

Setting standards for severity of common symptoms in oncology using the PROMIS item banks and expert judgment

David Cella · Seung Choi · Sofia Garcia · Karon F. Cook · Sarah Rosenbloom · Jin-Shei Lai · Donna Surges Tatum · Richard Gershon

Accepted: 4 June 2014 / Published online: 18 June 2014
© Springer International Publishing Switzerland 2014

Abstract

Background Although the use of patient-reported outcome measures (PROs) has increased markedly, clinical interpretation of scores remains lacking. We developed a method to identify clinical severity thresholds for pain, fatigue, depression, and anxiety in people with cancer.

Methods Using available Patient-Reported Outcomes Measurement Information System (PROMIS) item bank response data collected on 840 cancer patients, symptom vignettes across a range of symptom severity were developed and placed on index cards. Cards represented symptom severity at five-point intervals differences on the *T* score metric [mean = 50; standard deviation (SD) = 10]. Symptom vignettes for each symptom were anchored on these standardized scores at 0.5 SD increments across the full range of severity. Clinical experts, blind to the PROMIS score associated with each vignette, rank-ordered the vignettes by severity, then arrived at consensus regarding which two vignettes were at the upper and lower boundaries of normal and mildly symptomatic for each symptom. The procedure was repeated to identify cut scores separating mildly from moderately symptomatic, and moderately from severely symptomatic scores. Clinician severity rankings were then compared to the *T* scores upon which the vignettes were based.

Results For each of the targeted PROs, the severity rankings reached by clinician consensus perfectly matched the numerical rankings of their associated *T* scores. Across all symptoms, the thresholds (cut scores) identified to differentiate normal from mildly symptomatic were near a *T* score of 50. Cut scores differentiating mildly from moderately symptomatic were at or near 60, and those separating moderately from severely symptomatic were at or near 70.

Conclusions The study results provide empirically generated PROMIS *T* score thresholds that differentiate levels of symptom severity for pain interference, fatigue, anxiety, and depression. The convergence of clinical judgment with self-reported patient severity scores supports the validity of this methodology to derive clinically relevant symptom severity levels for PROMIS symptom measures in other settings.

Keywords PROMIS · Patient-reported outcomes · Symptom severity levels · Standard setting · Cancer

Introduction

Patient-reported outcome (PRO) assessment has become commonplace in clinical research and is emerging as an activity of interest in clinical practice monitoring [1–6]. Recent advances in the application of item-response theory (IRT) models have improved both the precision of PRO scores and the efficiency by which they are assessed [7]. One example of this is the Patient-Reported Outcomes Measurement Information System (PROMIS) initiative that has produced over 50 IRT-calibrated pools of items (item banks) developed to assess a range of physical, mental, and social health concepts, including symptoms and function

D. Cella (✉) · S. Choi · S. Garcia · K. F. Cook · S. Rosenbloom · J.-S. Lai · R. Gershon
Department of Medical Social Sciences, Northwestern University, 633 N. St. Clair, 19th Floor, Chicago, IL 60611, USA
e-mail: d-cella@northwestern.edu

D. S. Tatum
Meaningful Measurement Inc., 3170 North Sheridan, #1008, Chicago, IL 60657, USA

(www.nihpromis.org) [8–10]. The PROMIS item banks, and their associated tools, including static short forms and computerized adaptive testing (CAT), are all scored on a *T* score metric that has a mean of 50 and the standard deviation is 10 [10, 11]. The PROMIS metric is anchored to a general US population sample that matched the distribution of the 2000 census with respect to sex, age, and race/ethnicity. The advantage of this metric is that scores allow comparison to a reference population of interest. For example, a symptom score of 60 is one standard deviation worse than the reference US general population sample.

Normative comparisons provide a helpful context for scores, but they do not provide information regarding what patient-reported severity level would warrant clinical attention. Norm-referenced scores do not by themselves inform clinicians as to level of severity or, for that matter, the clinical meaningfulness of a specific magnitude of worsening or improvement. Expert referencing of severity to symptom scores can bring clinical meaning to the numeric score and enable better interpretation of change. Some have begun to address this problem empirically, using responses from single-item, 0–10 numeric rating scale (NRS) pain severity measures [12–16]. Zelman and colleagues [15] developed a metric for a “day of manageable pain control,” based on the 0–10 NRS. Classification methods for other symptoms have also been proposed: Using regression-based analyses linking 0–10 NRS reports to activity limitation and other external criteria (anchors), optimal cut points for fatigue severity levels have been suggested for a 0–10 single-item fatigue scale [17]. Similarly, Given and colleagues estimated cut points for mild, moderate, and severe levels of 16 cancer-related symptoms by associating 0–10 severity scores with self-reported levels of interference from each of the 16 symptoms [18]. These efforts represent useful advances; but they are limited in their application to relatively coarse single-item scales, and in the use of statistical methods alone, rather than clinician or patient judgment, to set threshold levels for severity terms such as mild, moderate, and severe. Such scales, and classification systems based upon their scores, have important drawbacks. Whereas they are appropriate for very narrow concepts measured for brief periods, such as pain or fatigue intensity, single-item scales are a poor choice for measuring more complex dimensions such as interference caused by pain and fatigue, or depression [19]. Multi-item measures typically have greater reliability and validity in measuring these more complex dimensions.

Classification systems based on multi-item measures are common in educational and psychological testing, but rare in health-outcome measurement. Referred to as “standard setting,” these empirical methods identify valid and defensible cut scores that could be used for high-stakes

decision making. For example, to assign standards for educational achievement (e.g., passing a course), Angoff [20] asked expert judges (educators) to estimate the probability that a barely proficient (passing) student would correctly answer each question on a test. The performance standard on the total score score is then determined by aggregating these per-question estimates. Lewis et al. [21] and Impara and Plake [22] proposed a standard setting approach in which a booklet of test questions lists items in order from easiest to most difficult. Experts then placed a “bookmark” at the test item in the booklet that a barely passing student became less likely to answer correctly than incorrectly. Similarly, using a 75-item reading comprehension test as the basis, Baghaea compared actual student performance to rater judgments of the minimum ability level required, on a six-point proficiency scale, to answer each question correctly expected item-level performance, obtaining very high concordance on the same metric [23].

The purpose of this study was to adapt and apply a standard setting method, derived from educational and psychological testing, to identify cut points for classifying the severity of four symptoms: pain, fatigue, anxiety, and depression, as measured by the PROMIS item banks.

Method

Subjects

Expert panelists were 22 attendees at an investigator and consultant meeting held in conjunction with NCI grant number CA60068. The purpose of the meeting was to prepare for a randomized clinical trial of a cancer symptom monitoring intervention. The primary task of the meeting was to establish clinical thresholds (“cut scores”) for symptom severity as a companion to an intervention guide for the treatment of pain, fatigue, depression, and anxiety in people with advanced cancer. The intervention guidelines were drawn from the National Comprehensive Cancer Network Guidelines for management of these four symptoms. Each of these guidelines specifies graded levels of intervention for mild, moderate, and severe symptom presentation. Given the plan was to use PROMIS CAT measures for each of these symptoms, it was an essential prerequisite that we have clinical cut scores to distinguish these clinical levels from each other.

To be eligible to serve as a panelist, an expert had to have a minimum 3 years’ experience treating over 100 people with cancer who present with the target symptom. In all, there were 22 experts across the four symptoms (22 pain; 22 fatigue; 22 depression; 21 anxiety). If eligible, panelists could serve on more than one expert panel. Of the 22 experts, there were 11 psychologists, 4 oncologists, 2

oncology nurses, 2 occupational therapists, 2 physical therapists, and a pharmacist.

Patient symptom vignettes were drawn from PROMIS item bank calibrations derived from a testing of PROMIS pain, fatigue, anxiety, and depression item banks. These four item banks were developed with mixed clinical and general population samples and calibrated on a large US general population sample [24–26]. Our cancer-focused study team reviewed, revised, and augmented these banks to increase cancer relevance across these four symptoms, for use in our study. Decisions regarding item and bank modification were driven by qualitative data (patient focus groups and cognitive interviews) gathered in a cancer-specific PROMIS qualitative research supplement as well as item performance in the PROMIS general population field test [8, 10]. For example, we replaced somatic-related depression items with more affect-related items for increased relevance within a cancer population. Cancer patient data for this expert panel standard setting exercise were collected using two accrual methods: in clinic ($n = 339$) and a customized online survey through national support societies (e.g., Y-Me, Gilda's Club; $n = 501$). After completing their randomly assigned bank testing (each assessment consisted of two item banks), participants were invited to complete additional item bank testing, and several agreed. As a result, 2,055 item bank assessments from 840 unique participants (most participants completed more than 1 assessment) provided sufficient cancer patient item calibration data for fatigue ($n = 512$), pain ($n = 529$), depression ($n = 507$), and anxiety ($n = 507$). Characteristics of the patient sample are detailed in Table 1. Symptom item bank calibrations, including per-item information on the most likely response based upon symptom severity across the full range of the symptom, were used to create patient vignettes for each symptom severity level. We checked the comparability of group-based T score estimations derived from these cancer-specific calibrations to T score estimations derived from the standard PROMIS item calibrations. In all cases, estimations were very similar, never deviating by more than two T score units (i.e., 0.2 SD).

Measures

Four PROMIS measures were used for the current study—pain interference, fatigue, anxiety, and depression. All PROMIS item banks were derived from large pools of existing, modified and new items, developed with mixed qualitative and quantitative methods [24–26]. Item responses were modeled (calibrated) using the graded response model [27]. Derived scores were transformed to the PROMIS T score metric ($X = 50$; $SD = 10$), as referenced to the 2000 US census on age, sex, and race/ethnicity [11]. High

scores reflect worse symptoms. Once items are calibrated using an IRT model, the most likely (probable) response for each item can be derived at any given level of the trait (symptom) being measured. This information was used to build the symptom vignette cards (Table 2 for example).

Procedures

Developing score level vignettes

The standard setting approach for this study was a modification of methods applied in education by Lewis and colleagues [21] and Impara and Plake [22]. For each of the four PROMIS measures, we identified score locations on the PROMIS T score metric that were five points (0.5 SD units) apart. The number of identified locations depended on the distribution of the outcome in the PROMIS standardization sample. Eight severity levels were identified for pain interference (T scores ranged from 45 to 80). For anxiety and depression, ten severity levels were identified corresponding to T scores of 40–85. Eleven severity levels were identified for fatigue (T score range 35–85). Next, for each PROMIS outcome measure, five items were selected from the item bank, and for each of these items, the “most likely responses” were identified for each of the 8–11 locations on the PROMIS T score metric. As an illustration, Fig. 1 shows the most likely responses to each depression item, based on a person's locations on the PROMIS metric. For example, the most likely response to the item, “I felt lonely even when I was with other people” for a person with a T score of 50 is 1 (“never”). For someone with a score of 55, the most likely response is 2 (“rarely”); and for a person with a score of 60, the most likely response is 3 = sometimes (see Fig. 1). Using similar response probability tables for items of each PROMIS bank, vignettes consisting of five items and the most likely responses were created for the target scores. The five items for each vignette were selected to maximize content coverage. For example, for the fatigue vignettes, we selected items that targeted both physical and cognitive fatigue. The items also were selected to maximize the diversity of most likely responses (e.g., “not at all” to “very much.”) so that experts could not easily rank the vignettes based solely on the response pattern.

Rating score level vignettes

Each vignette (five items, with the most likely response to each item for a given T score level) was printed on 3” × 5” card stock of the same color as other vignettes in the same symptom domain. Raters were kept blind to the T score value associated with each card; they worked only with the five items and responses on each card. Table 2

Table 1 Sample characteristics ($n = 840$)*

Characteristic	Fatigue ($n = 512$) %	Pain ($n = 529$) %	Depression/anxiety ($n = 507$)
Age			
<40	9	9	9
41–50	21	22	21
51–64	48	49	49
65–79	18	18	17
80+	3	3	3
Age (mean)	56.0 (SD = 12.07)	55.7 (SD = 12.03)	55.7 (SD = 12.12)
Gender			
Male	28	26	28
Female	72	74	72
Race			
White	81	82	82
Black or African American	15	15	15
Asian	2	1	2
American Indian/Alaska native	0	1	1
Native Hawaiian/other Pacific Islander	0	0	0
Multiple races or “Other”	3	3	3
Ethnicity			
Spanish/Hispanic/Latino	2	2	2
Not Spanish/Hispanic/Latino	98	98	98
Educational attainment			
Less than high school	5	5	5
High school graduate/GED	13	14	12
Some college	28	30	32
College degree	32	30	31
Advanced degree (MA, PhD, MD)	22	21	20
Continuum of care*			
Active tx early stage	22	22	22
Active tx advanced stage	13	12	12
Active tx unknown stage	11	11	11
Post tx within 5 years dx	32	32	34
Post tx 5+ years dx	21	22	20
Cancer site**			
Breast	33	36	35
Urologic	19	19	19
Gynecologic	10	9	10
Thoracic/head and neck	8	8	7
Gastrointestinal	10	9	10
Hematologic	12	12	11
Other/unknown	8	7	8

* Total $n = 840$. Because most patients completed at least two item banks
tx treatment

presents a sample anxiety (blue card) vignette. The vignette cards for each domain were shuffled and placed in an envelope prior to presentation to the experts.

All participants were asked to complete the card sorting exercise for all domains. Experts were given an envelope

with the case example cards and were instructed to sort the cards from least severe to most severe. Each panel member privately reviewed the cards and rank-ordered them in terms of the degree of symptom severity represented by each item-response cluster. They were reminded to look at

each item and the response individually, rather than adding or averaging response choices alone.

Once the cards were sorted, experts were asked to assign a rank to each card (“1” being least severe) and record it on the exercise sheets provided for each domain. Experts were encouraged to provide a unique ranking for each card, however, more than one card could be assigned a common

ranking if the expert believed that certain vignette “patients” were equivalent in terms of symptom severity. After rankings were recorded on the exercise sheet, experts were asked to place bookmarks to differentiate the cards into different severity categories (normal, mild, moderate, and severe) by drawing three vertical lines, one delineating each of the following: (1) a separation between the cards representing a normal level of the symptom and a mild level of the symptom; (2) a separation between the cards representing a mild level of the symptom and a moderate level of the symptom; and (3) a separation between the cards representing a moderate level of the symptom and a severe level of the symptom. Treatment interventions associated with each severity level (mild, moderate, severe) were also listed by relevant experts as part of the clinical trial preparation. Individual data were entered into spreadsheets as experts completed the exercise.

Next, experts participated in one of three (pain; fatigue; anxiety/depression) domain-specific group discussions to achieve consensus on vignette severity rankings and clinical cut points. Experts were divided by sub-specialization for these meetings, such that the pain group was comprised primarily of oncologists and nurses, the anxiety/depression group was

Table 2 Sample vignette card as presented to experts for ranking and bookmarking

<i>Anxiety (blue card)</i>						
1	I felt anxious	Never	Rarely	Sometimes	Often	Always
2	I felt upset	Never	Rarely	Sometimes	Often	Always
3	I felt worried	Never	Rarely	Sometimes	Often	Always
4	I felt uneasy	Never	Rarely	Sometimes	Often	Always
5	I felt tense	Never	Rarely	Sometimes	Often	Always

Bold font indicates the most likely item response among people with anxiety *T* score = 55. These *bold font* responses were circled to depict a patient with an anxiety *T* score of 55 (score value was not provided to experts)

Fig. 1 Response probability table for PROMIS depression item bank

Item Content	T Score														
	30	35	40	45	50	55	60	65	70	75	80	85	90		
I felt lonely even when I was with other people	1	1	1	1	1	2	3	3	4	4	5	5	5		
I felt worthless	1	1	1	1	1	1	3	3	4	5	5	5	5		
I felt that I had nothing to look forward to	1	1	1	1	1	2	3	4	4	4	5	5	5		
I felt helpless	1	1	1	1	1	2	3	3	4	5	5	5	5		
I withdrew from other people	1	1	1	1	1	2	3	3	4	4	5	5	5		
I felt that nothing could cheer me up	1	1	1	1	1	2	3	3	4	4	5	5	5		
I had mood swings	1	1	1	1	2	3	3	4	4	5	5	5	5		
I felt that I was not as good as other people	1	1	1	1	1	1	3	3	4	4	5	5	5		
I felt like crying	1	1	1	1	2	3	3	4	4	5	5	5	5		
I felt sad	1	1	1	2	2	3	3	4	4	4	5	5	5		
I felt that I wanted to give up on everything	1	1	1	1	1	1	2	3	4	4	5	5	5		
I felt that I was to blame for things	1	1	1	1	1	2	3	3	4	4	5	5	5		
I felt like a failure	1	1	1	1	1	2	3	3	4	4	5	5	5		
I had trouble feeling close to people	1	1	1	1	1	2	3	3	4	4	5	5	5		
I felt disappointed in myself	1	1	1	1	2	2	3	4	4	5	5	5	5		
I felt that I was not needed	1	1	1	1	1	1	3	3	4	4	5	5	5		
I felt lonely	1	1	1	1	1	3	3	4	4	4	5	5	5		
I felt depressed	1	1	1	1	2	3	3	4	5	5	5	5	5		
I felt discouraged about the future	1	1	1	1	2	3	3	4	4	5	5	5	5		
I found that things in my life were overwhelming	1	1	1	1	2	3	3	4	4	5	5	5	5		
I felt unhappy	1	1	1	2	2	3	3	4	4	5	5	5	5		
I felt unloved	1	1	1	1	1	1	3	3	4	4	5	5	5		
I felt I had no reason for living	1	1	1	1	1	1	1	3	3	5	5	5	5		
I felt hopeless	1	1	1	1	1	2	3	3	4	5	5	5	5		
I felt pessimistic	1	1	1	1	2	2	3	3	4	5	5	5	5		
I felt that my life was empty	1	1	1	1	1	1	3	3	4	5	5	5	5		
I felt emotionally exhausted	1	1	1	1	2	3	3	4	4	5	5	5	5		
I felt like I needed help for my depression	1	1	1	1	1	2	3	3	5	5	5	5	5		
I felt angry	1	1	1	1	2	3	3	3	4	4	4	5	5		
I felt irritable	1	1	1	2	2	3	3	4	4	4	5	5	5		

comprised primarily of psychologists, and the fatigue group was comprised of an even mixture of clinical specialists. Each group had at least one physical or occupational therapist. All domain groups were led by experienced facilitators who began with brief questions aimed to elicit participants' general impressions of the exercise, understanding how participants made decisions regarding sorting/ranking cards, and understanding how they determined cut points. Flip charts were used to create a summary of how domain group participants ranked each of the case example cards. Open discussion was initiated by noting where there were areas of disagreement, and group members who disagreed were encouraged to discuss the rationale for their rankings with others.

After the domain groups concluded their task and achieved consensus around ranking and cut scores, the entire group of 22 experts convened for large open discussion. The explicitly identified goal was to achieve final consensus. Results of individual rankings were added to the public summary through use of a flip chart, and domain groups determined whether or not the large group results should alter their ranking and cut score determinations. Open discussion was encouraged. Domain groups also presented lists of interventions for each severity level. The full group discussion of domain group recommendations for ranking and cut scores

was summarized for each domain group which was in turn charged with proposing final consensus recommendations.

Results

Median and mean individual rankings of symptom severity are plotted in Figs. 2, 3, 4, and 5. Beginning with Fig. 2, the top panel presents mean and median rankings of the 22 panel experts for the 8 pain vignettes. The horizontal dotted lines depict the final consensus for lower bound of mild, moderate, and severe pain interference. The lower left panel displays the distribution of pain interference scores in the cancer sample, with dotted vertical lines indicating cut scores separating normal, mild, moderate, and severe pain interference. The lower right panel provides proportions of patients in each severity category (Figs. 2, 3, 4, and 5).

Figure 3 displays the results for fatigue. As with pain, the top panel presents mean and median rankings of the panel experts for the 11 fatigue vignettes, and the horizontal dotted lines represent final consensus for lower bound of mild, moderate, and severe fatigue. The lower panels display distribution of fatigue scores in cancer sample (left) and proportions of patients in each severity

Fig. 2 Top panel plots the vignette *T* score (*x* axis) against the median and mean card rankings according to expert consensus (*y* axis). Dotted horizontal lines reflect the expert consensus on bookmarks separating the severity of symptom vignettes (mild; moderate; severe). Experts were blind to vignette *T* score values throughout the exercise. Lower left panel displays the distribution of pain scores (*y* axis) by *T* score (*x* axis), with vertical lines separating clinical categories (none; mild; moderate; severe). Lower right panel indicates the number and proportion of patients in each of the four clinical categories

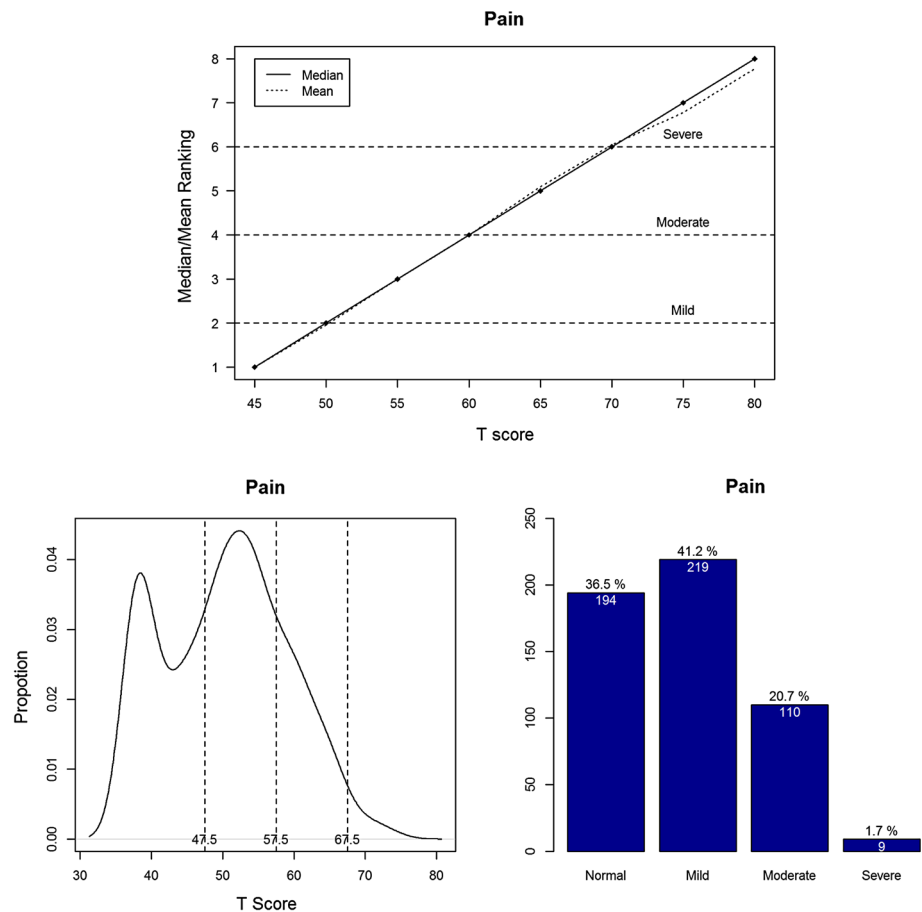
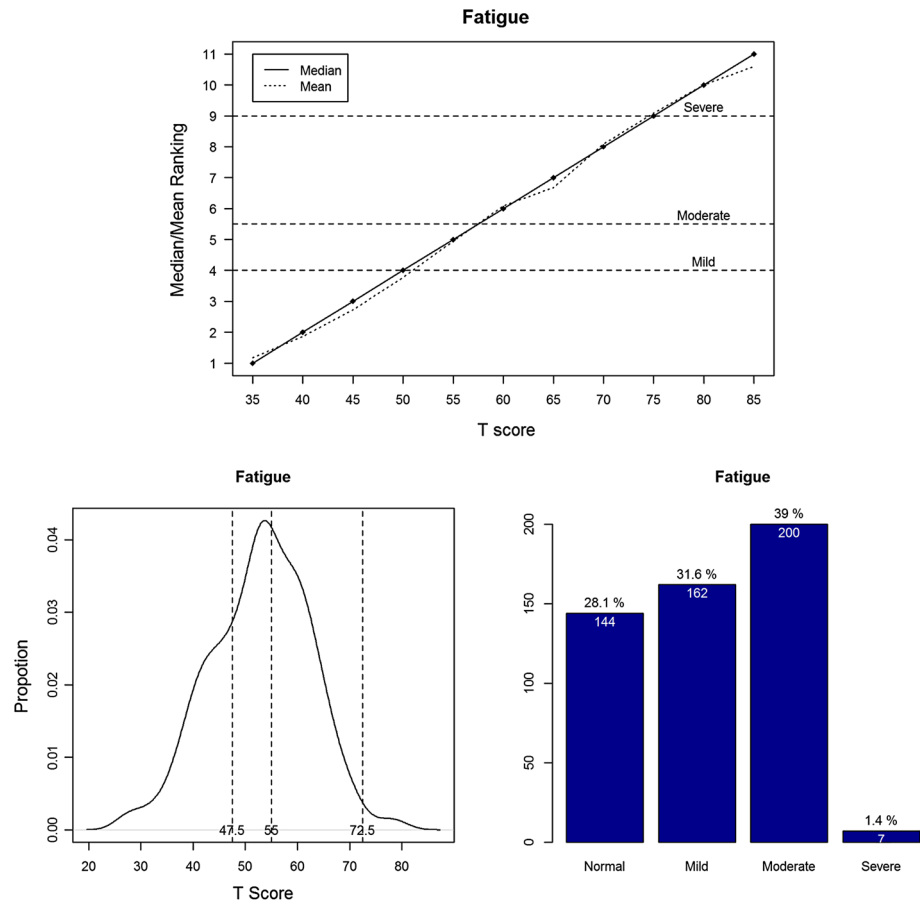


Fig. 3 Top panel plots the vignette T score (x axis) against the median and mean card rankings according to expert consensus (y axis). Dotted horizontal lines reflect the expert consensus on bookmarks separating the severity of symptom vignettes (mild; moderate; severe). Experts were blind to vignette T score values throughout the exercise. Lower left panel displays the distribution of fatigue scores (y axis) by T score (x axis), with vertical lines separating clinical categories (none; mild; moderate; severe). Lower right panel indicates the number and proportion of patients in each of the four clinical categories



category (right). Figures 4 and 5 present the same panels of data for anxiety and depression, respectively.

As evident in these figures, there was, in every case, complete agreement of expert rank (y axis) with T score (x axis). Individual rater disagreement at the extremes of the symptoms did not occur. Even in the more challenging midrange of symptom severity, disagreement was rare. Domain group consensus through these discussions regarding ranking and cut scores was therefore easily achieved for all four symptoms (see Figs. 2, 3, 4, and 5).

Consensus was reached for anxiety on vignette ranking and severity cut scores with minimal discussion. For depression, discussion centered on whether the T score = 70 card should be “moderate” or “severe;” in the end it was retained in the “moderate” category (We note that panelists were kept blind to the actual numeric T score designation of all index card vignettes.). For the fatigue domain, the cards remained in the same order as originally set by the domain group, however, the mild-moderate cut score was shifted 0.5 SD lower (between the $T = 55$ and $T = 60$ cards). The pain rankings and cut scores were not changed from the original domain group recommendations.

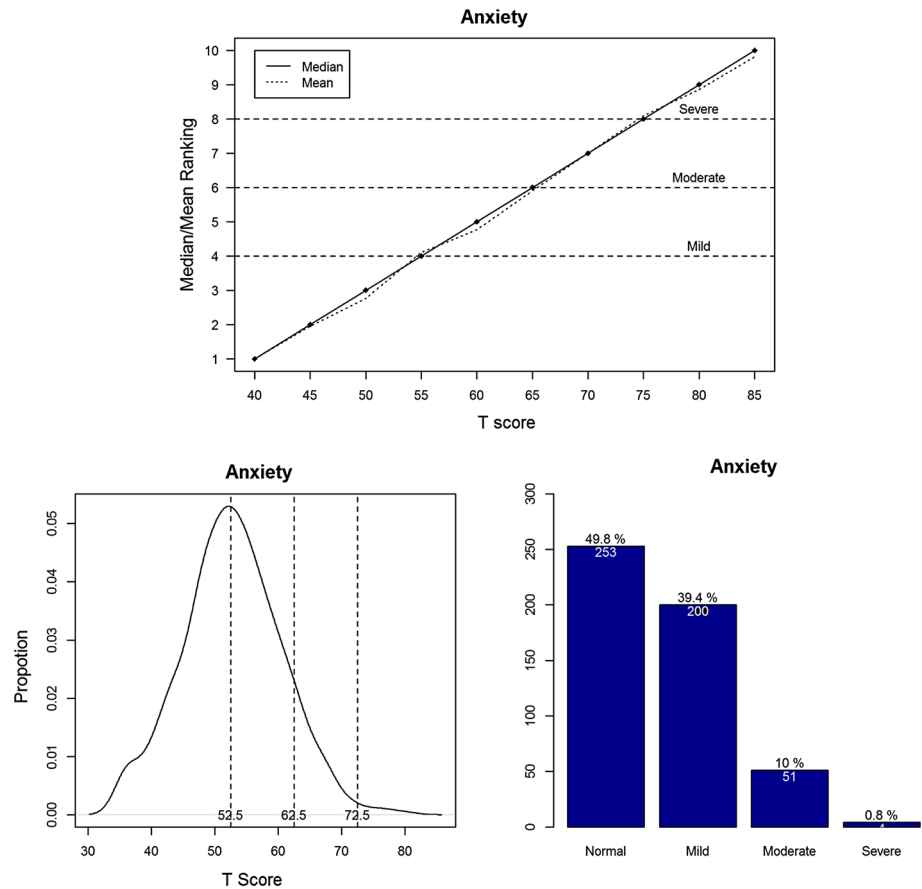
Small and large group discussions allowed for changes of vignette order or cut scores, but very few changes were made

during these discussions. The most animated conversations centered on where to draw the lines for normal, mild, moderate, and severe categories. After discussion, expert cut score consensus was achieved on all domains. Specific interventions to match severity levels were also reviewed and confirmed, but are not the focus of this manuscript.

Discussion

We describe a practical and generalizable method for deriving clinically relevant cut scores (“thresholds”) to distinguish mild, moderate, and severe symptom severity scores across the full range of the symptom measured by four PROMIS item banks. The thresholds we derived enable users of the PROMIS anxiety, depression, pain, and fatigue item banks to differentiate clinically resonant severity levels for diagnostic classification and treatment planning. We refer to the method, drawn from educational testing, as “standard setting,” because it emphasizes consensus-based expert clinician standards for symptom severity. Importantly, the vignettes used to derive these standards are generated based upon patient responses to PROMIS items. Ordering of expert clinician rankings of

Fig. 4 Top panel plots the vignette *T* score (*x* axis) against the median and mean card rankings according to expert consensus (*y* axis). Dotted horizontal lines reflect the expert consensus on bookmarks separating the severity of symptom vignettes (mild; moderate; severe). Experts were blind to vignette *T* score values throughout the exercise. Lower left panel displays the distribution of anxiety scores (*y* axis) by *T* score (*x* axis), with vertical lines separating clinical categories (none; mild; moderate; severe). Lower right panel indicates the number and proportion of patients in each of the four clinical categories



symptom vignettes was perfectly aligned with PROMIS *T* score levels across the full range of each symptom. This helps provide mutual confirmation of the clinical meaning of each symptom vignette, and helps provide a basis for defining severity thresholds based upon PROMIS *T* scores.

Because the standard setting exercise was based on intervals of five *T* score units (35, 40, 45, 50, etc.), precise cut scores are indicated at the midpoint of these anchored values (e.g., 52.5; 67.5). These are reflected in Figs. 2, 3, 4 and 5. For ease of clinical use, we set provisional severity thresholds proximal to the first score identified at the given symptom severity. This resulted in the following recommendations:

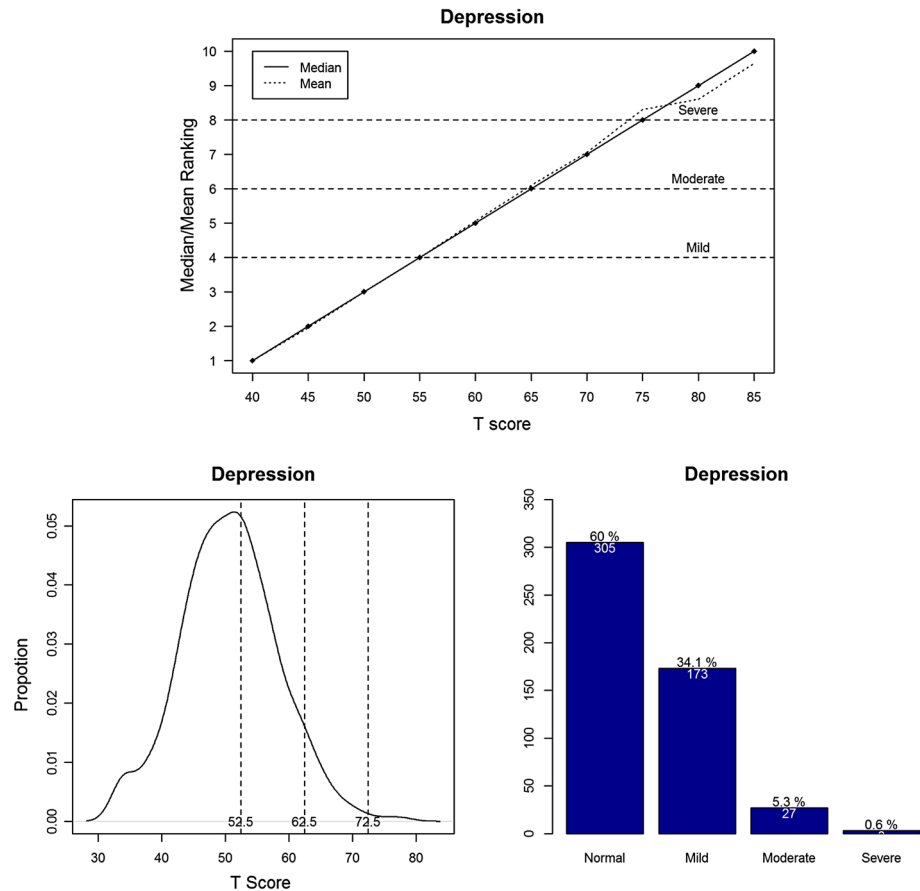
Pain	<50 normal; 50–59 mild; 60–69 moderate; ≥70 severe
Fatigue	<50 normal; 50–54 mild; 55–74 moderate; ≥75 severe
Anxiety	<55 normal; 55–64 mild; 65–74 moderate; ≥75 severe
Depression	<55 normal; 55–64 mild; 65–74 moderate; ≥75 severe

Efforts to derive clinically relevant thresholds for symptom severity have used a variety of methodologies, including the use of distribution-based statistics and

clinical anchors against which differences in the score can be compared to differences in the anchor [12–14, 16–18]. These “statistical” methods are useful demonstrations of validity. However, because they do not include expert judgment to assign clinical severity labels in the anchoring process, the assigned severity labels remain arbitrary. In addition, these earlier studies in symptom ratings rarely place the scores and the severity rankings on the same underlying scale, or metric, although this is done routinely in the educational setting. Surprisingly few studies have attempted to anchor PRO score differences to expert judgment regarding clinical severity on the concept being measured. One effort by King et al. [28] used experts to define score thresholds drawn from a meta-analysis of effect sizes based on all available evidence on a given PRO instrument. While clearly an advance, this methodology is labor intensive, sample dependent, and vulnerable to inter-rater unreliability based on varying familiarity with the PRO instrument and other factors.

Development of a method for identifying, labeling, and treatment planning around specified PRO scores would have great utility both for clinicians and researchers. Classification systems derived from such a method could be used to identify what constitutes clinically relevant score differences that distinguish intensity of interventions,

Fig. 5 *Top panel* plots the vignette *T* score (*x* axis) against the median and mean card rankings according to expert consensus (*y* axis). *Dotted horizontal lines* reflect the expert consensus on bookmarks separating the severity of symptom vignettes (mild; moderate; severe). Experts were blind to vignette *T* score values throughout the exercise. *Lower left panel* displays the distribution of depression scores (*y* axis) by *T* score (*x* axis), with *vertical lines* separating clinical categories (none; mild; moderate; severe). *Lower right panel* indicates the number and proportion of patients in each of the four symptom categories



and perhaps then setting standards for return to a “normal” range. Clinicians who have long experience using a given measure often develop an intuitive sense of what constitutes clinically relevant score differences. What is needed, however, is a method that draws from the rich experience of clinicians without requiring extensive familiarity with a particular PRO measure. This modified standard setting methodology, modified from educational testing, produced face valid and consensus-based thresholds for symptom severity, usually with little argument or debate. In rare instances when a clear line was not easily drawn, the expert compromise was to label these cases as “bridges” from one category to another. These transitional points require further investigation. One fruitful next step would be to bring patients themselves into the standard setting exercise, soliciting their opinion about severity based upon the patient perspective. Indeed, most clinical management decisions are made based upon patient report of symptom severity and degree of bother.

Limitations

This study had some notable limitations. For example, we opted to begin the exercise with individual judgments, and

then present the average of these judgments to small expert groups for reconciliation of any individual differences. Other methods to arrive at symptom severity rankings might have produced different results; however, we note that in all four cases, the final consensus ranking of symptom severity matched the ranking of the observed PROMIS score, to which all panelists were blinded. Another potential limitation was the restriction of items on each vignette card to five, rather than presenting the full bank. We cannot rule out the influence that a unique selection of five item/response clusters might shape rankings or severity ratings for a given symptom. One option is to create multiple vignettes for each trait level and randomly assign different vignettes to different raters. Use of multiple case packets per domain, employing different combinations of items, could have helped confirm these results.

We note that a five-point change in a *T* score, roughly equivalent to what has been considered a medium effect size, may be larger than what is a “minimally important” difference. In the middle of the score sample distributions, these broad *T* score ranges can have a big effect on the proportions of patients in a given category. For example, changing a *T* score threshold from 50 to 55 can have an

effect on prevalence of “mild” symptomatology as much as 20 %. Clearly, more work to “zero in” on valid cut points can help establish or modify the generalizability of these initial results.

Other potential limitations are that we employed only one group of experts. An independent group, or one with expertise in different conditions other than cancer, might have come up with different cut scores. Our experts were from a tertiary care facility and comprehensive cancer center comprised of fairly strong advocates for symptom management as an essential component of quality cancer care. This may have had the effect of lowering the thresholds for symptom severity, below a level that would be considered clinically relevant in the wider cancer treatment community. Finally, although we used patient response data to inform the symptom vignettes, our method did not include patients in the standard setting judgments themselves. It would be interesting, and important, to note whether or not patients share a common sense of symptom severity with the clinical experts who treat them. Perhaps, as mentioned, patient input can help settle expert difference of opinion when it occurs regarding degree of clinical severity of these common cancer symptoms. We encourage pursuit of this research in future efforts to engage in patient-centered outcomes research.

Conclusion

Here, we describe and report results of a standardized method for differentiating severity levels of patient-reported symptoms based on clinical judgment. The method was applied to PROMIS symptom item banks and produced logical clinical thresholds that retained tight comparability to actual patient report.

Acknowledgments Several colleagues contributed time and expertise to this effort, and we wish to acknowledge their effort. Clinical expert break-out group facilitators included some of the coauthors and also Nan Rothrock, PhD; and Zeeshan Butt, PhD. Expert raters included some of the coauthors and also Amy Peterman, PhD; Janine Gauthier, PhD; Lynne Wagner, PhD; Kimberly Davis, PhD; Margaret Whalen, RN; Gershon Locker, MD; Carmen Griza, MD; Jin-Shei Lai, PhD, OTR/L; Allen Heinemann, PhD; Lauren Mermel Welles, PT; Robin Mieli, MA, OTR/L; Kimberly Brennan, PT; Nan Rothrock, PhD; Zeeshan Butt, PhD; David Victorson, PhD; Judith Paice, PhD, RN, FAAN; Rose Catchatourian, MD; George Carro, PharmD; Mousami Shah, MD. Study coordination was done by Jacquelyn George and Rachel Hanrahan. Supported by a grant from the National Cancer Institute (R01 CA60068; Cella, PI).

References

- Carlson, L. E., Groff, S. L., Maciejewski, O., & Bultz, B. D. (2010). Screening for distress in lung and breast cancer outpatients: A randomized controlled trial. *Journal of Clinical Oncology*, 28(33), 4884–4891.
- Snyder, C., Jensen, R., Geller, G., Carducci, M., & Wu, A. (2010). Relevant content for a patient-reported outcomes questionnaire for use in oncology clinical practice: Putting doctors and patients on the same page. *Quality of Life Research*, 19(7), 1045–1055.
- Bennett, A. V., Jensen, R. E., & Basch, E. (2012). Electronic patient-reported outcome systems in oncology clinical practice. *CA: A Cancer Journal for Clinicians*, 62(5), 336–347.
- Abernethy, A. P., Herndon II, J. E., Wheeler, J. L., Day, J. M., Hood, L., Patwardhan, M., et al. (2009). Feasibility and acceptability to patients of a longitudinal system for evaluating cancer-related symptoms and quality of life: Pilot study of an e/tablet data-collection system in academic oncology. *Journal of Pain and Symptom Management*, 37(6), 1027–1038.
- Berry, D. L., Blumenstein, B. A., Halpenny, B., Wolpin, S., Fann, J. R., Austin-Seymour, M., et al. (2011). Enhancing patient-provider communication with the electronic self-report assessment for cancer: A randomized trial. *Journal of Clinical Oncology*, 29(8), 1029–1035.
- Basch, E., & Abernethy, A. P. (2011). Supporting clinical practice decisions with real-time patient-reported outcomes. *Journal of Clinical Oncology*, 29(8), 954–956.
- Cella, D., Gershon, R., Lai, J.-S., & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, 16(Suppl 1), 133–141.
- Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., et al. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap Cooperative Group during its first two years. *Medical Care*, 45(5 Suppl 1), S3–S11.
- Riley, W. T., Rothrock, N., Bruce, B., Christodolou, C., Cook, K., Hahn, E. A., et al. (2010). Patient-Reported Outcomes Measurement Information System (PROMIS) domain names and definitions revisions: further evaluation of content validity in IRT-derived item banks. *Quality of Life Research*, 19(9), 1311–1321.
- Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., et al. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, 63(11), 1179–1194.
- Liu, H., Cella, D., Gershon, R., Shen, J., Morales, L. S., Riley, W., et al. (2010). Representativeness of the PROMIS internet panel. *Journal of Clinical Epidemiology*, 63(11), 1169–1178.
- Jensen, M. P., Smith, D. G., Ehde, D. M., & Robinsin, L. R. (2001). Pain site and the effects of amputation pain: Further clarification of the meaning of mild, moderate, and severe pain. *Pain*, 91(3), 317–322.
- Paul, S. M., Zelman, D. C., Smith, M., & Miaskowski, C. (2005). Categorizing the severity of cancer pain: Further exploration of the establishment of cutpoints. *Pain*, 113(1), 37–44.
- Serlin, R. C., Mendoza, T. R., Nakamura, Y., Edwards, K. R., & Cleeland, C. S. (1995). When is cancer pain mild, moderate or severe? Grading pain severity by its interference with function. *Pain*, 61(2), 277–284.
- Zelman, D., Hoffman, D., Reed, P., Edwards, L., & Dukes, E. (2003). Quantifying a “day of manageable pain control:” The relationship between pain severity and patient daily functional expectations. *Journal of Pain*, 4(Suppl 1), 1.
- Zelman, D. C., Dukes, E., Brandenburg, N., Bostrom, A., & Gore, M. (2005). Identification of cut-points for mild, moderate and severe pain due to diabetic peripheral neuropathy. *Pain*, 115(1), 29–36.
- Wang, X. S., Zhao, F., Fisch, M. J., O’Mara, A. M., Cella, D., Mendoza, T. R., et al. (2014). Prevalence and characteristics of moderate to severe fatigue. *Cancer*, 120(3), 425–432.

18. Given, B., Sikorskii, A., Given, C. W., Jeon, S., McCorkle, R., Champion, V., et al. (2008). Establishing mild, moderate, and severe scores for cancer-related symptoms: How consistent and clinically meaningful are interference-based severity cut-points? *Journal of Pain and Symptom Management*, 35(2), 126–135.
19. Hays, R. D., Reise, S., & Calderón, J. L. (2012). How much is lost in using single items? *Journal of General Internal Medicine*, 27(11), 1402–1403.
20. Angoff, W. H., & Thorndike, R. L. (1971). Scales, norms and equivalent scores. In *Educational measurement*. Washington, D.C.: American Council on Education.
21. Lewis, D., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. Paper presented at the Council of Chief State School Officers National Conference on Large Scale Assessment Phoenix, AZ.
22. Impara, J. C., & Plake, B. S. (1977). Standard setting: An alternative approach. *Journal of Educational Measurement*, 34, 355–368.
23. Baghaei, P. (2007). Applying The Rasch Rating-Scale model to set multiple cut-offs. *Rasch Measurement Transactions*, 20(4), 1075–1076.
24. Lai, J. S., Cella, D., Choi, S. W., Junghaenel, D. U., Christodolou, C., Gershon, R., et al. (2011). How item banks and their application can influence measurement practice in rehabilitation medicine: A PROMIS fatigue item bank example. *Archives of Physical Medicine and Rehabilitation*, 92(10 Supplement), S20–S27.
25. Amtmann, D., Cook, K. F., Jensen, M. P., Chen, W.-H., Choi, S., Revicki, D., et al. (2010). Development of a PROMIS item bank to measure pain interference. *Pain*, 150(1), 173–182.
26. Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS): depression, anxiety, and anger. *Assessment*, 18(3), 263–283.
27. Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, 45(5 Suppl 1), S22–S31.
28. King, M. T., Cella, D., Osoba, D., Stockler, M., Eton, D., Thompson, J., et al. (2010). Meta-analysis provides evidence-based interpretation guidelines for the clinical significance of mean differences for the FACT-G, a cancer-specific quality of life questionnaire. *Patient Related Outcome Measures*, 1(1), 119–126.