Minimally important differences were estimated for six PROMIS-Cancer scales in advanced-stage cancer patients

Kathleen J. Yost[a], **David T. Eton**[a], **Sofia F. Garcia**[b,c], and **David Cella**[b,c,d]

[a] Department of Health Sciences Research, Mayo Clinic, Rochester, MN

[b] Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL

[c] Department of Psychiatry and Behavioral Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL

[d] Institute for Healthcare Studies, Northwestern University Feinberg School of Medicine, Chicago, IL

Abstract

Objective—We combined anchor- and distribution-based methods to establish minimally important differences (MIDs) for six PROMIS-Cancer scales in advanced-stage cancer patients.

Study Design and Setting—Participants completed six PROMIS-Cancer scales and 23 anchor measures at an initial (n=101) and follow-up (n=88) assessment 6 to 12 weeks later. Three *a priori* criteria were used to identify useable cross-sectional and longitudinal anchor-based MID estimates. The mean standard error of measurement was also computed for each scale. The focus of the analysis was on IRT-based MIDs estimated on a T-score scale. Raw score MIDs were estimated for comparison purposes.

Results—Many cross-sectional (64%) and longitudinal (73%) T-score anchor-based MID estimates were excluded because they did not meet *a priori* criteria. The following are recommended T-score MID ranges: 17-item Fatigue (2.5–4.5), 7-item Fatigue (3.0–5.0), 10-item Pain Interference (4.0–6.0), 10-item Physical Functioning (4.0–6.0), 9-item Emotional Distress-Anxiety (3.0–4.5), and 10-item Emotional Distress-Depression (3.0–4.5). Effect sizes corresponding to these MIDs averaged between 0.40 and 0.63.

Conclusions—This study is the first to address MIDs for PROMIS measures. Studies are currently being conducted to confirm these MIDs in other patient populations and to determine whether these MIDs vary by patients' level of functioning.

1. Introduction

Patient-reported outcomes (PROs) range from specific concepts such as a single symptom (i.e., perceived pain) to more general or multidimensional concepts such as health-related quality of life (HRQL). Clinical researchers and clinicians interested in incorporating PRO

assessments into their work have long desired brief yet precise PRO measures. In recent years, the Patient-Reported Outcomes Measurement Information System (PROMIS) Network, a National Institutes of Health Roadmap Initiative, has advanced PRO measurement by developing item banks for measuring major self-reported health domains affected by chronic illness [1–3]. An item bank is a collection of calibrated items from which short form measures and computer-adaptive tests can be derived. Scores on short forms derived from the same item bank are calibrated on the same measurement metric and can therefore be compared.

The National Cancer Institute (NCI) provided supplemental PROMIS funding to ensure that the PROs developed by the network were valid for cancer patients and survivors. Input was solicited from domain experts and patients to improve the cancer relevance of PROMIS measures of fatigue, pain, physical function, and emotional distress [4]. The resulting measures will be referred to herein as PROMIS-Cancer scales.

In addition to brevity and precision, PROs used in research and clinical practice must also be interpretable. One tool for enhancing the interpretability of PROs is the minimally important difference (MID), which we define as a difference in score that is large enough to have implications for a patient's treatment or care [5]. However, for certain treatment settings and objectives, such as in patients with advanced-stage disease where palliation is the intent of treatment, the MID is patient-centered and may have no specific reference to the clinical aspect of the patient's change [5]. Our primary objective was to develop preliminary IRT-based MIDs for six PROMIS-Cancer scales that were created as short form versions of item banks. A secondary objective was to develop a reference table linking IRT-based MIDs to raw score MIDs.

2. Methods

2.1. Patients

Patients were recruited at two cancer centers in the Chicago metropolitan area: a private, suburban and a public, urban hospital. Eligible patients were at least 18 years old, had a diagnosis of advanced-stage cancer (Stages III or IV), and were able to read and understand English. In an attempt to capture the spectrum of care for advanced-stage cancer, aside from hospice care, eligible patients could be receiving any cancer treatment or follow-up care. The study was approved by the institutional review boards for the participating sites and informed consent was obtained for all participants, who received $20 compensation for each completed assessment.

2.2. Procedure

Patients completed two assessments: one at baseline and one 6 to 12 weeks later. Assessments were completed in clinic using touch screen computers. At baseline, patients completed PROMIS-Cancer scales and anchor measures (described below). Sociodemographic (e.g., age, gender, ethnicity) and clinical (e.g., treatment, stage of disease) variables were captured through self-report or from medical records.

2.3. Measures

We estimated MIDs for the following six PROMIS-Cancer scales: 17-item fatigue (Fatigue-17); 7-item PROMIS fatigue (Fatigue-7) [4]; 10-item Pain Interference (PainInt-10), 10-item Physical Function (PhysFunc-10), 9-item Anxiety (Anxiety-9), and 10-item Depression (Depression-10). The six PROMIS-Cancer scales are presented in supplementary on-line appendices. Items had 5-point ordinal rating scales, and all PROMIS-Cancer scales were scored such that a higher score represents higher levels of the concept

(e.g., higher Fatigue-7 score indicates more fatigue, higher PhysFunc-10 score indicates better physical functioning).

2.4. Combination anchor- and distribution-based approach

Using anchor-based and distribution-based methods that we [6–12] and others [13–15] have published, we identified MIDs for static, fixed-length forms. IRT-based scores were determined by mapping item responses on a given PROMIS-Cancer scale to the item calibrations in the corresponding item bank. The item calibrations were based on data for over 2000 patients with different cancer types and stages of disease that were collected as part of the PROMIS-Cancer supplement from the NCI [4]. The resulting item parameters based on data from cancer populations were then linked to the corresponding PROMIS general population calibrations using a common item equating procedure [16]. As a result of the linking procedure, scores from the PROMIS-Cancer banks and the corresponding PROMIS general banks are placed on a common scale and, hence, are comparable. IRT-based scores (i.e., theta) can be transformed to T-scores, with a mean of 50 and a standard deviation of 10 in the reference population. Raw scores were calculated as the prorated sum of the item responses, and were computed if more than 50% of the items on the scale were answered. MIDs were estimated using similar methodology for both T-scores and raw scores so that conversions from one to the other can be made by end-users. The emphasis of this paper is on the T-score MIDs because IRT-based T-scores are a better reflection of the underlying concept being measured.

2.4.1. Distribution-based approaches—Distribution-based measures rely on the statistical distributions of PRO data, including effect size measures [17,18] and the standard error of measurement (SEM) [15]. To be confident in the MID, we must confirm that the magnitude of the MID is larger than the measurement error [19] or the minimally detectable difference [20] of the scale. The SEM is a measure of precision of the scale and can be interpreted as the smallest difference or change score likely to reflect a true difference or change rather than measurement error. It also reflects the minimally detectable difference in a scale [20,21]. Therefore, the lower bound on the MID range based on anchor-based estimates was compared to the SEM. If the SEM was greater than the lower bound of the MID range, then the lower bound was increased to be larger than the SEM. In IRT analysis, each person has a standard error associated with his/her score. Thus, for the T-score MIDs, the SEM in this study was measured as the mean standard error across the sample, whereas for the raw score MIDs, the SEM was computed per convention (i.e., $SEM = SD(1 - r_{xx})^{1/2}$ where SD is the standard deviation of the scale score and r_{xx} is the reliability of the scale).

2.4.2. Anchor-based approaches—Anchor-based approaches can be cross-sectional or longitudinal. Cross-sectional approaches involve comparing PRO scores for patients in clinically-distinct groups, such as categories of performance status rating. Longitudinal approaches involve comparing changes in PRO scores to patient-reported assessments of change over time (either prospectively or retrospectively determined) [22] or to clinically-relevant measures such as response to treatment [7,9]. We collected data on 23 clinically-relevant, self-reported anchors for the cross-sectional and longitudinal anchor-based analyses. Table 1 illustrates the anchors that were used to estimate MIDs for the different scale scores. We paired anchors with PROMIS-Cancer scales based on (1) precedence, that is, whether the anchor had been used in previous research to establish MIDs for a similar domain, and (2) our confidence in their clinical relevance [23] for estimating the MID for a given scale.

We defined three *a priori* criteria for useable anchor-based estimates. The first was a Spearman correlation between an anchor (e.g., Brief Pain Inventory) and PROMIS-Cancer

scale of at least 0.3 [10,23]. If an anchor was collapsed into categories, we computed the Spearman correlation between the PRO score and the collapsed categories. The second criterion required a sample size of at least 10 in the clinically-distinct group (e.g., ECOG performance status 0, 1, 2, or 3) or change score group (e.g., minimally better, minimally worse) that was used to calculate a cross-sectional or longitudinal MID, respectively. We believed, based on some of our previous MID experience, that a difference score or change score based on fewer observations would be too unstable. The final criterion was that the anchor-based estimate had a corresponding effect size within a plausible range of 0.2–0.8; that is, estimates with effect sizes <0.2 are unlikely to be "important" and estimates with effect sizes >0.8 are unlikely to be "minimal." If an anchor-based estimate did not meet all of these criteria, it was not considered in the final MID determination.

2.4.3. Cross-sectional anchor-based analysis—In the cross-sectional analysis, anchors were used to categorize patients into multiple clinically-distinct groups. Many different anchors can be used for this purpose, provided individuals can be classified into distinct categories that are both clinically relevant but also minimally different. Score differences between adjacent, clinically-distinct categories represent estimates of the MID. Effect sizes for these estimates were computed by dividing the adjacent category score difference by the overall SD for the sample [11].

2.4.3.1. Classifying patients using cross-sectional anchors: For anchors with a 5-point ordinal rating scale, each category was considered a clinically-distinct group. For anchors with an 11-point response scale, we referred to Butt et al. [24] who determined cut-off scores to identify clinically-distinct groups for 11-point single-item measures of fatigue, pain, anxiety, and depression. Using these criteria, three severity groups were formed: 0–3 = none/mild; 4–6 = moderate; 7–10 = severe. Mean scale scores were computed for each of the three categories and differences in mean scores across adjacent categories (e.g., none/mild vs. moderate; moderate vs. severe) were considered estimates of the MID.

For multi-item anchors such as the Hospital Anxiety and Depression Scale (HADS), Brief Pain Inventory (BPI), and Functional Assessment of Chronic Illness Therapy (FACIT)-Fatigue subscale, established cutpoints for scores were used to categorize patients into clinically-distinct groups [25–27]. Differences in mean PROMIS-Cancer scale scores across adjacent, distinct categories of the FACIT-Fatigue, HADS or BPI were considered estimates of the MID. To our knowledge, cutpoints for distinguishing multiple (i.e., more than 2) clinically distinct groups have not been published for the SF-36 10-item physical function subscale (SF-36 PF); thus, it was not used as an anchor in the cross-sectional analysis. Cross-sectional analyses were conducted separately using Assessment 1 and Assessment 2 data.

2.4.4. Longitudinal anchor-based analysis

2.4.4.1. Classifying patients using prospective anchors: Prospective anchors were measured longitudinally, i.e., at Assessment 1 (T1) and Assessment 2 (T2). For single-item anchors with a 5-point response scale, change scores can range from −4 to +4. We considered a 1-point change, either positive (improvement) or negative (decline) to be clinically meaningful [10]. For single-item anchors with 11-point response scales, change scores can range from −10 to +10. While there are no recognized guidelines for interpreting meaningful change on an 11-point scale, Farrar et al. [28] have identified a 2-point improvement as clinically meaningful on an 11-point pain scale. Thus, for the two 11-point pain items (global pain, BPI worst pain), we classified patients improving by 2–3 points as "a little better" and those declining by 2–3 points as "a little worse." Mean changes in the PROMIS-Cancer scale scores corresponding to these anchor changes were considered

estimates of the MID. Although these anchor-change categories are based on findings for a pain scale, we extended them to any anchor using an 11-point scale.

Established MIDs for multi-item anchors were used to identify patients who have experienced a meaningful change in a scale score. An MID of 1.5 has been estimated for the HADS subscales [29] and the MID for the FACIT-Fatigue subscale is 3–4 points [6,30]. Proposed MIDs for the SF-36 PF subscale, which has a 0–100 score range, include 7–8 points based on one SEM [15,31] and 10 points based on a consensus panel of experts using a Delphi method [32]. Based on these findings, we considered a range of 8–10 points as the MID for the SF-36 PF subscale. Patients were classified as "a little better" or "a little worse" if their HADS, FACIT-Fatigue or SF-36 PF scores increased or decreased by at least the lower end of the published MID ranges (i.e., 1.5 points for HADS, 3 points for FACIT-Fatigue and 8 points for SF-36 PF), but no more than 2 times the MID (i.e., 3 points for HADS, 6 points for FACIT-Fatigue and 16 points for SF-36 PF). Estimates of the MID were the mean PROMIS-Cancer scale change scores in the "a little better" and "a little worse" anchor change categories. We are not aware of an established MID for the BPI Pain Interference multi-item scale; thus, we estimated the MID as ½ standard deviation [33] (1.4 points in this dataset), which is a liberal approximation (i.e., erring on the high end) of the MID when empirical data are lacking.

2.4.4.2. Classifying patients using retrospective anchors: The global rating of change (GRC) was first suggested as a clinical anchor by Jaeschke and colleagues [34] and has been implemented in several of our previous MID studies [7,12,22]. The GRC items in this study were worded specifically for each PROMIS-Cancer scale; that is, the meaning of each domain measured by the scale (physical function, pain, etc.) was briefly described at Assessment 2 and then patients rated the degree of change they have experienced on each of these domains since Assessment 1. Responses were scored on a 7-point scale ranging from −3 = "very much worse" to +3 = "very much better." Mean change scores on the PROMIS-Cancer scales corresponding to GRC item responses of +1 or +2 ("a little better," "moderately better") and −1 or −2 ("a little worse," "moderately worse") were considered estimates of the MID. Due to the large number of anchor-based MID estimates calculated, we summarized the results using nonparametric statistics, namely medians and interquartile ranges.

3. Results

A total of 101 patients completed the first assessment and 88 completed the second. Participants were predominantly female, non-Hispanic White, and receiving chemotherapy (Table 2). Participants had worse fatigue, physical function and anxiety than a general cancer population as indicated by mean Assessment 1 T-scores greater than 50 (or less than 50 for physical function) as shown in Table 3. Levels of pain and depression were comparable to the general cancer population. The most frequent responses to anchor items tended to be for the middle categories except for pain, anxiety and depression for which respondents experienced mild levels. The Assessment 1 mean FACIT-Fatigue score of 34.0 was between reported values for anemic (23.9) and non-anemic (40.0) cancer patients [27].

3.1. Distribution-based Analysis

Although the T-score metric has a mean of 50 and standard deviation of 10 in the reference sample, the observed standard deviations in this sample were less than 10 for all short forms and ranged from 6.7 for Fatigue-17 to 9.4 for PainInt-10 (Table 3). Almost all mean standard errors were less than 1/3 standard deviation (Table 4), reflecting good precision of the T-scores.

3.2. Cross-sectional Anchor-based Analyses

Spearman correlations between anchors and short form T-scores were greater than 0.3 for all anchors at both assessments except between ECOG performance status and Anxiety-9 at Assessment 2. An average of 45 MID estimates (range 37 – 52) was calculated for each of the six PROMIS-Cancer scales over both assessments. An average of 36% of the estimates that were calculated for each short form (range 27%–41%) met our *a priori* criteria for determining the MID. The most common reason for excluding an MID estimate was because the sample size for one or both of the adjacent categories being compared was less than 10. Very few estimates were discarded because the effect size for the adjacent category score difference was less than 0.2, but quite a few were discarded because the effect size was greater than 0.8. The medians for the usable cross-sectional MID estimates ranged from 4.0 for Depression-10 to 5.7 for PhysFunc-10. The effect sizes corresponding to these medians ranged from 0.48 for Depression-10 to 0.61 for Fatigue-17. The minimum, maximum, median and interquartile ranges of useable cross-sectional estimates for each short form are presented in Table 5.

3.3. Longitudinal Anchor-based Analyses

An average of 17 MID estimates (range 14–20) was calculated for each PROMIS-Cancer scale in the longitudinal analysis. Across the six scales, an average of only 27% (range 6% – 40%) calculated estimates met *a priori* criteria for determining the MIDs. The main reason estimates were discarded was a Spearman correlation for an anchor change score and short form change score of less than 0.3. Very few estimates were discarded based on a sample size less than 10 in the anchor change score category or because the effect size for the short form change score was less than 0.2. No estimates were discarded due to an effect size for the change score greater than 0.8. The medians for the useable longitudinal MID estimates were considerably lower than those from the cross-sectional analysis and ranged from 2.4 for Fatigue-7 to 3.5 for PainInt-10. Effect sizes corresponding to the median MIDs ranged from 0.29 for Fatigue-7 to 0.42 for Anxiety-9. The minimum, maximum, median and interquartile ranges of useable longitudinal estimates for each short form are presented in Table 6.

3.4. Summary of Distribution- and Anchor-based Estimates

All usable anchor-based estimates of the MID are plotted in Fig. 1. As with any empirically derived value, there is uncertainty and variability associated with MIDs. To reflect this, we recommend MID ranges rather than single point estimates [8–10,12]. In this study, the interquartile ranges shown in Fig. 1 rounded to the nearest half-integer were used to inform the recommended MIDs ranges for each PROMIS-Cancer scale. For example, the interquartile range for the Depression-10 scale T-scores was 2.6 – 4.3 points (Fig. 1), which we rounded to 2.5 – 4.5.

The final step was to compare the lower bounds of the MID ranges to the SEMs in Table 4. The lower bounds for the Anxiety-9 and Depression-10 scales rounded to the nearest half-integer were both 2.5 points, which is smaller than their SEMs of 2.6 and 2.8 points, respectively. Therefore, the lower bounds on the MID ranges for these two scales were increased to the next half-integer of 3.0 points to ensure that the MID exceeds the measurement error. The lower bounds of the MID ranges for all other scales were greater than the SEMs. Recommended T-score MIDs are summarized in Table 7.

In our previous work, the MID ranges were presented as whole integers to facilitate interpretation of an individual patient's score, which can only change by a whole number on a raw score scale. However, on an IRT-based T-score scale, an individual patient's score can change by less than a whole integer. Thus, the recommended T-score MID ranges in Table 7

are not constrained to being bounded by whole integers. Useable raw score MIDs estimates are reported in Fig. 2 and recommended raw score MID ranges rounded to the nearest whole integer are summarized in Table 7.

4. Discussion

We combined multiple anchor-based estimates from a sample of 101 advanced cancer patients to determine the MIDs for six PROMIS-Cancer scales. We defined three criteria for identifying appropriate anchor-based MID estimates. Finally, we compared the MIDs to the SEM to ensure that the MID ranges exceeded a standard unit of measurement error for each scale. IRT-based T-score MIDs and raw score MIDs were estimated. Through this exercise we were able to begin to estimate PROMIS MIDs, at least as applied to this patient population.

We observed T-score standard deviations at Assessment 1 less than 10 for all short forms, which was smaller than expected. It is possible that patients in this study, who were restricted to Stage III or IV disease, were more homogeneous with respect to their PROs than were the Stage I through IV cancer patients who participated in the calibration study.

The cross-sectional T-score MID estimates were larger than the longitudinal estimates and a large number of estimates were discarded from the cross-sectional analysis because the effect sizes were too large (i.e., >0.8). This may be due to the definitions of clinically-distinct groups in the cross-sectional analysis. That is, it is possible that the adjacent groups represented more than a minimally important difference in the domain measured by the scale. This was apparent even in commonly used cross-sectional anchors such as patient-reported ECOG performance status and general health. For example, for four of the six PROMIS-Cancer scales at Assessment 1 and for three of the six scales at Assessment 2, cross-sectional anchor-based estimates based on the comparisons of the fair vs. good categories of general health were discarded because the effect sizes were greater than 0.8. This is not unusual. We often observed very large PRO score differences (e.g., effect sizes >1.0) when comparing adjacent categories of physical functioning or general health anchors with 4–5 severity levels [6,9,10]. Thus, it is possible that these categories, while representing clinically-distinct groups, do not represent a ***minimal*** clinical distinction. There was no discernable pattern regarding whether certain anchors performed better than others overall in the cross-sectional analyses.

The longitudinal analyses did not yield many usable estimates for the MID, primarily due to weak Spearman correlations between anchor change scores and short form change scores. The MID estimates that were calculated tended to be quite small. A great deal of change in PROs often occurs in the earlier phases of treatment when side effects are first experienced by patients [12]. Over time, side effects are managed and/or patients may adapt to them, both of which can contribute to the stabilization of PRO scores over time. In the present study, patients were eligible for enrolment at any time during treatment and follow-up care to better reflect the spectrum of care for advanced-stage cancer. Therefore, it is possible that outcomes such as pain, fatigue, physical function and emotional distress for many patients in our sample had already stabilized at the time of the first assessment. As a result, little change from the first to second assessment would be experienced by those patients, which may further explain why the longitudinal MID estimates were smaller than the cross-sectional MID estimates. The GRC items, which are some of the more prevalent and longstanding anchors used in MID analyses [14,21,34], were not useful in this analysis. Only the fatigue GRC item had sufficiently large correlations with the two fatigue scales to yield useable estimates. The other four GRC items produced no useable MID estimates for their respective

PROMIS-Cancer scales. Problems with GRC items as anchors have been noted by us [12] and others [35].

Turner et al. [21] recently recommended that investigators use anchor-based methods to determine MIDs for health-related PROs. However, we observed numerous problems with anchors (e.g., poor correlation with PROs, inability to distinguish *minimal* clinically distinct groups) that can threaten one's ability to be comfortable with proposed MIDs until sufficient data have been amassed. Particularly problematic were the legacy transition rating anchors (i.e., GRC). In light of the potential problems that we noted with anchors, placing primacy on anchor-based methods seems unsupported until better methodology exists for identifying appropriate anchors for an MID analysis. At present, no one method is without limitations. Thus, we recommend triangulation across multiple methods [6–12] in which anchor- and distribution-based estimates are considered together.

Despite being scored on the same scale, recommended T-score MIDs varied by PROMIS-Cancer scale. This may represent variability in relevance of anchors across domains, differences in precision of measurement of these particular selected scales in this particular patient population, or true variability in magnitude of change required to surpass a threshold for meaningfulness in a given measured concept.

A strength of this study was the large number of variables available (between six and ten per PROMIS-Cancer scale) to use in the anchor-based analyses. Therefore, we were able to discard estimates based on *a priori* criteria and still have plenty of useable estimates on which to base the MIDs. Our study is not without limitations. The sample size of 101 and 88 at the first and second assessments, respectively, was small relative to other published MID studies. As a result of the small sample, a substantial number of cross-sectional MID estimates were discarded because they failed to meet the *a priori* criterion of at least 10 subjects in each adjacent, clinically distinct anchor category. Replication of this analysis with a larger sample is ongoing. We used a correlation of 0.3 to determine whether an anchor was appropriate for estimating the MID for a given scale. It is possible that this criterion was not stringent enough; others have recommended a correlation of at least 0.5 between GRC anchors and PRO change scores [21]. The sample of patients with advanced-stage cancer evaluated here experienced little longitudinal change in PRO scores and was potentially homogeneous; thus, results may not be generalizable to cancer patients as a whole. The range of time between the first and second assessments (6–12 weeks) was quite variable and may have contributed to the small PROMIS-Cancer scale change scores and to the weak correlations observed between anchor and scale change scores. In addition to triangulation across methods, MIDs should also be triangulated across multiple samples [10,23]. Thus, the MIDs estimated here should be confirmed in other samples of cancer patients. Finally, and possibly most importantly, the MID may vary as a function of the level of impairment experienced by patients [36]. In other words, the MID may vary by location on the severity continuum. To address this concern, one would need a sample large enough to support separate MID analyses within subsamples defined by T-scores. For example, three separate MIDs could be determined for patients with T-scores ≤45, >45 – <55 and 55+. The sample in this study was too small to accomplish this. However, this study represents the first of many that will address MIDs in PROMIS-Cancer and general PROMIS scales. Studies are currently being conducted that will allow us to answer the question of whether MIDs for PROMIS scales differ by level of functioning.

5. Conclusion

PROMIS is an IRT-based measurement system that measures self-reported physical, mental and social health using large item banks that drive brief, precise assessment referenced to

United States general population norms [1–3]. These normative T-scores are standardized with the mean set at 50 and standard deviation set at 10 units. This is the first paper to estimate the minimally important difference of those T-scores, using cross-sectional and longitudinal data. We estimated MID ranges for five PROMIS domains: Fatigue, pain, depression, anxiety, and physical functioning. Interquartile ranges of those estimated MIDs were in the range of 3–5 points for fatigue, anxiety and depression, and 4–6 points for pain and physical function. This corresponds with approximately one-third to two-third standard deviation units for each scale, and offers a beginning point for interpretation of difference and change with PROMIS measures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Liu H, Cella D, Gershon R, Shen J, Morales LS, Riley W, Hays RD. Representativeness of the Patient-Reported Outcomes Measurement Information System Internet panel. J Clin Epidemiol. 63(11):1169–78. [PubMed: 20688473]

2. 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. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. J Clin Epidemiol. 63(11):1179–94. [PubMed: 20685078]

3. Rothrock NE, Hays RD, Spritzer K, Yount SE, Riley W, Cella D. Relative to the general US population, chronic diseases are associated with poorer health-related quality of life as measured by the Patient-Reported Outcomes Measurement Information System (PROMIS). J Clin Epidemiol. 63(11):1195–204. [PubMed: 20688471]

4. Garcia SF, Cella D, Clauser SB, Flynn KE, Lad T, Lai JS, Reeve BB, Smith AW, Stone AA, Weinfurt K. Standardizing patient-reported outcomes assessment in cancer clinical trials: a patient-reported outcomes measurement information system initiative. J Clin Oncol. 2007; 25(32):5106–12. [PubMed: 17991929]

5. Wyrwich KW, Bullinger M, Aaronson N, Hays RD, Patrick DL, Symonds T. Estimating clinically significant differences in quality of life outcomes. Qual Life Res. 2005; 14(2):285–95. [PubMed: 15892420]

6. Cella D, Eton DT, Lai JS, Peterman A, Merkel DE. Combining anchor and distribution based methods to derive minimal clinically important differences on the Functional Assessment of Cancer Therapy (FACT) Anemia and Fatigue scales. J Pain Symptom Manage. 2002; 24(6):547–561. [PubMed: 12551804]

7. Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. Qual Life Res. 2002; 11(3):207–21. [PubMed: 12074259]

8. Cella D, Eton DT, Fairclough DL, Bonomi P, Heyes AE, Silberman C, Wolf M, Johnson D. What is a clinically meaningful change on the Functional Assessment of Cancer Therapy - Lung (FACT-L): Results from the Eastern Cooperative Oncology Group (ECOG) Study 5592. J Clin Epidemiol. 2002; 55:285–295. [PubMed: 11864800]

9. Eton DT, Cella D, Yost KJ, Yount SE, Peterman AH, Neuberg DS, Sledge GW, Wood WC. A combination of distribution- and anchor-based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. J Clin Epidemiol. 2004; 57(9):898–910. [PubMed: 15504633]

10. Yost KJ, Cella D, Chawla A, Holmgren E, Eton DT, Ayanian JZ, West DW. Minimally important differences were estimated for the Functional Assessment of Cancer Therapy-Colorectal (FACT-C) instrument using a combination of distribution- and anchor-based approaches. J Clin Epidemiol. 2005; 58(12):1241–51. [PubMed: 16291468]

11. Yost KJ, Eton DT. Combining distribution- and anchor-based approaches to determine minimally important differences: The FACIT experience. Eval Health Prof. 2005; 28(2):172–191. [PubMed: 15851772]

12. Yost KJ, Sorensen MV, Hahn EA, Glendenning GA, Gnanasakthy A, Cella D. Using Multiple Anchor- and Distribution-Based Estimates to Evaluate Clinically Meaningful Change on the Functional Assessment of Cancer Therapy-Biologic Response Modifiers (FACT-BRM) Instrument. Value in Health. 2005; 8(2):117–127. [PubMed: 15804320]

13. de Vet HC, Ostelo RW, Terwee CB, van der Roer N, Knol DL, Beckerman H, Boers M, Bouter LM. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Qual Life Res. 2007; 16(1):131–42. [PubMed: 17033901]

14. Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol. 1998; 16(1):139–144. [PubMed: 9440735]

15. Wyrwich KW, Nienaber NA, Tierney WM, Wolinsky FD. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. Med Care. 1999; 37(5):469–478. [PubMed: 10335749]

16. Stocking ML, Lord FM. Developing a common metric in item response theory. Applied Psychological Measurement. 1983; 7:201–210.

17. Cohen, J. Statistical Power Analysis for the Behavioral Sciences. 2. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

18. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care. 1989; 27(3 Suppl):S178–S189. [PubMed: 2646488]

19. de Vet HC, Terluin B, Knol DL, Roorda LD, Mokkink LB, Ostelo RW, Hendriks EJ, Bouter LM, Terwee CB. Three ways to quantify uncertainty in individually applied "minimally important change" values. J Clin Epidemiol. 2010; 63(1):37–45. [PubMed: 19540720]

20. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes. 2006; 4:54. [PubMed: 16925807]

21. Turner D, Shunemann H, Griffith L, Beaton D, Griffiths A, Critch J, Guyatt G. The minimal detectable change cannot reliably replace the minimal important difference. J Clin Epidemiol. 2010; 63:28–36. [PubMed: 19800198]

22. Cella D, Bullinger M, Scott C, Barofsky I. Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life. Mayo Clin Proc. 2002; 77(4):384–392. [PubMed: 11936936]

23. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008; 61(2):102–9. [PubMed: 18177782]

24. Butt Z, Wagner LI, Beaumont JL, Paice JA, Peterman AH, Shevrin D, Von Roenn JH, Carro G, Straus JL, Muir JC, Cella D. Use of a single-item screening tool to detect clinically significant fatigue, pain, distress, and anorexia in ambulatory cancer practice. J Pain Symptom Manage. 2008; 35(1):20–30. [PubMed: 17959345]

25. Snaith, RP.; Zigmond, AS. HADS: Hospital Anxiety and Depression Scale. Windsor: NFER Nelson; 1994.

26. Pain Research Group. Brief Pain Inventory. MD Anderson; 2008. http://www3.mdanderson.org/depts/prg/bpi.htm

27. Cella D, Lai JS, Chang CH, Peterman A, Slavin M. Fatigue in cancer patients compared with fatigue in the general United States population. Cancer. 2002; 94(2):528–38. [PubMed: 11900238]

28. Farrar JT, Young JP Jr, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain. 2001; 94(2):149–58. [PubMed: 11690728]

29. Puhan MA, Frey M, Buchi S, Schunemann HJ. The minimal important difference of the hospital anxiety and depression scale in patients with chronic obstructive pulmonary disease. Health Qual Life Outcomes. 2008; 6:46. [PubMed: 18597689]

30. Patrick DL, Gagnon DD, Zagari MJ, Mathijs R, Sweetenham J. Assessing the clinical significance of health-related quality of life (HrQOL) improvements in anaemic cancer patients receiving epoetin alfa. Eur J Cancer. 2003; 39(3):335–45. [PubMed: 12565986]

31. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol. 1999; 52(9):861–873. [PubMed: 10529027]

32. Wyrwich KW, Fihn SD, Tierney WM, Kroenke K, Babu AN, Wolinsky FD. Clinically important change in health-related quality of life for patients with chronic obstructive pulmonary disease. J Gen Intern Med. 2003; 18(March):196–202. [PubMed: 12648251]

33. Norman GR, Sloan JA, Wyrwich KW. Interpretation of Changes in Health-related Quality of Life: The Remarkable Universality of Half a Standard Deviation. Med Care. 2003; 41(5):582–92. [PubMed: 12719681]

34. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Controlled Clinical Trials. 1989; 10(4):407–415. [PubMed: 2691207]

35. Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. J Clin Epidemiol. 2002; 55(9):900–8. [PubMed: 12393078]

36. Wise RA, Brown CD. Minimal clinically important differences in the six-minute walk test and the incremental shuttle walking test. Copd. 2005; 2(1):125–9. [PubMed: 17136972]
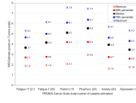
Fig. 1.
Summary of useable anchor- and distribution-based T-score MID estimates
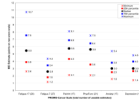
Fig. 2.
Summary of useable anchor- and distribution-based raw score MID estimates*
*Note: The possible raw score range varies across PROMIS-Cancer scale due to different number of items per scale; thus, direct comparisons of raw score MIDs across scales is not meaningful.

Table 1

PROMIS–Cancer scales matched to relevant anchors

Anchor	Rating scale	PROMIS-Cancer scale					
		Fatigue-17	Fatigue-7	PainInt-10	PhysFunc-10	Anxiety-9	Depression-10
ECOG performance status	5 pt	X	X	X	X	X	X
General health	5 pt	X	X	X	X	X	X
GRC-Fatigue	7 pt	X	X				
GRC-Pain	7 pt			X			
GRC-Physical	7 pt				X		
GRC-Anxiety	7 pt					X	
GRC-Depression	7 pt						X
Global physical health	5 pt	X	X	X	X		X
Global mental health	5 pt					X	X
Global physical functioning	5 pt	X	X		X		
Global pain	11 pt			X			
Global fatigue	5 pt	X	X		X		
Global anxiety/depression	5 pt					X	X
Global fatigue	11 pt	X	X				
Global physical limitations	11 pt	X	X		X		
Global anxiety	11 pt					X	
Global depression	11 pt						X
FACIT-Fatigue	multi-item	X	X		X		
SF-36 PF-10	multi-item	X	X		X		
BPI worst pain 24 hrs	11 pt			X			
BPI pain interference	multi-item			X			
HADS-anxiety	multi-item					X	
HADS-depression	multi-item						X

pt: point, GRC: global rating of change, FACIT: Functional Assessment of Chronic Illness Therapy, PF-10, 10-item physical function subscale, BPI: Brief Pain Inventory, HADS: Hospital Anxiety and Depression Scale

PainInt: Pain Interference, PhysFunc: Physical Function,

Table 2

Assessment 1 sample characteristics

Characteristic	N=101
Age	
Mean (SD)	59.6 (12.0)
Range	38 – 84
Male	44.6%
Race/ethnicity	
Non-Hispanic White	66%
Non-Hispanic Black	25%
Non-Hispanic Asian	3%
Hispanic (any race)	4%
Missing/other	3%
Cancer Type	
Breast	21.8%
Colorectal	18.8%
Gynecological	12.9%
Lung	11.9%
Prostate	5.9%
Head and Neck	5.9%
Other	13.9%
Missing/Unknown	9.9%
Treatment in past month	
Chemotherapy only	74.3%
Chemo- and radiation therapy	9.9%
Other mixed modalities	13.8%
Missing	2.0%
Highest grade completed	
Less than high school graduate	10.9%
High school graduate/GED	20.8%
Some college	25.7%
College graduate	26.7%
Advanced degree	15.8%
Household Income	
<$20,000	35.6%
$20,000 – $49,999	15.8%
$50,000 – $99,999	16.8%
$100,000+	30.7%
Missing	1.0%

Table 3

Assessment 1 PROMIS-Cancer scale scores and selected[*] anchor distributions (n=101)

PROMIS-Cancer scale	Mean T-Scores	T-Score SD
17-item Fatigue	54.5	6.7
7-item Fatigue	53.6	7.7
10-item Pain Interference	51.7	9.4
10-item Physical Function	46.1	8.9
10-item Anxiety	53.2	7.5
9-item Depression	50.5	8.3

Anchor	Response Scale	%
Patient-reported ECOG performance status	Normal (0)	19.8
	Some symptoms (1)	47.5
	Bed rest < 50% of day (2)	22.8
	Bed rest > 50% of day (3)	9.9
Overall fatigue	None	9.0
	Mild	32.0
	Moderate	50.0
	Severe	9.0
	Very severe	0.0
Overall pain (11 point scale collapsed)	None/Mild (0–3)	61.2
	Moderate (4–6)	24.5
	Severe (7–10)	14.3
Overall physical health	Excellent	3.0
	Very good	20.8
	Good	42.6
	Fair	30.7
	Poor	3.0
Overall mental health	Excellent	17.8
	Very good	34.7
	Good	24.8
	Fair	21.8
	Poor	1.0

[*] For brevity, five of the 23 anchors are shown

Table 4

IRT-based mean standard error of measurement for PROMIS-Cancer scales

	Fatigue-17	Fatigue-7	PainInt-10	PhysFunc-10	Anxiety-9	Depression-10
Assessment 1	1.7	2.6	2.4	2.4	2.6	2.8
Assessment 2	1.8	2.7	2.4	2.5	2.6	2.8

Table 5

Summary of cross-sectional T-score MID estimates computed for assessments 1 and 2

PROMIS-Cancer scale	Fatigue-17 14 (26.9%)		Fatigue-7 17 (32.7%)		PainInt-10 15 (39.5%)		PhysFunc-10 19 (39.6%)		Anxiety-9 15 (40.5%)		Depression-10 14 (35.0%)	
Number of useable estimates (%)*												
	Points	Effect Sizes	Points	Effect Sizes	Points	Effect Sizes	Points	Effect Sizes	Points	Effect Sizes	Points	Effect Sizes
Minimum	2.7	0.34	2.1	0.24	2.1	0.22	3.1	0.34	2.3	0.3	1.8	0.21
25th Percentile	3.7	0.49	4.1	0.48	4.3	0.46	4.5	0.48	3.2	0.42	3.1	0.38
Median	4.2	0.61	4.7	0.60	5.2	0.53	5.7	0.60	4.2	0.55	4.0	0.48
75th Percentile	4.7	0.68	5.5	0.71	6.1	0.65	6.5	0.71	4.7	0.61	4.7	0.56
Maximum	5.3	0.79	6.1	0.76	7.6	0.78	7.5	0.79	5.7	0.74	5.3	0.63

PainInt: Pain Interference, PhysFunc: Physical Function

*
The total number of estimates calculated varied by PROMIS-Cancer scale

Table 6

Summary of longitudinal T-score MID estimates

PROMIS-Cancer scale	Fatigue-17		Fatigue-7		PainInt-10		PhysFunc-10		Anxiety-9		Depression-10	
Number of useable estimates (%)*	7 (35.0%)		8 (40.0%)		2 (14.3%)		1 (5.6%)		5 (35.7%)		4 (28.6)	
	Points	Effect Sizes	Points	Effect Sizes	Points	Effect Sizes	Points	Effect Sizes	Points	Effect Sizes	Points	Effect Sizes
Minimum	1.9	0.27	1.9	0.24	2.2	0.23	3.0	0.33	1.6	0.22	2.1	0.25
25th Percentile	2.5	0.36	2.2	0.27					2.0	0.27	2.4	0.29
Median	2.6	0.38	2.4	0.29	3.5	0.37	3.0	0.33	3.1	0.42	2.7	0.34
75th Percentile	3.0	0.44	3.3	0.41					3.2	0.44	3.2	0.40
Maximum	3.6	0.52	4.2	0.51	4.8	0.5	3.0	0.33	4.7	0.64	3.7	0.46

PainInt: Pain Interference, PhysFunc: Physical Function

*The total number of estimates calculated varied by PROMIS-Cancer scale.

Table 7

Recommended IRT-based T-score MIDs and Raw Score MIDs for PROMIS-Cancer Short Forms in Advanced Cancer Patients

Short Form	T-Score MID Points	T-Score MID Effect Sizes[*]	Raw Score MID Points	Raw Score MID Effect Sizes[§]
17-item Fatigue	2.5 – 4.5	0.37 – 0.67	4.0 – 8.0	0.33 – 0.65
7-item Fatigue	3.0 – 5.0	0.39 – 0.65	2.0 – 3.0	0.38 – 0.57
Pain Interference	4.0 – 6.0	0.43 – 0.64	4.0 – 7.0	0.39 – 0.69
Physical Function	4.0 – 6.0	0.45 – 0.67	4.0 – 6.0	0.42 – 0.63
Emotional Distress-Anxiety	3.0 – 4.5	0.40 – 0.60	3.0 – 4.0	0.45 – 0.60
Emotional Distress-Depression	3.0 – 4.5	0.36 – 0.54	3.0 – 4.0	0.43 – 0.57

[*] Calculated as the T-Score MID divided by the Assessment 1 T-score standard deviation

[§] Calculated as the Raw Score MID divided by the Assessment 1 Raw Score standard deviation