Published in final edited form as:

J Clin Epidemiol. 2016 May ; 73: 89–102. doi:10.1016/j.jclinepi.2015.08.038.

PROMIS® measures of pain, fatigue, negative affect, physical function, and social function demonstrate clinical validity across a range of chronic conditions

Karon F. Cook[a,*], **Sally E. Jensen**[a,b], **Benjamin D. Schalet**[a], **Jennifer L. Beaumont**[a], **Dagmar Amtmann**[c], **Susan Czajkowski**[d], **Darren A. Dewalt**[e], **James F. Fries**[f], **Paul A. Pilkonis**[g], **Bryce B. Reeve**[h], **Arthur A. Stone**[i], **Kevin P. Weinfurt**[j], and **David Cella**[a]

[a]Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, 633 N. St. Clair, 19th Floor, Chicago, IL 60611, USA

[b]Department of Surgery (Division of Organ Transplant), Northwestern University Feinberg School of Medicine, 251 E. Huron, Galter 3-150, Chicago, IL, USA

[c]Department of Rehabilitation Medicine, University of Washington School of Medicine, 4907 25th Ave NE, Seattle, WA 98105, USA

[d]National Heart Lung and Blood Institute (NHLBI), Building 31, Room 5A52, 31 Center Drive MSC 2486, Bethesda, MD 20892, USA

[e]Division of General Internal Medicine and Cecil G. Sheps Center for Health Services Research, University of North Carolina School of Medicine, CB# 7590, 725 Martin Luther King Jr. Blvd., Chapel Hill, NC 27599-7590, USA

[f]Department of Medicine, Stanford University School of Medicine, 300 Pasteur Dr H3580, Stanford, CA 94305, USA

[g]Department of Psychiatry, University of Pittsburgh School of Medicine, 3811 O'Hara Street, Pittsburgh, PA 15213, USA

[h]Department of Health Policy and Management, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, CB# 7590, 725 Martin Luther King Jr. Blvd., Chapel Hill, NC 27599-7590, USA

[i]University of Southern California, 635 Downey Way, Los Angeles, CA 9008, USA

[j]Department of Psychiatry, Duke University School of Medicine, 2400 Pratt Street, Durham, NC 27705, USA

[*]Corresponding author. Tel.: 713-291-3918; fax: 312-503-4800. karon.cook@northwestern.edu (K.F. Cook).

Conflict of interest: K.F.C. is an unpaid officer of the PROMIS Health Organization; D.A.D. is an unpaid member of the board of directors of the PROMIS Health Organization; B.B.R. is an unpaid member of the board of directors of the PROMIS Health Organization. A.A.S. declares a potential conflict as Senior Scientist with the Gallup Organization and as a Senior Consultant with ERT, inc.; K.W. is an unpaid member of the board of directors of the PROMIS Health Organization; D.C. is an unpaid member of the board of directors and officer of the PROMIS Health Organization. All other authors declare no conflict of interest.

Supplementary data

Supplementary data related to this article can be found online at http://dx.doi.org/10.1016/j.jclinepi.2015.08.038.

Abstract

Objective—To present an overview of a series of studies in which the clinical validity of the National Institutes of Health's Patient Reported Outcome Measurement Information System (NIH; PROMIS) measures was evaluated, by domain, across six clinical populations.

Study Design and Setting—Approximately 1,500 individuals at baseline and 1,300 at follow-up completed PROMIS measures. The analyses reported in this issue were conducted post hoc, pooling data across six previous studies, and accommodating the different designs of the six, within-condition, parent studies. Changes in T-scores, standardized response means, and effect sizes were calculated in each study. When a parent study design allowed, known groups validity was calculated using a linear mixed model.

Results—The results provide substantial support for the clinical validity of nine PROMIS measures in a range of chronic conditions.

Conclusion—The cross-condition focus of the analyses provided a unique and multifaceted perspective on how PROMIS measures function in "real-world" clinical settings and provides external anchors that can support comparative effectiveness research. The current body of clinical validity evidence for the nine PROMIS measures indicates the success of NIH PROMIS in developing measures that are effective across a range of chronic conditions.

Keywords

Responsiveness; Validity; Psychometrics; Outcomes research; Patient-reported outcomes

1. Introduction

In a succinct seven words, Lee Sechrest summed up the formidable challenge that faces researchers who use and develop psychometric instruments: "Validity of measures is no simple matter." [1] Although researchers often describe a scale as "valid" or having been "validated," validity is contextual. It resides, not in the instrument itself, but in the use of the scores. A simple example clarifies the relationship between validity and appropriate use based on context. Consider a measure of depressive symptoms whose scores are found to successfully predict clinical depression. Such a finding supports the validity of using the measure's scores to screen individuals for depression; using scores of this measure to predict substance abuse, however, is unlikely to be as successful. Poor predictability of the scores in the latter case does not "invalidate" the measure any more than success in the former confers global validity to the measure.

In health measurement, among the challenges that make validity "no simple matter" is the application of measures in diverse populations and for a range of purposes. Scores are used to follow people over time, evaluate interventions, compare the effectiveness of treatments, and quantify the impact of disease on quality of life. To evaluate how effective measures are for these purposes, it is critical to administer them in clinical contexts and evaluate their performance. This is not to say that other considerations are irrelevant. The quality of an instrument is influenced by the soundness of the methods used to develop it, and evaluation of these methods is a critical component in assessment of a measure's validity. But the level

of validity evidence generated in the typical measurement development study often is rudimentary and inadequate for establishing the validity of using scores in clinical contexts [2].

For patient-reported outcome (PRO) measures, there are a number of particularly relevant validity evaluations. It is important to know how well scores on measures perform in quantifying the impact of disease and health problems on domains important to patients, in comparing effectiveness of treatments and management strategies, and in tracking the longitudinal course of disease. Subjecting well-constructed PRO tools to these critical tests of clinical validity is an essential step in the maturation of a new measure.

Well-constructed, generalizable, and clinically relevant PRO measures can be very useful when conducting comparative effectiveness research (CER). CER is defined as research "designed to inform health care decisions by providing evidence on the effectiveness, benefits, and harms of different treatment options" [3]. PRO scores increasingly serve as end points in treatment efficacy and effectiveness studies. They can be used to define responding (or progressing) patients in clinical trials. Magnitude of change in the PRO score of an individual that is required to classify as improved or worsened is specified a priori. Treatments are then compared with regard to differences in proportions of responders, progressors, or both. Responder analysis is appealing because it embeds meaningful change into the consideration of statistical significance. To conduct a responder analysis, however, one must answer the difficult question, "How should response to treatment be operationalized?" Statistical approaches have been critiqued because of the absence of an external anchor and the lack of consensus and empirical support [4]. A more patient-centered approach is to estimate meaningful change by anchoring to responses to a one-item global rating of change (GROC). However, the GROC has been criticized at several levels because of its vulnerability to response bias [5]. Clinical validity studies that evaluate changes in scores across different conditions and contexts could provide more defensible anchors for responder analysis, supporting the science of CER.

This article presents an overview of the five-article series published in this issue. [An additional article in this issue is devoted to examining the ecological validity of various Patient Reported Outcome Measurement Information System (PROMIS) measures across five different populations]. These publications document progress in building a body of clinical validity evidence for nine measures from the National Institutes of Health's (NIH) PROMIS [6–12]. Collectively, the findings substantially increase knowledge of the appropriate and meaningful applications of these PROMIS measures. In addition, they present an innovative approach in which measures are evaluated and compared across multiple chronic diseases and conditions. We further discuss how these findings may be used to support comparative effective research.

2. Background

During the first period of NIH PROMIS funding (2004–2009), several longitudinal studies were undertaken. Each was conducted in one of six clinical conditions: chronic heart failure (CHF), chronic obstructive pulmonary disease (COPD), rheumatoid arthritis (RA), cancer,

back pain, or major depression. These studies, the “parent studies” for this series of articles, addressed both substantive and psychometric research questions. For example, the back pain study evaluated the impact of spinal injections on individuals with back and/or leg pain and also investigated the responsiveness of PROMIS pain measures. The depression study evaluated the impact of standard treatments (medication, psychotherapy, or the combination of the two) in a sample of persons with clinical depression and also investigated the psychometric functioning of PROMIS measures of emotional distress (depression, anxiety, and anger). As these studies are published, their contributions will be of interest to the general measurement community and to researchers whose work is in one of the six clinical populations that were targeted.

We recognized a potential additional contribution for these studies. The data collected across the six clinical populations provided a unique opportunity to extend evaluation of the psychometric function of PROMIS measures by examining them, not within, but across clinical conditions. Psychometric evaluations, including the PROMIS studies described above, generally occur within a single clinical condition. When psychometric comparisons do occur, the comparisons typically are made among different measures of the same construct, but within a single clinical population. The tradition of conducting psychometric research within clinical research silos is a barrier to multiperspective evaluations that could be informative not only regarding the properties of measures, but also about the clinical character of different conditions. PROMIS is unique in that it is comprised of a team of investigators whose collective interests span a broad array of clinical conditions. Thus, the usual barrier to cross-condition investigations was removed.

With the combined data collected by the PROMIS studies, we seized the opportunity to conduct a set of “cross-cutting studies” that compared individual PROMIS measures across multiple clinical populations obtaining a multifaceted perspective on their clinical validity. The cross-cutting articles in this issue provide a unique perspective on the clinical validity of nine PROMIS measures.

3. Overview of studies

3.1. Organization and scope of studies

The cross-cutting articles in this series are organized by PRO domain rather than by clinical population. This is consistent with the PROMIS measurement philosophy that emphasizes domain-specific, rather than disease-specific measurement. Table 1 reports which domains and subdomains were measured and which corresponding PROMIS measures were administered to each of the six clinical samples. (PROMIS measures themselves are available in an online Appendix at www.jclinepi.com.) As evident in Table 1, a single domain may be represented by one or more subdomains as measured by PROMIS. In some of the clinical populations, all nine PROMIS domains covered in these articles were administered. In other samples, a subset was administered.

3.2. Cross-study sample characteristics

The individual cross-cutting articles report sample characteristics by PROMIS domain and measure. Table 2 is an omnibus presentation of the demographic and clinical characteristics of the combined data. This information is presented by clinical condition rather than by PROMIS domain.

3.3. Analytic approach

As already noted, for the current series of cross-cutting articles, we combined data obtained from separate studies completed within specific clinical populations. These parent studies differed in research design, analytic approach, and research questions. We developed an analytic plan that fit the needs of the current set of analyses and that incorporated the different research designs applied in the parent studies. Some of the parent PROMIS studies, but not all, included an intervention. For intervention studies, changes in PROMIS scores over time were evaluated with particular attention to results for more proximal outcomes (e.g., PROMIS depression scores in the depression study). When the parent study design allowed, differences between known groups (e.g., COPD stable and COPD exacerbation) were evaluated. Although the design of several of these studies (RA, back pain, COPD, and major depression) included multiple follow-up assessment points, the analyses reported in the cross-cutting articles were restricted to results from baseline and the latest follow-up point.

Essential to our ability to compare measures across different clinical population was selection of appropriate clinical anchors. We identified candidate anchors with input from investigators of the parent studies. To maximize consistency across the present studies, consensus was reached both for global anchors (e.g., the item, “Has there been any change in your overall health since you started the study?” from the COPD study) and domain-specific anchors (e.g., “How has your fatigue changed?” from the RA parent study). The set of anchors used across studies is reported in Table 3.

4. Summary of individual studies

4.1. Back pain

Recruitment and all procedures were approved by the University of Washington Institutional Review Board [13]. All participants provided informed consent.

4.1.1. Participants—A sample of 218 participants with back or leg pain was recruited from the University of Washington Spine Center at Harborview Medical Center (n = 131), Advanced Pain Medicine in Tacoma, Washington (n = 80), Virginia Mason Medical Center (n = 2), Roosevelt Medical Center (n = 2), Group Health Cooperative (n = 1), and from web site postings (n = 2). Inclusion criteria included having back and/or leg pain for at least 6 weeks and being scheduled for spinal injection (i.e., epidural steroid injection, facet joint injection, or sacroiliac joint injection). Patients were excluded if they had lumbar surgery within the last year or had unstable neurological symptoms, cauda equine syndrome, cancer, spinal cord injury; vertebral fractures; or multiple sclerosis.

Table 2 lists the demographic and clinical characteristics of the participants in this study. The sample's gender composition was 56% female and primarily white, non-Hispanic (85%). The median age group was 55–59 years. Seventy-three percent of participants reported that their worst back pain during the past 7 days was 8 or above on a scale from 0 to 10.

4.1.2. Measures—All nine PROMIS domains and subdomains were administered to the back pain sample at baseline and at 3 months posttreatment. Global ratings of change were collected using the question, "Compared to your first appointment, how would you rate your current back pain?" Responses of "much better" and "somewhat better" were collapsed to define "better"; "no change" was defined as "stable"; "somewhat worse" and "much worse" were collapsed to define "worse."

4.1.3. Clinical validity questions—By collecting the scores of back pain participants before and after an intervention that was expected to have a favorable clinical effect, we were able to evaluate the responsiveness of the PROMIS measures. Because the interventions targeted pain, the greatest responsiveness was expected in scores from PROMIS measures of pain behavior and pain interference. Because all PROMIS measures were administered, there was opportunity to evaluate the responsiveness of PROMIS scores on other quality of life domains more distal to the intervention's primary target of pain.

4.2. Cancer

Recruitment and procedures were approved by the North Shore University Health System's Institutional Review Board [14]. All participants provided informed consent.

4.2.1. Participants—Participants were 310 individuals with cancer recruited from North Shore University Health System. The sample included outpatients with any kind of cancer who were beginning a new cancer treatment. Table 2 displays participants' sociodemographic and clinical characteristics. The sample was predominantly female (61%) and white, non-Hispanic (81%). The median age group was 50–54 years.

4.2.2. Measures—PROMIS measures of physical function, fatigue, pain interference, depression, and anxiety were administered at enrollment in the study and at follow-up (6–12 weeks). In addition to PROMIS measures, participants rated their levels of change with respect to each outcome. Five such items were presented with the stem, "Since the last time you filled out a questionnaire, your level of [depression, anxiety, fatigue, physical function, pain] is…" Responses were coded 1 (very much better), 2 (moderately better), 3 (a little better), 4 (about the same), 5 (a little worse), 6 (moderately worse), and 7 (very much worse). Responses were combined as follows: 1–3 = "better"; 4 = "about the same"; and 5–7 = "worse."

4.2.3. Clinical validity questions—Cancer participants completed PROMIS measures before or at any stage of the cancer treatment. Clinical expectations for this group were less clear cut than for those clinical samples in which a specific intervention was assigned and administered. Given the heterogeneity of where a patient may be relative to treatment,

patients' score changes in the assessed domains of physical function, fatigue, pain interference, depression, and anxiety were expected to vary. Some participants would worsen due to disease progression or active treatment; some would be stable (unchanged); and others would improve after completing treatment. These improving and worsening patients would be identified by their individual global ratings of change asked at the second assessment. Based on prior work on the asymmetry of meaningful change, we predicted those patients who reported improvement in a domain would have lower effect sizes when compared to those who reported worsening in a domain [15,16]. Self-ratings of change allowed comparison of the correspondence between the magnitude of PROMIS score changes and participants' self-perceived changes in outcomes.

4.3. Chronic heart failure

Institutional Review Boards at the University of Pitts-burgh and Duke University approved all procedures, and all participants gave informed consent.

4.3.1. Participants—Patients ($N = 80$) who had severe CHF and were receiving a heart transplant were recruited from the medical centers at the University of Pittsburgh ($n = 60$) and Duke University ($n = 20$). Patients were eligible based on their placement on the heart transplant waiting list and the attending cardiologist's judgment that heart failure presented the greatest medical limitation on participant's daily function. After exclusion of participants who did not receive a transplant as scheduled, a total of 60 patients are included for analysis.

The sample was primarily male (80%) and white, non-Hispanic (86%). The median age group was 50–54 years. Ninety-eight percent of participants were classified as having "marked" or "severe" limitations (New York Heart Association Functional Classifications III–IV) [17].

4.3.2. Measures—PROMIS measures of physical function, fatigue, depression, and satisfaction with participation in discretionary social activities were administered to all patients. Because heart transplantation often produces substantial changes in cardiac functioning over a relatively short period, patients were assessed at two time points: at baseline, when patients were put on the heart transplant registry, and at 8–12 weeks following the transplant procedure. In addition to PROMIS measures, participants' perceptions of changes in outcomes were collected using four questions regarding the domains assessed with PROMIS measuresd—"How has your (depression, fatigue, ability to carry out your everyday physical activities, ability to carry out your usual social activities, and roles) changed since your heart transplant?" Responses of "got a lot better" and "got a little better" were collapsed to define the category "better"; "stayed the same" defined "same"; "got a little worse," and "got a lot worse" were collapsed to define "worse."

4.3.3. Clinical validity questions—The measurement of patients on PROMIS domains before and after heart transplant provided an opportunity to evaluate the responsiveness of PROMIS measures after an intervention with known clinical effectiveness. The clinical expectation was that, on average, participants would improve in the assessed domains of physical function, fatigue, depression, and anxiety. Statistically significant changes in the

expected direction would be evidence for the clinical validity of the targeted PROMIS measures. The postsurgery timing was selected as the minimum time following surgery at which a clinically significant improvement in functioning is typically observed.

4.4. Chronic obstructive pulmonary disease

The University of North Carolina, North Shore University Health System, University of Pittsburgh, and Duke University Institutional Review Boards approved the recruitment of participants and study procedures [18]. All participants provided informed consent.

4.4.1. Participants—A sample was recruited consisting of 185 persons with COPD. Participants met criteria for a clinical history of COPD according to the Global Initiative for Chronic Obstructive Lung Disease (Pauwels et al.) [19] definition and had at least a 10 pack/y history of smoking. Participants were considered stable if they had been exacerbation free for a minimum of 2 months before enrollment. Both stable and exacerbating participants were enrolled. Patients were recruited through a variety of clinics and hospitals at participating institutions (University of North Carolina, $n = 88$; North Shore University Health System, $n = 8$; University of Pittsburgh, $n = 47$; and Duke University, $n = 42$). Participating sites recruited patients through clinic visit logs, use of a COPD registry, and/or patients hospitalized for an acute exacerbation of COPD. Patients were approached by a clinician, their designee, or research assistant (as required by site-specific regulations), who provided an explanation of the study and obtained informed consent.

At baseline, slightly less than half of the sample ($n = 85$) was classified as COPD exacerbation, whereas the remainder ($n = 100$) was classified as COPD stable. Individuals developing a new exacerbation during the study period were censored at the time of exacerbation diagnosis, resulting in 79 patients who completed the study with confirmed stable COPD throughout and 46 patients who were experiencing an exacerbation at baseline and no subsequent exacerbations during the course of the study. The demographic characteristics of the COPD-stable and COPD-exacerbation groups did not differ substantially. Forty-four percent of the COPD-stable group and 39% of the COPD-exacerbation group were female, with both groups composed of primarily white, non-Hispanic participants (72% and 73%, respectively). The median COPD-stable age group was 60–64 years, and the median COPD-exacerbation age group was 55–59 years. At baseline, 19 percent of COPD-stable participants and 43% of COPD-exacerbation reported a Medical Research Council breathlessness rating of grade 4, “I stop for breath after walking 100 yards or after a few minutes on the level” or grade 5, “I am too breathless to leave the house” [20].

4.4.2. Measures—PROMIS measures of physical function, pain interference, pain behavior, depression, anxiety, anger, fatigue, satisfaction with participation in discretionary social activities, and satisfaction with participation in social roles were administered. Participants completed assessments at baseline, followed by weekly short assessments, and a 3-month comprehensive follow-up assessment that included a global assessment of change at follow-up. For the clinical validity questions addressed in this series of cross-cutting articles, the final 3-month assessment was compared to baseline.

Participants reported changes in general health by responding to the question, “Has there been any change in your overall health since you started the study?” using the following response options: 1 (“very much worse”), 2 (“moderately worse”), 3 (“a little worse”), 4 (“about the same”), 5 (“a little better”), 6 (“moderately better”), or 7 (“very much better”). At each time point, participants responded to global questions about the domains assessed by PROMIS measures. Scores on these items at baseline were subtracted from scores at follow-up to operationalize classification as worse, same, or better. The item used for defining changes in depression and anxiety was “How often have you been bothered by emotional problems such as feeling anxious, depressed or irritable?” To define changes in fatigue the item, “How would you rate your fatigue on average?” was administered. For physical function, the item administered was “To what extent are you able to carry out your everyday physical activities such as walking, climbing stairs, carrying groceries, or moving a chair?” For social function, the item was “In general, please rate how well you carry out your usual social activities and roles [etc.]” For pain interference and pain behavior, the item was “How would you rate your pain on average?” For depression, anxiety, pain, and fatigue, individuals who had higher follow-up scores than baselines scores were classified as “worse,” lower follow-up scores defined “better,” and equal values at baseline and followup defined “same.”

4.4.3. Clinical validity questions—The purpose of examining the PROMIS measures in this clinical setting was to evaluate the success of scores in discriminating persons with and without COPD exacerbation. The clinical expectation was that, on average, participants who had an exacerbation would have worse scores than persons without an exacerbation [21,22]. Because both stable and participants with an exacerbation were evaluated at two time points, the sensitivity of PROMIS measures over time was evaluated. The expectation was that those with an exacerbation at enrollment would improve in outcomes over the 3-month period, but there were no such expectations for those who began and completed the study classified as stable.

4.5. Rheumatoid arthritis

The Stanford University Institutional Review Board approved the study and informed consent was obtained for all participants [23,24].

4.5.1. Participants—PROMIS measures of physical function, fatigue, and pain interference were administered to 521 individuals with RA at baseline, 6 months, and 12 months. Measures evaluated longitudinal changes in a clinical population receiving routine care. All participants met the American College of Rheumatism criteria for RA. Our sample of patients was drawn from two sources: (1) the Aging Medical Information System and (2) the Stanford RA Registry, which provided patients who are enrolled in generic studies. Table 2 reports the demographics of the sample. Briefly, most were female (81%) and white, non-Hispanic (88%). The median age group was 65–69 years. Participants were seeing their rheumatologist for care and may or may not have been receiving a clinical intervention.

4.5.2. Measures—PROMIS measures of pain, physical function, and fatigue were administered to the RA sample at baseline and at a final follow-up of 12 months. In addition,

participants provided global ratings of change in outcomes by answering the questions: (1) How has your fatigue changed? (2) How has your ability to carry out your everyday physical activities such as walking, climbing stairs, carrying groceries, or moving a chair changed? And (3) How has your pain changed? Responses were 1–5. For all three domains, responses of 1 (“got a lot better”) and 2 (“got a little better”) were considered “better;” 3 (“stayed the same”) considered “the same;” and 4 (“got a little worse”) and 5 (“got a lot worse”) considered “worse.”

4.5.3. Clinical validity questions—The purpose of examining the PROMIS measures in this clinical setting was to evaluate the longitudinal trajectory in outcomes of persons with RA. Clinical expectations for this group were less clear cut than for those clinical samples in which a known intervention was assigned and administered. Although participants were receiving regular care that could include intervention, they also were experiencing a disease that is known to worsen over time. The inclusion of this sample allowed us to evaluate PROMIS measures in a heterogeneous chronic condition sample whose outcomes were thought to be more reflective of real-world RA populations not enrolled in clinical research trials. Self-ratings of change allowed comparison of the correspondence between the magnitude of score changes and participants’ self-perceived changes in outcomes.

4.6. Major depressive disorder

The University of Pittsburgh’s Institutional Review Board approved the study, and all participants provided informed consent [25].

4.6.1. Participants—A sample of 196 participants was recruited from outpatient treatment clinics at Western Psychiatric Institute and Clinic (WPIC, $n = 136$), Dubois Regional Medical Center Behavioral Health Services ($n = 29$), and private practitioners located in the Pittsburgh metropolitan area ($n = 31$). Participant recruitment procedures involved the distribution of IRB-approved study promotional materials at WPIC and its affiliates. The promotional materials encouraged potential participants to call or to visit the research staff, at which time an initial screening took place. Participants were excluded if they were undergoing current psychiatric inpatient treatment or had a lifetime history of any psychotic disorder (e.g., schizophrenia, schizoaffective disorder) or bipolar disorder documented in medical records or reported during the Structured Clinical Interview for DSM-IV [26]. Most patients received a combined treatment of medication and psychotherapy (64%), whereas smaller proportions received medication only (28%) or psychotherapy only (8%). Participants completed PRO measures at baseline and after 3 months.

Participant sociodemographic and clinical characteristics are described in Table 1. Most of the patients were female (74%) and white, non-Hispanic (78%). The median age group was 45–49 years. Forty-four percent of participants reported a Center for Epidemiologic Studies Depression Scale [27] score of 28 or above.

4.6.2. Measures—The following PROMIS measures were administered at baseline and at 3-month follow-up: physical function, fatigue, pain behavior, pain interference, depression, anxiety, anger, satisfaction with participation in social roles, and satisfaction with

participation in discretionary social activities. In addition, participants rated their change with respect to depression by answering the question, “Compared to your first appointment, how would you rate your current level of depression?” Responses ranged from −2 to +2. Responses of −2 and as −1 were coded as “worse,” 0, “same” and 1 and 2 as “better.” In addition, participants rated their general health by responding to the item, “Is your health excellent, very good, good, fair, or poor?” Positive and negative increments were coded as “better” and “worse,” respectively. No change was coded as “same.”

4.6.3. Clinical validity questions—The purpose of administering PROMIS measures in major depressive disorder (MDD) was to evaluate the responsiveness of PROMIS measures to MDD treatment. The clinical expectation was that participants’ scores would improve most on the most proximal subdomains of depression and anxiety.

5. Analyses

As reported above, different clinical samples lent themselves to different research questions, but all analyses contributed to understanding the clinical validity of the PROMIS domain measures. Another point of analytic continuity was the use of global and clinical anchors (such as perceived improvement in health) that were consistent within clinical groups and comparative across populations. Table 3 identifies the selected global and clinical anchors by clinical sample.

For each PROMIS instrument and each clinically anchored subgroup, the authors of each article report the change in T-scores and the standardized response mean (SRM). The SRM is the ratio of the mean change to the standard deviation of that change [28], which is a form of Cohen’s effect size index [29]. A minimally important difference in PRO measures was operationalized as an effect size of 0.30 [30,31]. PROMIS instruments use the convention of higher scores indicating more of the domain for which the measure is named.

Some studies (e.g., depression, back pain) included an intervention. For these studies, changes in PROMIS scores over time were evaluated with particular attention to results for more proximal outcomes (e.g., PROMIS pain scores in the back pain study, PROMIS depression scores in the depression study).

When appropriate, differences between known groups (e.g., COPD stable and COPD exacerbation) were evaluated. This was accomplished using linear mixed models estimated with random subject effects to account for the correlation among repeated observations within individuals [32,33]. A mixed model is advantageous because all available data can be used, including that from participants who did not provide data at both time points [34,35]. Least squares means, standard errors, and 95% confidence intervals were estimated from the models.

At baseline, completion rates for the individual PROMIS measures ranged from 95% to 100%, whereas 79–95% of study participants completed the measures at follow-up. We compared baseline characteristics of participants with complete data to those with incomplete data, pooling across conditions. There were some differences, indicating that missing data were not missing completely at random (MCAR). It is not possible to test for

the appropriateness of the missing at random (MAR) assumption compared to missing not at random. We judged the MAR assumption, conditional on the observed data (i.e., baseline scores), to be reasonable in this setting. The mixed effects models implemented in the analyses assume MAR (conditional on observed data in the model) and incorporate all available data. All analyses appropriate for MAR data are also appropriate for MCAR data. An alternative MAR strategy is multiple imputations; however, because baseline scores are historically the strongest predictor of follow-up scores, multiple imputations with additional variables in the model was not expected to improve substantially on the mixed model analyses.

6. Summary

This series of articles provides practical information on the responsiveness of several PROMIS domains across six clinical validation studies. Cumulatively, they report clinical validity findings for nine PROMIS measures, representing five PROMIS domains, evaluated across six clinical conditions, and including approximately 1,500 individuals at baseline and 1,300 at follow-up. The cross-condition focus of the analyses provides a unique and multifaceted perspective on how PROMIS measures function in “real-world” clinical settings. The results support the use of PROMIS measures across several chronic conditions.

Perhaps, the most practically useful research products in this series of studies are the graphs that show differences in PROMIS scores across time and across clinical condition and clinical subgroups. The graphs not only demonstrate the psychometric properties of the PROMIS measure being evaluated, but they express differences in clinical characteristics of the included chronic conditions. The graphs are succinct presentations of the validity of using scores from a given PROMIS measure for specific clinical questions.

Because the validity results for each PROMIS domain were based on multiple clinical samples, no single study can collect all the relevant clinical validity evidence for a given measure. The same observation can be made about any single series of studies. However, the cross-cutting studies reported in this issue provide substantial support for the clinical validity of nine PROMIS measures in a range of chronic conditions. We think the results have implications for the practical application of PROMIS tools. In addition, we think the analytic strategy is informative, practical, and innovative. We are hopeful that this unique, across-condition analytic approach will be repeated in future psychometric studies using both retrospective and prospective data.

6.1. Limitations

The analyses reported in this issue were conducted post hoc, pooling data across six distinct studies. We developed an analytic plan that accommodated the different designs of the six, within-condition, parent studies and maximized our ability to make relevant comparisons across conditions. A stronger approach would have been to develop, a priori, a data collection method, research design, and analytic approach that focused specifically on the purpose of the current set of studies—the cross-condition, psychometric evaluation of the PROMIS measures.

We used retrospective reports of change and differences in global scores from baseline to follow-up. Limiting the analyses to these anchors is a limitation of the studies. (A related issue is the fact that multiple cut-points could have been chosen to trichotomize change into "worse," "same," and "better.") Anchoring estimates of change based on global self-reports have intuitive appeal, but it also has limits. Retrospective, self-reported global change ratings have been shown to be vulnerable to response bias [36]; and they may be more related to current status than to change [5,37]. An ideal design would include estimates based on multiple anchors. For some conditions, external anchors such as clinician ratings, functional capacity tests, and work status could be incorporated. Provision of a range of estimates based on multiple anchors recognizes that there is no gold standard anchor, nor "true" estimate of meaningful change, a fact obscured when researchers report a measure's "minimally important difference" or "clinically meaningful difference" as a single score value. A more sound approach is to triangulate estimates of meaningful score changes to estimate a defensible range of estimates [38]. Although it is not possible to capture "the truth," there are ways to "get it surrounded." Toward this end, we recommend future studies that include a larger range of clinical anchors and contexts.

There is a few additional limitations of these studies. Although PROMIS was developed including numerous people across a variety of races and ethnicity, the participants in these studies are predominately non-Hispanic whites. In addition, many of the authors were PROMIS investigators who participated in the development of the measures being evaluated. The PROMIS process was extended and challenging, and it is likely that the investigators involve retain some "pride of ownership" that could influence their presentations and interpretations of results. This is an issue in any validity study conducted by a measure's developers. The authors recognize this potential influence, acknowledge it, and have attempted to minimize it.

6.2. Implications for CER

Despite the limitations in the studies in this issue, the results have significant potential for supporting CER. The observed score changes and score differences among clinical subpopulations could serve as defensible estimates of what constitutes clinically meaningful change and, by extension, provide empirical support for responder definitions. Again, we emphasize that there is no single, definitive score change that constitutes a meaningful or important change for all settings or for all purposes. But a strength of this study is its evaluation of the measures in a wide range of clinical populations, providing several perspectives for defining meaningful changes in scores and responders to interventions. For example, Schalet et al. (Schalet et al., 2014; submitted data) (this volume) report changes in PROMIS anxiety scores in four clinical populations including COPD. The mean PROMIS anxiety scores for individuals experiencing a COPD exacerbation was 60.2. After resolution, the mean dropped to 55.9—a difference of 4.3 points. This data point could be very informative for use in a clinical trial of a self-management intervention in COPD that is expected to reduce anxiety in participants. When used along with other available anchors and supported by distributional estimates, the information helps justify a responder definition.

In summary, the findings presented in cross-cutting studies in this issue provide unique and multifaceted perspectives on how PROMIS measures function in a range of clinical populations. The results support CER and add to the growing body of validity evidence for PROMIS measures used in a range of chronic conditions. Future studies can help define the strengths and limitations of the PROMIS measures by including additional clinical populations and a wider range of clinical anchors.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

NIH Science Officers on this project have included Deborah Ader, PhD, Vanessa Ameen, MD (deceased), Susan Czajkowski, PhD, Basil Eldadah, MD, PhD, Lawrence Fine, MD, DrPH, Lawrence Fox, MD, PhD, Lynne Haverkos, MD, MPH, Thomas Hilton, PhD, Laura Lee Johnson, PhD, Michael Kozak, PhD, Peter Lyster, PhD, Donald Mattison, MD, Claudia Moy, PhD, Louis Quatrano, PhD, Bryce Reeve, PhD, William Riley, PhD, Peter Scheidt, MD, Ashley Wilder Smith, PhD, MPH, Susana Serrate-Sztein, MD, William Phillip Tonkins, DrPH, Ellen Werner, PhD, Tisha Wiley, PhD, and James Witter, MD, PhD. The contents of this article use data developed under PROMIS. These contents do not necessarily represent an endorsement by the US Federal Government or PROMIS. See www.nihpromis.org for additional information on the PROMIS initiative.

Funding: PROMIS was funded with cooperative agreements from the National Institutes of Health (NIH) Common Fund Initiative (Northwestern University, PI: David Cella, PhD, U54AR057951, U01AR052177, R01CA60068; Northwestern University, PI: Richard C. Gershon, PhD, U54AR057943; American Institutes for Research, PI: Susan (San) D. Keller, PhD, U54AR057926; State University of New York, Stony Brook, PIs: Joan E. Broderick, PhD, and Arthur A. Stone, PhD, U01AR057948, U01AR052170; University of Washington, Seattle, PIs: Heidi M. Crane, MD, MPH, Paul K. Crane, MD, MPH, and Donald L. Patrick, PhD, U01AR057954; University of Washington, Seattle, PI: Dagmar Amtmann, PhD, U01AR052171; University of North Carolina, Chapel Hill, PI: Harry A. Guess, MD, PhD (deceased), Darren A. DeWalt, MD, MPH, U01AR052181; Children's Hospital of Philadelphia, PI: Christopher B. Forrest, MD, PhD, U01AR057956; Stanford University, PI: James F. Fries, MD, U01AR052158; Boston University, PIs: Alan Jette, PT, PhD, Stephen M. Haley, PhD (deceased), and David Scott Tulsky, PhD (University of Michigan, Ann Arbor), U01AR057929; University of California, Los Angeles, PIs: Dinesh Khanna, MD (University of Michigan, Ann Arbor), and Brennan Spiegel, MD, MSHS, U01AR057936; University of Pittsburgh, PI: Paul A. Pilkonis, PhD, U01AR052155; Georgetown University, PIs: Carol. M. Moinpour, PhD (Fred Hutchinson Cancer Research Center, Seattle), and Arnold L. Potosky, PhD, U01AR057971; Children's Hospital Medical Center, Cincinnati, PI: Esi M. Morgan DeWitt, MD, MSCE, U01AR057940; University of Maryland, Baltimore, PI: Lisa M. Shulman, MD, U01AR057967; and Duke University, PI: Kevin P. Weinfurt, PhD, U01AR052186).

References

1. Sechrest L. Validity of measures is no simple matter. Health Serv Res. 2005; 40:1584–604. [PubMed: 16178997]
2. Lohr KN, Aaronson NK, Alonso J, Audrey Burnam M, Patrick DL, Perrin EB, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. Clin Ther. 1996; 18:979–92. [PubMed: 8930436]
3. U.S. Department of Health and Human Services. What is comparative effectiveness research 2013: [2/11/2013]. Available at http://effectivehealthcare.ahrq.gov/index.cfm/what-is-comparative-effectiveness-research1/. Accessed March 10, 2016
4. Wyrwich KW, Norquist JM, Lenderking WR, Acaster S. Methods for interpreting change over time in patient-reported outcome measures. Qual Life Res. 2013; 22:475–83. [PubMed: 22528240]
5. Schmitt J, Di Fabio RP. The validity of prospective and retrospective global change criterion measures. Arch Phys Med Rehabil. 2005; 86:2270–6. [PubMed: 16344022]
6. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. 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. 2010; 63:1179–94. [PubMed: 20685078]

7. Liu H, Cella D, Gershon R, Shen J, Morales LS, Riley W, et al. Representativeness of the PROMIS Internet Panel. J Clin Epidemiol. 2010; 63:1169–78. [PubMed: 20688473]
8. Rothrock N, Hays R, 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. 2010; 63:1195–204. [PubMed: 20688471]
9. Magasi S, Ryan G, Revicki D, Lenderking W, Hays R, Brod M, et al. Content validity of patient-reported outcome measures: perspectives from a PROMIS meeting. Qual Life Res. 2012; 21:739–46. [PubMed: 21866374]
10. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The patient-reported outcomes measurement information System (PROMIS): progress of an NIH roadmap cooperative group during its first two years. Med Care. 2007; 45:S3–11.
11. DeWalt DA, Rothrock N, Yount S, Stone AA, PROMIS Cooperative Group. Evaluation of item candidates: the PROMIS qualitative item review. Med Care. 2007; 45:S12–21. [PubMed: 17443114]
12. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the patient-reported outcomes measurement information System (PROMIS). Med Care. 2007; 45:S22–31. [PubMed: 17443115]
13. Karp JF, Yu L, Friedly J, Amtmann D, Pilkonis PA. Negative affect and sleep disturbance are associated with response to epidural steroid injections for spine-related pain. Arch Phys Med Rehabil. 2014; 95:309–15. [PubMed: 24060493]
14. Yost KJ, Eton DT, Garcia SF, Cella D. Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-cancer scales in advanced-stage cancer patients. J Clin Epidemiol. 2011; 64:507–16. [PubMed: 21447427]
15. Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality-of-life scores: differences between improved and worsening. Qual Life Res. 2002; 11:207–21. [PubMed: 12074259]
16. Ringash J, Bezjak A, O'Sullivan B, Redelmeier D. Interpreting differences in quality of life: the FACT-H&N in laryngeal cancer patients. Qual Life Res. 2004; 13:725–33. [PubMed: 15129883]
17. New York Heart Association Criteria Committee. Dolgin, M. Nomenclature and criteria for diagnosis of diseases of the heart and great vessels. 9th. Boston: Little, Brown; 1994.
18. Dewalt, DA. Validation of PROMIS banks with COPD exacerbations 2012:[cited 2013 October 29]. Available at http://clinicaltrials.gov/ct2/show/NCT00784342?term5COPD+AND+dewalt&rank=2. Accessed March 10, 2016
19. Pauwels RA, Buist AS, Calverley PM, Jenkins CR, Hurd SS. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop summary. Am J Respir Crit Care Med. 2001; 163:1256–76. [PubMed: 11316667]
20. Fletcher C. Standardised questionnaire on respiratory symptoms: a statement prepared and approved by the MRC Committee on the Aetiology of Chronic Bronchitis (MRC breathlessness score). Br Med J. 1960; 2:1665. [PubMed: 13688719]
21. Andersson I, Johansson K, Larsson S, Pehrsson K. Long-term oxygen therapy and quality of life in elderly patients hospitalised due to severe exacerbation of COPD. A 1 year follow-up study. Respir Med. 2002; 96(11):944–9. [PubMed: 12418593]
22. Doll H, Duprat-Lomon I, Ammerman E, Sagnier P-P. Validity of the St George's respiratory questionnaire at acute exacerbation of chronic bronchitis: comparison with the Nottingham health profile. Qual Life Res. 2003; 12:117–32. [PubMed: 12639059]
23. Fries JF, Krishnan E, Rose M, Lingala B, Bruce B. Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. Arthritis Res Ther. 2011; 13(5):R147. [PubMed: 21914216]
24. Hays RD, Spritzer KL, Fries JF, Krishnan E. Responsiveness and minimally important difference for the Patient-Reported Outcomes Measurement Information System (PROMIS) 20-item physical

functioning short form in a prospective observational study of rheumatoid arthritis. Ann Rheum Dis. 2015; 74(1):104–7. [PubMed: 24095937]

25. Pilkonis, P. Validating PROMIS instruments in depression 2013:[cited 2013 October 29]. Available at http://clinicaltrials.gov/ct2/show/NCT00784199?term=Major+Depressive+Disorder+AND+PROMIS&rank=1. Accessed March 10, 2016
26. First, MB.; Spitzer, RL.; Gibbon, M.; Williams, JBW. Structured clinical interview for DSM-IV axis I disorders: Patient Edition SCID-I/P 2ed. Washington, DC: American Psychiatric Press; Biometrics Research Dept., New York State Psychiatric Institutute; 1997.
27. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas. 1977; 1:385–401.
28. Fayers, PM.; Machin, D. Quality of life: the assessment, analysis and interpretation of patient-reported outcomes. 2. Chichester, England; Hoboken, NJ: John Wiley & Sons; 2007.
29. Cohen, J. Statistical power analysis for the behavioral sciences. 2nd. Hillsdale, N.J.: L. Erlbaum Associates; 1988.
30. Revicki DA, Cella D, Hays RD, Sloan JA, Lenderking WR, Aaronson NK. Responsiveness and minimal important differences for patient reported outcomes. Health Qual Life Outcomes. 2006; 4:1–5. [PubMed: 16393335]
31. Yost KJ, Eton DT. Combining distribution-and anchor-based approaches to determine minimally important differences: the FACIT experience. Eval Health Prof. 2005; 28(2):172–91. [PubMed: 15851772]
32. Verbeke, G.; Molenberghs, G. Linear mixed models for longitudinal data. New York, NY: Springer-Verlag; 2000.
33. Hedeker, DR.; Gibbons, RD. Longitudinal data analysis. Hoboken, N.J.: Wiley-Interscience; 2006.
34. Troxel AB, Fairclough DL, Curran D, Hahn EA. Statistical analysis of quality of life with missing data in cancer clinical trials. Stat Med. 1998; 17:653–66. [PubMed: 9549814]
35. Little, RJA.; Rubin, DB. Statistical analysis with missing data. Hoboken, NJ: John Wiley & Sons, Inc; 2002.
36. Ross M. Relation of implicit theories to the construction of personal histories. Psychol Rev. 1989; 96:341.
37. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol. 1997; 50:869–79. [PubMed: 9291871]
38. 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; 61:102–9. [PubMed: 18177782]

What is new?

Key findings

- There is substantial clinical validity evidence that National Institutes of Health's Patient Reported Outcome Measurement Information System was successful in developing measures that are effective across a range of chronic conditions.

What this adds to what was known?

- The results add to what is known about the properties of the targeted PROMIS measures.
- The weight of evidence supports the appropriateness of using PROMIS measures across varied clinical populations.

What is the implication and what should change now?

- Findings support the fitness of PROMIS measures in clinical and comparative effectiveness research.
- Results should be considered when establishing responder criteria for comparative effectiveness research.

Table 1

Domains and subdomains measured by clinical population

PROMIS framework			Clinical populations					
Component	**Domain**	**Subdomain**	**Chronic obstructive pulmonary disease**	**Back pain**	**Major depressive disorder**	**Chronic heart failure**	**Rheumatoid arthritis**	**Cancer**
Physical Health	*Physical function*	–	✓	✓	✓	✓	✓	✓
	Fatigue	–	✓	✓	✓	✓	✓	✓
	Pain	*Pain behavior*	✓	✓	✓			
		Pain interference	✓	✓	✓		✓	✓
Mental health	Negative affect	*Depression*	✓	✓	✓	✓		✓
		Anxiety	✓	✓	✓			✓
		Anger	✓	✓	✓			
Social health	Satisfaction with participation in social roles and activities	*Satisfaction with participation in discretionary social activities*	✓	✓	✓	✓		
		Satisfaction with participation in social roles	✓	✓	✓			

Abbreviation: PROMIS, Patient Reported Outcome Measurement Information System.

Domains and subdomains in italics represent the concept that was measured by the corresponding instrument in each study. The domains and subdomains concepts were measured with short form or CAT instruments, as described in each content article in this issue. With the exception of social health, all short form and CAT items belong to the version 1.0 item banks. The social health items, however, were from version 2.0 item banks. See the online Appendix at www.jclinepi.com for details on the PROMIS instruments.

J Clin Epidemiol. Author manuscript; available in PMC 2016 December 01.

Table 2

PROMIS clinical validation study groups, demographics, and clinical characteristics

	Back pain (*n* = 218)	Cancer (*n* = 310)	Chronic heart failure (*n* = 60)	Chronic obstructive pulmonary disease, exacerbation (*n* = 46)	Chronic obstructive pulmonary disease, stable (*n* = 79)	Major depressive disorder (*n* = 196)	Rheumatoid arthritis (*n* = 521)
Female gender	121 (56%)	189 (61%)	12 (20%)	18 (39%)	35 (44%)	145 (74%)	422 (81%)
Age category, y							
18–29	17 (8%)	3 (1%)	2 (4%)	0	0	47 (24%)	2 (>1%)
30–39	17 (8%)	8 (3%)	5 (10%)	0	0	25 (13%)	18 (3%)
40–49	48 (22%)	59 (19%)	10 (20%)	4 (9%)	1 (1%)	30 (15%)	39 (7%)
50–59	55 (25%)	90 (29%)	16 (33%)	23 (50%)	24 (30%)	13 (7%)	114 (22%)
60–69	43 (20%)	69 (23%)	14 (29%)	10 (22%)	27 (34%)	56 (29%)	164 (31%)
70+	38 (17%)	77 (25%)	2 (4%)	9 (20%)	27 (34%)	24 (12%)	184 (35%)
Missing	0	4	11	0	0	1	0
Median age group	55–59	50–54	50–54	55–59	60–64	45–49	65–69
Race, ethnicity							
White, non-Hispanic	183 (84%)	243 (81%)	50 (86%)	33 (73%)	56 (72%)	145 (78%)	454 (88%)
Black, non-Hispanic	7 (3%)	33 (11%)	7 (12%)	12 (27%)	19 (24%)	26 (14%)	19 (4%)
Other, non-Hispanic	15 (7%)	10 (3%)	1 (2%)	0	2 (3%)	4 (2%)	15 (3%)
Hispanic, any race	10 (5%)	14 (5%)	0	0	1 (1%)	10 (5%)	27 (5%)
Missing	3	10	2	1	1	11	6
ECOG PSR							
0		102 (33%)					
1		135 (44%)					
2		47 (15%)					
3		21 (7%)					
Missing		5					
NYHA classification							
II (slight limitation)			1 (2%)				
III (marked limitation)			28 (50%)				
IV (severe limitation)			27 (48%)				
Missing			4				

	Back pain (*n* = 218)	Cancer (*n* = 310)	Chronic heart failure (*n* = 60)	Chronic obstructive pulmonary disease, exacerbation (*n* = 46)	Chronic obstructive pulmonary disease, stable (*n* = 79)	Major depressive disorder (*n* = 196)	Rheumatoid arthritis (*n* = 521)
COPD state							
Stable				0	79 (100%)		
Exacerbation resolving				8 (17%)	0		
Exacerbation current				38 (83%)	0		
MRC breathlessness rating							
1 (not breathless)				2 (4%)	14 (18%)		
2				15 (33%)	30 (38%)		
3				9 (20%)	20 (25%)		
4				14 (30%)	15 (19%)		
5 (too breathless to leave the house)				6 (13%)	0		
Worst back pain, past 7 days							
0–3	8 (4%)					91 (47%)	
4–6	25 (12%)					37 (19%)	
7	24 (11%)					12 (6%)	
8	49 (23%)					21 (11%)	
9	49 (23%)					17 (9%)	
10	60 (28%)					17 (9%)	
Missing	3					1	
Worst leg pain, past 7 days							
0–3	45 (21%)					116 (59%)	
4–6	36 (17%)					21 (11%)	
7	20 (9%)					13 (7%)	
8	41 (19%)					19 (10%)	
9	34 (16%)					10 (5%)	
10	40 (19%)					16 (8%)	
Missing	2					1	
CES-D score							
<16 (not depressed)	81 (38%)					18 (9%)	
16–21	60 (28%)					36 (18%)	
22–27	33 (15%)					56 (29%)	

J Clin Epidemiol. Author manuscript; available in PMC 2016 December 01.

	Back pain (*n* = 218)	Cancer (*n* = 310)	Chronic heart failure (*n* = 60)	Chronic obstructive pulmonary disease, exacerbation (*n* = 46)	Chronic obstructive pulmonary disease, stable (*n* = 79)	Major depressive disorder (*n* = 196)	Rheumatoid arthritis (*n* = 521)
28–34	20 (9%)					50 (26%)	
35+	19 (9%)					35 (18%)	
Missing	5					1	
HAQ Disability Index[a]							
Mean (SD)							0.90 (0.73)
Median (range)							0.75 (0–3)
0–1							295 (57%)
1–2							181 (35%)
2–3							45 (9%)

Abbreviations: PROMIS, Patient Reported Outcome Measurement Information System; ECOG PSR, Eastern Cooperative Oncology Group Performance Status Rating; NYHA, New York Heart Association; COPD, chronic obstructive pulmonary disease; MRC, Medical Research Council; CES-D, Center for Epidemiological Studies Depression Scale; HAQ, Health Assessment Questionnaire; SD, standard deviation.

[a]Scoring was not adjusted for aids and devices.

Table 3

Global and domain-specific anchors used across studies

	Back pain study	Cancer study	CHF study	COPD study	Depression study	Rheumatoid arthritis
Global anchors:						
	Is your health…? [excellent, very good, good, fair, poor]	In general, would you say your health is…?[excellent, very good, good, fair, poor]	How has your overall health changed since your heart transplant? [got a lot better, got a little better, stayed the same, got a little worse, got a lot worse]	Has there been any change in your overall health since you started the study? [very much better, moderately better, a little better, about the same, a little worse, moderately worse, very much worse]	Is your health…? [excellent, very good, good, fair, poor]	In general, would you say your current health is…? [excellent, very good, good, fair, poor]
Instrument						
Domain-specific anchors:						
Anger				How often have you been bothered by emotional problems such as feeling anxious, depressed or irritable? [never, rarely, sometimes, often, always]		
Anxiety		Since the last time you filled out a questionnaire, your level of anxiety is…[very much better, moderately better, a little better, about the same, a little worse, moderately worse, very much worse]		How often have you been bothered by emotional problems such as feeling anxious, depressed or irritable? [never, rarely, sometimes, often, always]		
Depression	Compared to your first appointment… How would you rate your current level of depression? [much better, somewhat better, no change, somewhat worse, much worse]	Since the last time you filled out a questionnaire, your level of depression is…[very much better, moderately better, a little better, about the same, a little worse, moderately worse, very much worse]	How has your depression changed since your heart transplant? [got a lot better, got a little better, stayed the same, got a little worse, got a lot worse]	How often have you been bothered by emotional problems such as feeling anxious, depressed or irritable?[never, rarely, sometimes, often, always]	Compared to your first appointment… How would you rate your current level of depression? [much better, somewhat better, no change, somewhat worse, much worse]	
Fatigue		Since the last time you filled out a questionnaire, your level of fatigue is…[very much better, moderately better, a little better, about the same, a little worse, moderately worse, very much worse]	How has your fatigue changed since your heart transplant? [got a lot better, got a little better, stayed the same, got a little worse, got a lot worse]	How would you rate your fatigue on average? [none, mild, moderate, severe, very severe]		How has your fatigue changed? [got a lot better, got a little better, stayed the same, got a little worse, Got a lot worse]

J Clin Epidemiol. Author manuscript; available in PMC 2016 December 01.

	Back pain study	Cancer study	CHF study	COPD study	Depression study	Rheumatoid arthritis
Physical function		Since the last time you filled out a questionnaire, your physical function is…[very much better, moderately better, a little better, about the same, a little worse, moderately worse, very much worse]	How has your ability to carry out your everyday physical activities changed since your heart transplant? [got a lot better, got a little better, stayed the same, got a little worse, got a lot worse]	To what extent are you able to carry out your everyday physical activities such as walking, climbing stairs, carrying groceries, or moving a chair? [completely, mostly, moderately, a little, not at all]		How has your ability to carry out your everyday physical activities such as walking, climbing stairs, carrying groceries, or moving a chair changed? [got a lot better, got a little better, stayed the same, got a little worse, got a lot worse]
Pain	Compared to your first appointment, how would you rate your current back pain? [much better, somewhat better, no change, somewhat worse, much worse]	Since the last time you filled out a questionnaire, your level of pain is…[very much better, moderately better, a little better, about the same, a little worse, moderately worse, very much worse]		How would you rate your pain on average? [0 to 10]	When you had back pain, how would you rate your average pain? [0 to 10]	How has your pain changed? [got a lot better, got a little better, stayed the same, got a little worse, got a lot worse]
Social			How has your ability to carry out your usual social activities and roles changed since your heart transplant? [got a lot better, got a little better, stayed the same, got a little worse, got a lot worse]	In general, how would you rate your satisfaction with your social activities and relationships? [excellent, very good, good, fair, poor]		

Abbreviations: CHF, chronic heart failure; COPD, chronic obstructive pulmonary disease.

For consistency of presentation, the order of response category options are reversed from the direction they were presented to patients.