening clinical utility of the Patient Health Questionnaire (PHQ-9 and PHQ-2) for depression in primary care: a diagnostic meta-analysis of 40 studies. BJPsych Open 2, 127–138. [PubMed: 27703765]

O'Connor E, Rossum RC, Henninger M, Groom HC, Burda BU, Henderson JT, Bigler KD, Whitlock EP, 2016 Screening for Depression in Adults: An Updated Systematic Evidence Review for the U.S. Preventive Services Task Force. Evidence Synthesis No. 128, Rockville, MD.

Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D, 2011 Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. Assessment 18, 263–283. [PubMed: 21697139]

Pilkonis PA, Yu L, Dodds NE, Johnston KL, Maihoefer CC, Lawrence SM, 2014 Validation of the depression item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS) in a three-month observational study. J Psychiatr Res 56, 112–119. [PubMed: 24931848]

Purvis TE, Neuman BJ, Riley LH, Skolasky RL, 2019 Comparison of PROMIS Anxiety and Depression, PHQ-8, and GAD-7 to screen for anxiety and depression among patients presenting for spine surgery. J Neurosurg Spine, 1–8.

Revicki D, Hays RD, Cella D, Sloan J, 2008 Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 61, 102–109. [PubMed: 18177782]

Schalet BD, Pilkonis PA, Yu L, Dodds N, Johnston KL, Yount S, Riley W, Cella D, 2016 Clinical validity of PROMIS Depression, Anxiety, and Anger across diverse clinical samples. J Clin Epidemiol 73, 119–127. [PubMed: 26931289]

Siu AL, Force USPST, Bibbins-Domingo K, Grossman DC, Baumann LC, Davidson KW, Ebell M, Garcia FA, Gillman M, Herzstein J, Kemper AR, Krist AH, Kurth AE, Owens DK, Phillips WR, Phipps MG, Pignone MP, 2016 Screening for Depression in Adults: US Preventive Services Task Force Recommendation Statement. JAMA 315, 380–387. [PubMed: 26813211]

Sunderland M, Batterham P, Calear A, Carragher N, 2018 Validity of the PROMIS depression and anxiety common metrics in an online sample of Australian adults. Qual Life Res 27, 2453–2458. [PubMed: 29872956]

Tang E, Ekundayo O, Peipert JD, Edwards N, Bansal A, Richardson C, Bartlett SJ, Howell D, Li M, Cella D, Novak M, Mucsi I, 2019 Validation of the Patient-Reported Outcomes Measurement Information System (PROMIS)-57 and −29 item short forms among kidney transplant recipients. Qual Life Res 28, 815–827. [PubMed: 30467780]

Vilagut G, Forero CG, Adroher ND, Olariu E, Cella D, Alonso J, investigators IN, 2015 Testing the PROMIS(R) Depression measures for monitoring depression in a clinical sample outside the US. J Psychiatr Res 68, 140–150. [PubMed: 26228413]

Von Korff M, Ormel J, Keefe FJ, Dworkin SF, 1992 Grading the severity of chronic pain. Pain 50, 133–149. [PubMed: 1408309]

Vos T, Flaxman AD, Naghavi M, Lozano R, Michaud C, Ezzati M, Shibuya K, Salomon JA, Abdalla S, Aboyans V, Abraham J, Ackerman I, Aggarwal R, Ahn SY, Ali MK, Alvarado M, Anderson HR, Anderson LM, Andrews KG, Atkinson C, Baddour LM, Bahalim AN, Barker-Collo S, Barrero LH, Bartels DH, Basanez MG, Baxter A, Bell ML, Benjamin EJ, Bennett D, Bernabe E, Bhalla K, Bhandari B, Bikbov B, Bin Abdulhak A, Birbeck G, Black JA, Blencowe H, Blore JD, Blyth F, Bolliger I, Bonaventure A, Boufous S, Bourne R, Boussinesq M, Braithwaite T, Brayne C, Bridgett L, Brooker S, Brooks P, Brugha TS, Bryan-Hancock C, Bucello C, Buchbinder R, Buckle G, Budke CM, Burch M, Burney P, Burstein R, Calabria B, Campbell B, Canter CE, Carabin H, Carapetis J, Carmona L, Cella C, Charlson F, Chen H, Cheng AT, Chou D, Chugh SS, Coffeng LE, Colan SD, Colquhoun S, Colson KE, Condon J, Connor MD, Cooper LT, Corriere M, Cortinovis

M, de Vaccaro KC, Couser W, Cowie BC, Criqui MH, Cross M, Dabhadkar KC, Dahiya M, Dahodwala N, Damsere-Derry J, Danaei G, Davis A, De Leo D, Degenhardt L, Dellavalle R, Delossantos A, Denenberg J, Derrett S, Des Jarlais DC, Dharmaratne SD, Dherani M, Diaz-Torne C, Dolk H, Dorsey ER, Driscoll T, Duber H, Ebel B, Edmond K, Elbaz A, Ali SE, Erskine H, Erwin PJ, Espindola P, Ewoigbokhan SE, Farzadfar F, Feigin V, Felson DT, Ferrari A, Ferri CP, Fevre EM, Finucane MM, Flaxman S, Flood L, Foreman K, Forouzanfar MH, Fowkes FG, Franklin R, Fransen M, Freeman MK, Gabbe BJ, Gabriel SE, Gakidou E, Ganatra HA, Garcia B, Gaspari F, Gillum RF, Gmel G, Gosselin R, Grainger R, Groeger J, Guillemin F, Gunnell D, Gupta R, Haagsma J, Hagan H, Halasa YA, Hall W, Haring D, Haro JM, Harrison JE, Havmoeller R, Hay RJ, Higashi H, Hill C, Hoen B, Hoffman H, Hotez PJ, Hoy D, Huang JJ, Ibeanusi SE, Jacobsen KH, James SL, Jarvis D, Jasrasaria R, Jayaraman S, Johns N, Jonas JB, Karthikeyan G, Kassebaum N, Kawakami N, Keren A, Khoo JP, King CH, Knowlton LM, Kobusingye O, Koranteng A, Krishnamurthi R, Lalloo R, Laslett LL, Lathlean T, Leasher JL, Lee YY, Leigh J, Lim SS, Limb E, Lin JK, Lipnick M, Lipshultz SE, Liu W, Loane M, Ohno SL, Lyons R, Ma J, Mabweijano J, MacIntyre MF, Malekzadeh R, Mallinger L, Manivannan S, Marcenes W, March L, Margolis DJ, Marks GB, Marks R, Matsumori A, Matzopoulos R, Mayosi BM, McAnulty JH, McDermott MM, McGill N, McGrath J, Medina-Mora ME, Meltzer M, Mensah GA, Merriman TR, Meyer AC, Miglioli V, Miller M, Miller TR, Mitchell PB, Mocumbi AO, Moffitt TE, Mokdad AA, Monasta L, Montico M, Moradi-Lakeh M, Moran A, Morawska L, Mori R, Murdoch ME, Mwaniki MK, Naidoo K, Nair MN, Naldi L, Narayan KM, Nelson PK, Nelson RG, Nevitt MC, Newton CR, Nolte S, Norman P, Norman R, O'Donnell M, O'Hanlon S, Olives C, Omer SB, Ortblad K, Osborne R, Ozgediz D, Page A, Pahari B, Pandian JD, Rivero AP, Patten SB, Pearce N, Padilla RP, Perez-Ruiz F, Perico N, Pesudovs K, Phillips D, Phillips MR, Pierce K, Pion S, Polanczyk GV, Polinder S, Pope CA 3rd, Popova S, Porrini E, Pourmalek F, Prince M, Pullan RL, Ramaiah KD, Ranganathan D, Razavi H, Regan M, Rehm JT, Rein DB, Remuzzi G, Richardson K, Rivara FP, Roberts T, Robinson C, De Leon FR, Ronfani L, Room R, Rosenfeld LC, Rushton L, Sacco RL, Saha S, Sampson U, Sanchez-Riera L, Sanman E, Schwebel DC, Scott JG, Segui-Gomez M, Shahraz S, Shepard DS, Shin H, Shivakoti R, Singh D, Singh GM, Singh JA, Singleton J, Sleet DA, Sliwa K, Smith E, Smith JL, Stapelberg NJ, Steer A, Steiner T, Stolk WA, Stovner LJ, Sudfeld C, Syed S, Tamburlini G, Tavakkoli M, Taylor HR, Taylor JA, Taylor WJ, Thomas B, Thomson WM, Thurston GD, Tleyjeh IM, Tonelli M, Towbin JA, Truelsen T, Tsilimbaris MK, Ubeda C, Undurraga EA, van der Werf MJ, van Os J, Vavilala MS, Venketasubramanian N, Wang M, Wang W, Watt K, Weatherall DJ, Weinstock MA, Weintraub R, Weisskopf MG, Weissman MM, White RA, Whiteford H, Wiersma ST, Wilkinson JD, Williams HC, Williams SR, Witt E, Wolfe F, Woolf AD, Wulf S, Yeh PH, Zaidi AK, Zheng ZJ, Zonies D, Lopez AD, Murray CJ, AlMazroa MA, Memish ZA, 2012 Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 380, 2163–2196. [PubMed: 23245607]

Wahl I, Lowe B, Bjorner JB, Fischer F, Langs G, Voderholzer U, Aita SA, Bergemann N, Brahler E, Rose M, 2014 Standardization of depression measurement: a common metric was developed for 11 self-report depression measures. J. Clin. Epidemiol 67, 73–86. [PubMed: 24262771]

Wyrwich KW, Nienaber NA, Tierney WM, Wolinsky FD, 1999a Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. Med Care 37, 469–478. [PubMed: 10335749]

Wyrwich KW, Tierney WM, Wolinsky FD, 1999b Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J. Clin. Epidemiol 52, 861–873. [PubMed: 10529027]

Yost KJ, Eton DT, 2005 Combining distribution- and anchor-based approaches to determine minimally important differences: the FACIT experience. Eval Health Prof 28, 172–191. [PubMed: 15851772]

Yost KJ, Eton DT, Garcia SF, Cella D, 2011 Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. J Clin Epidemiol 64, 507–516. [PubMed: 21447427]

Highlights

- A minimally important difference (MID) for the PROMIS depression T-score is 3 to 4 points.

- PROMIS depression scales of varying lengths have a similar MID.

- A PHQ-9 change of 1 point equates with a PROMIS T-score change of 1.25 points.

- PROMIS T-scores of 55, 60, 65, and 70 represent mild, moderate, moderately severe, and severe depression.
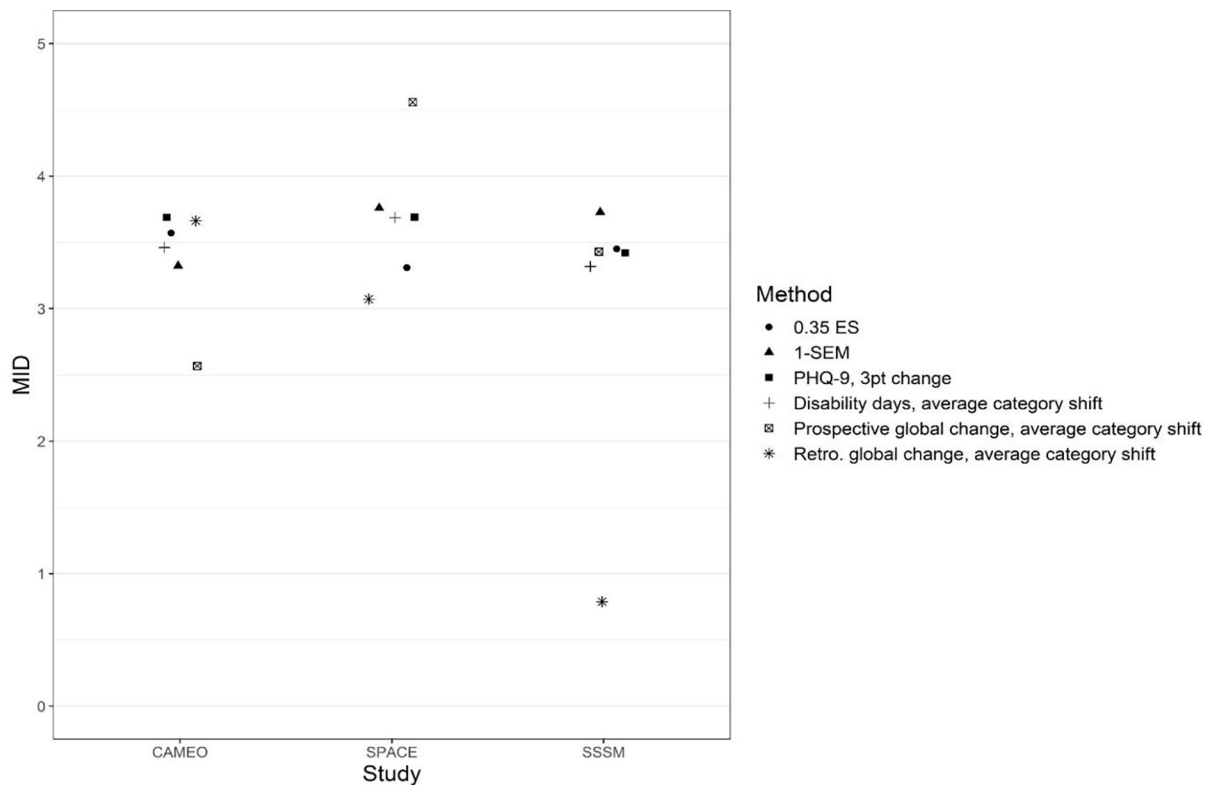
Figure 1.
Minimally Important Difference (MID) for PROMIS Depression Scales T-scores Using Six
Methods. For cross-sectional anchor-based estimates, only estimates from baseline data are
plotted. Average category shift" is the mean T-score change between adjacent categories for
the 3 categorical anchors (see Table 2). The MID estimates in the figure were the average
MIDs across the four PROMIS depression measures, as the 4 PROMIS depression measures
were comparable in MID estimates.

PHQ-9 = Patient Health Questionnaire 9-item depression scale. ES = effect size in standard
deviation units of the baseline scores. SEM = standard error of measurement. MID =
minimally important difference. Retro = retrospective

Figure 2.
Minimally Important Difference (MID) Estimates across Four PROMIS Depression
Measures of Varying Lengths.

Table 1.

Characteristics of Three Samples in the Randomized Controlled Trials (RCTs)[*]

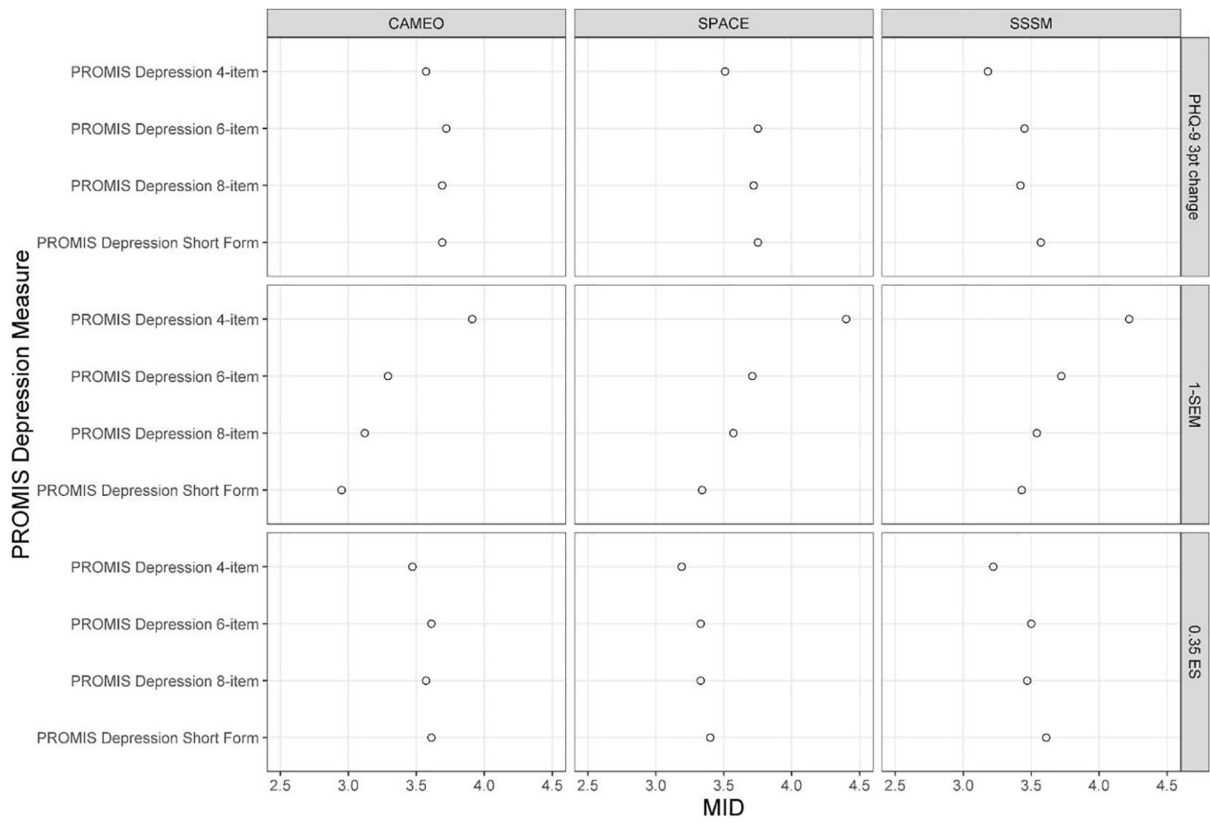|	Sample 1 CAMEO RCT (N₁=153)		Sample 2 SPACE RCT (N₂=240)		Sample 3 SSSM RCT (N₃=258)	
Clinical Population	Chronic low back pain		Chronic musculoskeletal pain		Stroke survivors	
Recruitment Setting	Primary care		Primary care		Neurology	
Age, mean (SD)	58.1	(9.3)	58.3	(13.7)	61.7	(10.8)
Male, n (%)	140	(91.5)	208	(86.7)	209	(81.0)
Race,[*] n (%)						
White	111	(72.5)	207	(86.2)	166	(64.3)
Black	37	(24.2)	18	(7.5)	78	(30.2)
Other	5	(3.3)	15	(6.3)	14	(5.4)
Education, n (%)						
Less than high school	8	(5.2)	6	(2.5)	31	(12.2)
High school	49	(32.0)	71	(29.6)	85	(33.3)
Technical school or some college	74	(48.4)	103	(42.9)	80	(31.4)
College degree or greater	22	(14.4)	60	(25.0)	59	(23.1)
Marital status, n (%)						
Married	81	(52.9)	135	(56.5)	135	(52.5)
Divorced	43	(28.1)	60	(25.1)	68	(26.5)
Other	29	(19.0)	44	(18.4)	54	(21.0)
PROMIS T-scores, mean (SD)						
Depression 4-item	53.5	(9.9)	50.3	(9.1)	51.3	(9.2)
Depression 6-item	53.2	(10.3)	49.9	(9.5)	50.5	(10.0)
Depression 8-item	53.0	(10.2)	49.6	(9.5)	50.3	(9.9)
Depression short-form	53.0	(10.3)	49.7	(9.7)	50.0	(10.3)
PHQ-9 depression score (possible range: 0–27), mean (SD)	11.1	(6.2)	6.2	(5.0)	7.7	(6.2)
Cross-sectional Global Ratings of Depression (0–4), mean (SD)	2.5	(1.0)	2.0	(0.9)	1.9	(1.0)
DSM-V depressive disorder, N (%)						
Major depression	68	(44.4)	36	(15.0)	66	(25.6)
Minor depression	21	(13.7)	23	(9.6)	21	(8.1)
Disability days in the past 4 weeks, mean (SD)	16.3	(8.6)	10.3	(9.0)	5.1	(7.7)

[*] CAMEO = **Ca**re **M**anagement for **E**ffective Use of **O**pioids RCT

SPACE = **S**trategies for **P**rescribing **A**nalgesics **C**omparative **E**ffectiveness RCT

SSSM = **S**troke **S**urvivor **S**elf-**M**anagement RCT

Table 2.

Minimally Important Difference (MID) for PROMIS Depression T-Score across Three Clinical Trials Using Different Methods[*]

MID Estimation Method	Sample 1 (CAMEO) (N₁=153)		Sample 2 (SPACE) (N₂=240)		Sample 3 (SSSM) (N₃=258)		Average Across Samples and Time
	Baseline	6 months	Baseline	3 months	Baseline	3 months	
Distribution-based analysis							
Effect size							
• 0.2 standard deviation	2.04		1.89		1.97		1.97
• 0.35 standard deviation	3.57		3.31		3.45		**3.44**
• 0.5 standard deviation	5.09		4.73		4.93		4.92
Standard error of measurement							
• 1-SEM	3.32		3.76		3.73		**3.60**
• 2-SEM	6.64		7.52		7.46		7.21
Cross-sectional anchors							
Legacy measure cross-walk							
• PHQ-9 change (3 points)	3.69	3.87	3.69	4.29	3.42	3.57	**3.76**
• PHQ-9 change (5 points)	6.15	6.45	6.15	7.15	5.70	5.95	6.26
Disability Days (past 4 weeks)							
• 0–7 to 8–14	3.10	6.55	4.43	4.80	9.23	4.85	5.49
• 8–14 to 15–21	4.13	1.63	3.65	1.38	1.08	6.13	3.00
• 15–21 to 22–28	3.15	6.73	2.98	6.05	−0.35	−2.88	2.61
Average category shift change[†]	3.46	4.97	3.69	4.08	3.32	2.70	**3.70**
Longitudinal anchors							
Prospective global change							
• Same to Better		3.44		5.69		3.73	4.29
• Same to Worse		−1.70		−3.42		−3.12	−2.75
Average category shift change[†]		2.57		4.56		3.43	**3.52**
Retrospective global change							
• Same to Better		2.74		2.29		0.30	1.78
• Same to Worse		−4.58		−3.84		−1.28	−3.23
Average category shift change[†]		3.66		3.07		0.79	**2.51**

[*] The MID estimates in the table are the averages MID estimates across the 4 PROMIS scale versions. The 4 versions were comparable in MID estimates

[†] Average category shift is the mean PROMIS T-score change between adjacent categories for these 3 categorical anchors

Table 3.

PROMIS T-scores Corresponding to Depression Severity Categories in Present and Prior Studies

Study *Method to Derive Categories*	PROMIS Depression T-Score Equivalent						
	Present Study[a] **Patient Global Mood**[b]	**Present Study**[a] **Patient PHQ-9 Range Lower Bound**[c]	**Gibbons 2011** *Patient PHQ-9 Range Lower Bound*[c]	**Amtmann 2014** *Patient PHQ-9 Range Lower Bound*[c]	**Choi 2014** *Patient PHQ-9 Range Lower Bound*[c]	**Pilkonis 2014** *Patient PHQ-9 Range Mean*[d]	**Cella 2014** *Clinical Experts*[e]
Mild Depression	52	47	42	53	53	55	55
Moderate Depression	58	53	52	59	60	60	65
Moderately Severe Depression	65	60	64	65	66	66	
Severe Depression	73	66	73	70	72	72	75

[a]Present study refers to data from all 3 trials combined in this paper.

[b]To the question about global mood in the past 7 days, responses of "mildly unhappy or down", "moderately unhappy or down", "severely unhappy or down", and "very severely unhappy or down" are classified as mild, moderate, moderately severe, and severe.

[c]T-score corresponding to PHQ-9 scores of 5, 10, 15, and 20 which represent the lower bound of mild, moderate, moderately severe, and severe depression.

[d]Mean T-score for patients with PHQ-9 scores of 5–9, 10–14, 15–19, and 20–27 which represent the ranges of mild, moderate, moderately severe, and severe depression.

[e]T-score corresponding to what clinical experts considered mild, moderate and severe symptoms by rating symptom vignettes from PROMIS item bank data in 507 cancer patients.