Title

Reliability and construct validity of PROMIS® measures for patients with heart failure who undergo heart transplant

Journal

Quality of Life Research, 24(11)

ISSN

0962-9343

Authors

Flynn, KE
Dew, MA
Lin, L
et al.

Publication Date

2015-11-01

DOI

10.1007/s11136-015-1010-y

Peer reviewed

Qual Life Res (2015) 24:2591–2599
DOI 10.1007/s11136-015-1010-y

Reliability and construct validity of PROMIS® measures for patients with heart failure who undergo heart transplant

Kathryn E. Flynn[1] · Mary Amanda Dew[2,3,4] · Li Lin[6] · Maria Fawzy[6] · Felicia L. Graham[6] · Elizabeth A. Hahn[8] · Ron D. Hays[9] · Robert L. Kormos[5] · Honghu Liu[9] · Mary McNulty[2] · Kevin P. Weinfurt[6,7]

Accepted: 4 May 2015 / Published online: 3 June 2015

Abstract
Purpose To evaluate the reliability and construct validity of measures from the Patient-Reported Outcomes Measurement Information System® (PROMIS®) for patients with heart failure before and after heart transplantation.
Methods We assessed reliability of the PROMIS short forms using Cronbach's alpha and the average marginal reliability. To assess the construct validity of PROMIS computerized adaptive tests and short-form measures, we calculated Pearson product moment correlations between PROMIS measures of physical function, fatigue, depression, and social function and existing PRO measures of similar domains (i.e., convergent validity) as well as different domains (i.e., discriminate validity) in patients with heart failure awaiting heart transplant. We evaluated the responsiveness of these measures to change after heart transplant using effect sizes.
Results Forty-eight patients were included in the analyses. Across the many domains examined, correlations between conceptually similar domains were larger than correlations between different domains of health, demonstrating construct validity. Health status improved substantially after heart transplant (standardized effect sizes, 0.63–1.24), demonstrating the responsiveness of the PROMIS measures. Scores from the computerized adaptive tests and the short forms were similar.
Conclusions This study provides evidence for the reliability and construct validity (including responsiveness to change) of four PROMIS domains in patients with heart failure before and after heart transplant. PROMIS measures are a reasonable choice in this context and will facilitate comparisons across studies and health conditions.

Keywords Congestive heart failure · Outcomes research · Patient-reported outcomes

Electronic supplementary material The online version of this article (doi:10.1007/s11136-015-1010-y) contains supplementary material, which is available to authorized users.

✉ Kevin P. Weinfurt
kevin.weinfurt@duke.edu

[1] Center for Patient Care and Outcomes Research, Department of Medicine, Medical College of Wisconsin, Milwaukee, WI, USA

[2] Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA

[3] Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA

[4] Department of Epidemiology and Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA

[5] Department of Surgery, University of Pittsburgh, Pittsburgh, PA, USA

[6] Duke Clinical Research Institute, Duke University School of Medicine, Box 17969, Durham, NC 27715, USA

[7] Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine, Box 17969, Durham, NC 27715, USA

[8] Department of Medical Social Sciences and Center for Patient-Centered Outcomes, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

[9] Department of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

Introduction

Heart failure is a common, chronic, and life-threatening condition associated with fatigue, dyspnea, and depression [1, 2]. Health status for patients with heart failure is routinely measured using physician estimation of patient function, New York Heart Association (NYHA) classification, exercise capacity, echocardiograms, and laboratory measures like B-type natriuretic peptide level and other biomarkers. Patient-reported outcome (PRO) measures are an important complement to these clinical indicators and are a key metric of cardiovascular health [3].

Several PRO measures are available for assessment of the disease-specific effects of heart failure [4], including the Minnesota Living With Heart Failure Questionnaire (MLHFQ) [5] and the Kansas City Cardiomyopathy Questionnaire (KCCQ) [6]. Such disease-specific measures may be more sensitive to changes in health than generic measures. For example, the MLHFQ and the KCCQ were found to be more responsive than the SF-12 to clinically important changes in heart failure [11, 12].

As a complement to disease-specific measures, generic health status instruments can facilitate comparisons of disease burden and treatment effectiveness across diseases. They may also be preferable for evaluating health status in patients who have multiple health conditions, in that they do not ask patients to attribute their symptoms or function to a single health condition. The National Institutes of Health (NIH) Patient-Reported Outcomes Measurement Information System® (PROMIS®) has developed PRO measures designed for use across multiple chronic diseases. PROMIS utilizes modern psychometric methods to enhance assessment and scoring of generic health-related quality of life. Unlike older generic measures, PROMIS measures can be assessed using computerized adaptive tests (CAT), which customize the items a participant sees by choosing each successive item based on the participant's response to the preceding item. This can result in a substantial reduction in respondent burden. Another advantage of the PROMIS measures is that they provide scores based on a common metric, normed to the US general population of adults.

The validity of the PROMIS measures has been evaluated in patients with inflammatory bowel disease [7], arthritis [8–10], and cancer [11], and among others, as well as in general US populations [12–14]. The purpose of this study was to provide evidence about the reliability and construct validity (including responsiveness to change) of the PROMIS measures in patients with heart failure who undergo heart transplant.

Materials and methods

Sample

Participants were recruited at the Duke University Medical Center and the University of Pittsburgh Medical Center. We recruited candidates for heart transplant who had United Network for Organ Sharing (UNOS) status 1A, 1B, or 2. Eligible patients were 18 years or older, were able to speak English, were able to provide informed consent, did not have a current diagnosis of psychosis or dementia, and were actively listed on the Organ Procurement and Transplantation Network heart transplant list. We reviewed participants' medical records at each participating site after the patients consented to participate in the study to confirm that the inclusion criteria were met. Participants received compensation of $80 in the form of a gift card for each completed assessment. The institutional review boards of the Duke University Health System and the University of Pittsburgh Medical Center approved this study, and all patients provided written consent to participate.

Procedures

Participants completed all study assessments both before heart transplant (i.e., "baseline," including any time after the patient was listed for transplant) and after transplant (i.e., "follow-up," 8–12 weeks after surgery, deemed by cardiologist coinvestigators as the minimum time after transplant at which a clinically significant improvement in functioning is typically observed). Patients had the option at each time point of completing the assessments by computer-assisted telephone interview or by themselves using a computer.

Measures

We collected baseline characteristics through both patient self-report and medical record review. We measured four domains expected to change after heart transplant: physical functioning, fatigue, satisfaction with discretionary social activities, and depression. For each PROMIS domain, a higher score represents more of that domain content (e.g., higher physical functioning scores reflect better physical functioning; higher fatigue scores reflect greater fatigue). PROMIS domain scores are expressed as T scores, for which a score of 50 corresponds to the US general population average with an SD of 10.

For each PROMIS domain, we first administered a CAT from the PROMIS version 1.0 item banks [14]. We used

the default settings on Assessment Center for PROMIS adult banks, which specify that at least 4–12 items be administered per domain and that CAT administration stop when the standard error of the estimated T score is 3.0 or lower (reliability = 0.91 or higher). We were also interested in evaluating the validity of the domains as measured by the PROMIS short forms, which are fixed-length measures of each domain that consist of items covering the full range of functioning. Accordingly, we administered short forms of each domain (physical function 10a, fatigue 7a, depression 8b, and satisfaction with discretionary social activities 7a). So as not to ask patients to answer the exact same questions twice, we excluded questions patients had just answered as part of the CAT. All PROMIS measures are available in Assessment Center[SM] (www.assessmentcenter.net).

As our goal was to understand selected PROMIS measures' validity in this population of patients, we also included selected items and subscales from psychometrically sound and commonly used extant questionnaires that are intended to measure similar constructs as the selected PROMIS domains. We used items and subscales from extant measures rather than whole measures, because (1) we wished to reduce subject burden and (2) our intent was to determine whether PROMIS measures yielded results that were consistent with conceptually similar items/subscales of well-accepted measures.

The KCCQ is a 23-item questionnaire designed to measure several important aspects of heart failure [5]. The KCCQ scales are scored from 0 to 100 (higher scores = better health status). For this study, we administered the KCCQ physical limitation subscale, the social limitation subscale, and the two fatigue items from the symptoms subscale.

The Medical Outcomes Study Short Form (SF-36 v.1) vitality scale is a 4-item subscale that measures how fatigued or energetic a person feels [15, 16]. We scored the scale using a T score metric with a mean of 50 and an SD of 10 in the US general population.

The 2-item Patient Health Questionnaire depression module (PHQ-2) is used as a screen for the presence of major depression, with questions about the frequency of depressed mood and anhedonia over the past 2 weeks [17]. A PHQ-2 score ranges from 0 to 6, with higher scores representing greater depression.

Finally, we collected data on clinical-based functional measures. Treating physicians recorded each patient's NYHA class, which we abstracted from the medical record. Patients performed a 6-min walk test [18] specifically for this study. In accordance with the test guidelines [19], a patient who was too sick to walk was assigned a 6-min walk distance of 0.

Hypotheses

Based on previous studies, we expected significant improvements after heart transplant in physical functioning, fatigue, depressive symptoms, and social functioning [20–22]. We hypothesized larger correlations between different measures of the same domain (such as fatigue measured by PROMIS and fatigue measured by the KCCQ) than correlations between different domains measured either with the same instrument (such as fatigue and depression measured by PROMIS) or with different instruments (such as fatigue measured by PROMIS and depression measured by the PHQ-2).

Statistical analysis

We summarized the data using means and SDs for continuous variables and frequencies and percentages for discrete variables. We used a 1-sample t test to compare PROMIS scores in the study sample to the US normative mean of 50. We computed reliability for the PROMIS short forms using Cronbach's alpha and the average marginal reliability. In IRT, the reliability of scores varies depending on the severity of the score. The average marginal reliability is the average reliability across all of the patients included in this study. We considered reliability of 0.70 or greater to be acceptable [23]. To evaluate construct validity, we calculated Pearson product moment correlations between PROMIS scores and their corresponding PRO or clinical measures for the baseline and follow-up assessments, as well as for the change from baseline. We did not calculate correlations with NYHA class due to small cell sizes. Correlations of 0.10, 0.30, and 0.50 were deemed small, medium, and large, respectively [24]. For both the reliability (Cronbach's alpha) and construct validity point estimates, we estimated the 95 % CI using the bootstrap method because it does not assume a normal distribution [25]. We also examined the magnitude of relationships across different domains both within and across measures using a multitrait, multimethod evaluation of convergent and discriminant validity [26] with four traits (i.e., domains) and two methods (i.e., PROMIS vs. other PROs). We indicated which values were not significantly different from 0 at $P \leq 0.05$. For the responsiveness analysis, we computed the effect size by dividing the mean change in score by the SD of individuals' baseline scores. We evaluated the magnitude of the effect sizes using standard criteria (i.e., 0.20 is a small effect size, 0.50 is a medium effect size, and 0.80 is a large effect size) [24]. We also estimated the 95 % CI of the effect size for each measure using the bootstrap method. Sample size estimates were based on 2-tailed $\alpha \leq 0.05$, statistical power greater than

80 %, and a correlation of 0.70 between pre- and post-transplant scores. We sought to detect effect sizes as small as 0.30 between pre- to post-transplant scores, which required 60 participants. We used SAS version 9.2 (SAS Institute, Cary, NC) and a 2-tailed significance level of $\alpha \leq 0.05$ for all assessments.

Results

Figure 1 shows the recruitment flow chart. Our analyses focused on patients who had a baseline assessment before heart transplant and a follow-up assessment after transplant. When the study timeline indicated that data collection should be stopped, some enrolled patients had not undergone transplant. We collected a second assessment from these patients; however, because they did not meet the study criteria for pre- and post-transplant assessment, we did not include them in the analyses. Table 1 shows the baseline characteristics of the 48 patients who make up the analytic sample. The median time from baseline assessment to transplant was 32 days (interquartile range 13–99), and the median time from transplant to follow-up assessment was 100 days (interquartile range 71–133).

Table 2 shows the average baseline and follow-up values for NYHA class and the PRO measures. Ninety percent of patients had NYHA class three or four symptoms before heart transplant, and more than half were unable to walk 200 m in 6 min. Compared with the general US population, patients in our sample at baseline had significantly worse physical functioning ($P < 0.001$), greater fatigue ($P < 0.001$), less satisfaction with discretionary social activities ($P < 0.001$), and average depressive symptoms ($P = 0.45$). There were substantial improvements in health status after transplant. Table 2 also shows that the PRO measures were responsive to change. There were large effect sizes (0.80 or higher) representing improvement in physical function, fatigue, and social function after transplant. There was a medium effect size (0.50 or greater) representing improvement in depression. Scores from the PROMIS CATs and the PROMIS short forms were similar.

To assess construct validity, we estimated correlations between the baseline, follow-up, and change-from-baseline values of the PROMIS CATs and short forms (Table 3) and the corresponding, conceptually similar measures. For physical functioning, correlations between the PROMIS measures and the KCCQ were large ($r = 0.68$–0.85). Correlations between 6-min walk test results and the PROMIS CAT were large at baseline and follow-up and medium for change from baseline; they were similarly sized between the 6-min walk and KCCQ physical function measure (0.53 at baseline, 0.63 at follow-up, and 0.35 for change). The PROMIS measures of fatigue had large correlations with the SF-36 vitality scale ($r = -0.75$ to -0.78) and the KCCQ individual fatigue items ($r = -0.57$ to -0.79). Correlations between the PROMIS CAT and the 6-min walk were large at baseline but small at follow-up and for change from baseline. Again the correlations were similarly sized between 6-min walk and the other patient-reported measures of fatigue including KCCQ fatigue 05

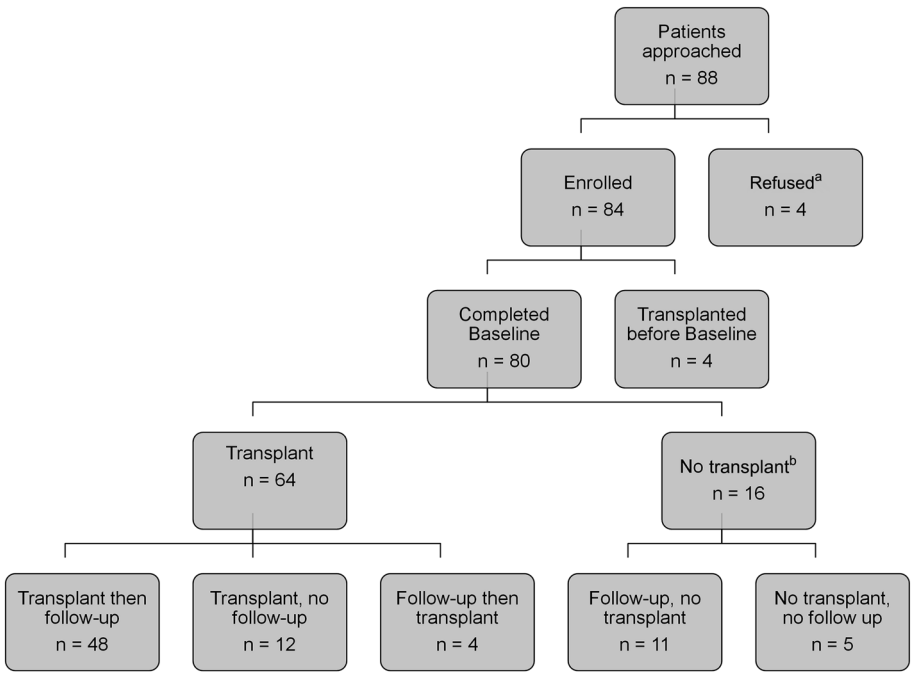

Fig. 1 Recruitment flow chart. [a]Reasons for refusal: patient not interested, too tired, or too sick. [b]Reasons for no transplant: death, unavailability of donor hearts, or the transplant team removed the patient from the list (United Network for Organ Sharing status 7)

Table 1 Patient characteristics at baseline

Characteristic	Patients (N = 48)[a]
Sex, no. (%), male	36 (75.0)
Age, mean (SD), y	51.8 (12.3)
Race, no. (%)	
Asian	1 (2.1)
Black or African American	5 (10.4)
White	42 (87.5)
Hispanic or Latino ethnicity, no. (%)	0 (0)
Educational attainment, no. (%)	
Less than high school	4 (8.3)
High school or equivalent	16 (33.3)
Some college, technical degree, or associate degree	18 (37.5)
College or postgraduate degree	9 (18.7)
Missing	1 (2.1)
Occupational status, no. (%)	
On disability	31 (64.6)
Retired	11 (22.9)
Unemployed, on leave of absence, or homemaker	15 (31.2)
Employed full time or part time	8 (16.7)
Missing	1 (2.1)
Relationship status, no. (%)	
Never married	7 (14.6)
Married or living with committed partner	28 (58.4)
Separated, divorced, or widowed	12 (25.0)
Missing	1 (2.1)
Household income, no. (%)	
< $20,000	17 (35.4)
$20,000–$49,999	16 (33.3)
>$50,000	10 (20.9)
Missing	5 (10.4)
Left ventricular assist device, no. (%)	9 (18.8)
Hospital inpatient, no. (%)	18 (37.5)
Receiving intravenous inotropes, no. (%)	21 (43.8)
Indication for heart transplant, no. (%)	
Coronary artery disease	4 (8.3)
Myopathy	39 (81.3)
Other	5 (10.4)
Recruitment site, no. (%)	
Duke University	9 (18.8)
University of Pittsburgh	39 (81.3)

[a] Percentages may not sum to 100 because of rounding

(0.57 at baseline, 0.21 at follow-up, and 0.23 for change), KCCQ fatigue 06 (0.62 at baseline, 0.28 at follow-up, and 0.14 for change), and SF-36 vitality (0.55 at baseline, 0.09 at follow-up, and 0.36 for change). Correlations between PROMIS depression and the PHQ-2 were large at baseline ($r = 0.65$ and 0.70) and for changes from baseline ($r = 0.53$ and 0.57) and were medium at follow-up ($r = 0.35$ and 0.42). The PHQ-2 items at follow-up had limited variability, with observed responses falling into only two categories. Finally, the PROMIS and KCCQ social function measures had large correlations ($r = 0.60$–0.74).

For both the PROMIS CATs and the PROMIS short forms at both baseline and follow-up, the average correlations between different measures of the same domain (i.e., monotrait-heteromethod) were larger than the average correlations between different domains within a measure (i.e., heterotrait-monomethod) and different domains

Table 2 Health status before and after heart transplant

Measure	Before transplant	After transplant	Effect Size[a] (95 % CI)
Clinical measures			
NYHA class 1, no. (%)	0	31 (64.6)	–
NYHA class 2, no. (%)	1 (2.1)	10 (20.8)	–
NYHA class 3, no. (%)	24 (50.0)	3 (6.3)	–
NYHA class 4, no. (%)	19 (39.6)	1 (2.1)	–
NYHA class missing, no. (%)	4 (8.3)	3 (6.3)	–
6-min walk test, mean (SD), m	186.6 (171.0)	291.8 (108.4)	0.73 (0.32–1.25)
Physical function, mean (SD)			
PROMIS physical function CAT	34.9 (6.1)	42.1 (6.5)	1.14 (0.63–1.95)
PROMIS physical function short form-10a	37.2 (5.2)	43.3 (6.0)	1.13 (0.64–1.88)
KCCQ physical limitation	52.3 (23.2)	80.8 (16.2)	1.21 (0.85–1.70)
Fatigue, mean (SD)			
PROMIS fatigue CAT	58.4 (11.2)	47.0 (9.3)	1.00 (0.56–1.66)
PROMIS fatigue short form-7a	57.6 (10.2)	47.8 (7.0)	0.96 (0.55–1.54)
KCCQ fatigue symptom (Item 5)	3.3 (2.1)	5.8 (1.4)	1.19 (0.80–1.77)
KCCQ fatigue symptom (Item 6)	2.8 (1.5)	4.5 (1.1)	1.09 (0.67–1.70)
SF-36 vitality	39.1 (13.1)	52.8 (8.7)	1.04 (0.66–1.56)
Depression, mean (SD)			
PROMIS depression CAT	51.2 (10.6)	44.8 (7.5)	0.63 (0.29–1.07)
PROMIS depression short form-8b	51.4 (8.9)	44.3 (7.7)	0.79 (0.41–1.29)
PHQ-2	1.7 (1.8)	0.4 (0.7)	0.71 (0.48–0.97)
Social function, mean (SD)			
PROMIS DSA CAT	43.9 (11.2)	53.0 (8.3)	0.80 (0.41–1.36)
PROMIS DSA short form-7a	42.1 (10.3)	50.5 (8.2)	0.81 (0.40–1.39)
KCCQ social limitation	43.2 (29.2)	76.8 (21.3)	1.24 (0.85–1.79)

CAT computerized adaptive test, *DSA* satisfaction with discretionary social activities, *KCCQ* Kansas City Cardiomyopathy Questionnaire, *NYHA* New York Heart Association, *PHQ* Patient Health Questionnaire, *SF-36* Medical Outcomes Study Short Form-36

[a] The effect size is the change in mean score divided by the SD at baseline

across different measures (i.e., heterotrait-heteromethod; Table 4). The full multitrait-multimethod matrices are available in the Appendix.

Table 5 shows the reliability of the PROMIS short forms at baseline and follow-up. All short forms demonstrated acceptable reliability.

The number of items administered in the PROMIS CATs ranged from 2 to 12. Across all assessments and domains, the median was four items, except for the follow-up depression assessment, where the median was six items. Correlations between the PROMIS CAT and short-form scores were large, ranging from 0.88 for physical functioning to 0.96 for satisfaction with discretionary social activities.

Discussion

This study provides evidence for the reliability and construct validity (including responsiveness to change) of 4 PROMIS domains in patients with heart failure before and after heart transplant. We observed large improvements across all of the measures, as expected in this clinical scenario (Table 2). The efficacy of transplant allowed us to examine validity for a wide range of disease morbidity among patients with heart failure; before transplant, 90 % of patients had NYHA 3 or 4, and after transplant, 85 % had NYHA 1 or 2. Furthermore, the magnitude of changes assessed by the PROMIS measures was strongly associated with the magnitude of changes assessed by conceptually similar measures (Table 3). The PROMIS short forms were reliable in these samples. Our comparisons of PROMIS short forms and CATs found that both provided highly sensitive estimates.

There is widespread interest in increasing the role of PROs to improve healthcare quality, yet concurrent concern due to the proliferation of disease-specific PRO measures, which limits researchers' ability to compare disease burden and treatment effectiveness in multiple contexts. There is value in standardizing PRO measurement across different settings. The NIH PROMIS Network

Table 3 Construct validity among health status measures

Measure	Correlation, r[a] (95 % confidence interval)		
	Baseline	Follow-up	Change
PROMIS physical function CAT			
KCCQ physical limitation	0.79 (0.70, 0.85)	0.77 (0.61, 0.88)	0.68 (0.48, 0.82)
6-min walk distance	0.67 (0.46, 0.80)	0.55 (0.15, 0.80)	0.47 (0.19, 0.66)
PROMIS physical function short form-10a			
KCCQ physical limitation	0.85 (0.75, 0.90)	0.75 (0.59, 0.88)	0.74 (0.55, 0.87)
PROMIS fatigue CAT			
KCCQ fatigue 05	0.74 (0.53, 0.84)	0.59 (0.33, 0.73)	0.70 (0.48, 0.81)
KCCQ fatigue 06	0.83 (0.57, 0.90)	0.59 (0.38, 0.75)	0.69 (0.49, 0.81)
SF-36 vitality	0.74 (0.53, 0.87)	0.78 (0.62, 0.87)	0.77 (0.61, 0.86)
6-min walk distance	0.59 (0.29, 0.78)	0.22 (0.14, 0.61)	0.45 (0.14, 0.68)
PROMIS fatigue short form-7a			
KCCQ fatigue 05	0.83 (0.69, 0.90)	0.55 (0.29, 0.72)	0.72 (0.52, 0.82)
KCCQ fatigue 06	0.84 (0.70, 0.91)	0.55 (0.30, 0.70)	0.63 (0.43, 0.76)
SF-36 vitality	0.77 (0.62, 0.87)	0.76 (0.61, 0.86)	0.76 (0.61, 0.85)
PROMIS depression CAT			
PHQ-2 depression	0.71 (0.50, 0.84)	0.21 (0.10, 0.56)	0.53 (0.24, 0.77)
PROMIS depression short form-8a	0.63 (0.43, 0.75)	0.44 (0.16, 0.66)	0.53 (0.31, 0.71)
PHQ-2 depression	0.65 (0.45, 0.76)	0.42 (0.13, 0.64)	0.57 (0.34, 0.74)
PROMIS DSA CAT			
KCCQ social limitation	0.70 (0.51, 0.82)	0.69 (0.48, 0.82)	0.61 (0.44, 0.75)
PROMIS DSA short form-7a			
KCCQ social limitation	0.74 (0.61, 0.83)	0.63 (0.40, 0.78)	0.60 (0.43, 0.72)

CAT computerized adaptive test, *DSA* satisfaction with discretionary social activities, *KCCQ* Kansas City Cardiomyopathy Questionnaire, *PHQ* Patient Health Questionnaire, *SF-36* Medical Outcomes Study Short Form-36

[a] From Pearson product moment correlation

Table 4 Summary of multitrait-multimethod matrices

Measures	Mean Correlation, r[a]		
	Monotrait-Heteromethod	Heterotrait-Monomethod	Heterotrait-Heteromethod
PROMIS CAT and other PROs[b]			
Baseline	0.76	0.66	0.62
Follow-up	0.63	0.50	0.44
PROMIS short form and other PROs[b]			
Baseline	0.79	0.70	0.63
Follow-up	0.61	0.48	0.42

CAT computerized adaptive test, *PRO* patient-reported outcome

[a] From Pearson product moment correlation

[b] Other PROs include the Kansas City Cardiomyopathy Questionnaire, the Medical Outcomes Study Short Form-36, and the Patient Health Questionnaire-2

has developed PRO measures intended to measure important domains of health across chronic diseases without substantial loss in sensitivity for any one disease. This study demonstrates that the PROMIS measures provided sensitivity (responsiveness) to change in a sample of patients with cardiac transplantation. This study also provides further support for the reliability and validity of the PROMIS measures. PROMIS measures are a reasonable choice in this context and will facilitate comparisons across studies and health conditions.

As with all item bank-based measures, the PROMIS domains can be assessed using either CATs or short forms.

Table 5 Reliability of PROMIS short forms

Measure	Cronbach's alpha (95 % confidence interval)		IRT estimated marginal reliability	
	Baseline	Follow-up	Baseline	Follow-up
PROMIS physical function 10a	0.84 (0.76, 0.89)	0.85 (0.76, 0.93)	0.92	0.91
PROMIS fatigue 7a	0.91 (0.85, 0.94)	0.77 (0.65, 0.84)	0.91	0.85
PROMIS depression 8b	0.91 (0.86, 0.95)	0.92 (0.83, 0.97)	0.90	0.73
PROMIS DSA 7a	0.95 (0.92, 0.97)	0.92 (0.86, 0.94)	0.93	0.92

DSA satisfaction with discretionary social activities

Generally, the results in this study were the same for CATs and short forms; however, the CAT scores demonstrated the same level of responsiveness and validity as the short forms with fewer items. This advantage in measure length is balanced by the need to use computers for administration and scoring.

Our study has limitations. First, there may be differences in PRO responses by mode of administration. Although a recent study found no statistically significant differences in PROMIS scores by multiple methods of administration, including interactive voice recording and personal computer [27], another study found differences by mode of administration (interview versus self-administration) [28]. Second, because donor heart availability necessitates a quick turnaround for heart transplant, it was not feasible to collect baseline assessments at the same time for all patients before transplant. Third, difficulties obtaining follow-up data for 12 patients meant that we did not reach our target sample size of 60 patients. However, because the effect sizes were substantially larger (0.63–1.24) than what we assumed in the power calculation (0.30), we had sufficient statistical power. Fourth, to limit patient burden, we did not administer the full KCCQ or SF-36. Fifth, although we were able to note the consistency of results among PROMIS and other PRO measures, the study was not designed or powered to evaluate statistical differences between the PROMIS and other PRO measures. Finally, an element that served as both a limitation and a strength was the limited variability in change-from-baseline scores; that is, nearly everyone made large improvements. This result was advantageous for the responsiveness analyses and allowed us to analyze the PROMIS measures in patients when they were experiencing severe functional limitations (pre-transplant) as well as when they were not (post-transplant). The large and consistent improvements in scores also support the generalizability of the findings because they show that the sample was typical of heart recipients, who are almost universally found to show major improvements in health and well-being from before to after transplant [20–22]. However, because correlations can be attenuated when there is limited variability on a given measure (here, change from baseline), we were able to observe only relatively low correlations between such change and other measures (e.g., depression at follow-up) in our examination of construct validity. The limited variability in change-from-baseline scores also prevented us from conducting the minimally important difference analyses that we planned. A related limitation is that since the changes observed were very large, additional work will be needed to understand the sensitivity of PROMIS measures to more subtle clinical changes.

Conclusions

This study provides evidence for the reliability and construct validity (including responsiveness to change) of four PROMIS domains in patients with heart failure before and after heart transplant. Researchers should feel comfortable choosing either PROMIS short-form or CAT measures in this context, and by doing so they will facilitate comparisons across studies and health conditions. At the same time, there are important disease-specific considerations when measuring health status in patients with heart failure, such as disease-specific symptoms (e.g., dyspnea) and concepts (e.g., heart failure-specific quality of life), which are not measureable within PROMIS. It is likely that including disease-specific measures along with generic measures will provide the most complete assessment of patient-reported health.

Acknowledgments This study was supported by grants U01AR052186 and U01AR052155 from the National Institute of Arthritis and Musculoskeletal and Skin Diseases. Dr. Flynn was supported in part by the Research and Education Program Fund, a component of the Advancing a Healthier Wisconsin endowment at the Medical College of Wisconsin. Dr. Hays was supported by grants P30AG028748 and P30AG021684 from the National Institute on Aging and grant P20MD000182 from the National Center on Minority Health and Health Disparities. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest None.

References

1. Barnes, S., Gott, M., Payne, S., Parker, C., Seamark, D., Gariball, S., & Small, N. (2006). Prevalence of symptoms in a community-based sample of heart failure patients. *Journal of Pain and Symptom Management, 32*(3), 208–216.
2. Dracup, K., Walden, J. A., Stevenson, L. W., & Brecht, M. L. (1992). Quality of life in patients with advanced heart failure. *Journal of Heart and Lung Transplantation, 11*(2 Pt 1), 273–279.
3. Rumsfeld, J. S., Alexander, K. P., Goff, D. C, Jr, Graham, M. M., Ho, P. M., Masoudi, F. A., et al. (2013). Cardiovascular health: The importance of measuring patient-reported health status: a scientific statement from the American Heart Association. *Circulation, 127*(22), 2233–2249.
4. Garin, O., Herdman, M., Vilagut, G., Ferrer, M., Ribera, A., Rajmil, L., et al. (2014). Assessing health-related quality of life in patients with heart failure: A systematic, standardized comparison of available measures. *Heart Failure Reviews, 19*(3), 359–367.
5. Rector, T. S., & Cohn, J. N. (1992). Assessment of patient outcome with the Minnesota Living with Heart Failure questionnaire: Reliability and validity during a randomized, double-blind, placebo-controlled trial of pimobendan. Pimobendan Multicenter Research Group. *American Heart Journal, 124*(4), 1017–1025.
6. Green, C. P., Porter, C. B., Bresnahan, D. R., & Spertus, J. A. (2000). Development and evaluation of the Kansas City Cardiomyopathy Questionnaire: A new health status measure for heart failure. *Journal of the American College of Cardiology, 35*(5), 1245–1255.
7. Kappelman, M. D., Long, M. D., Martin, C., DeWalt, D. A., Kinneer, P. M., Chen, W., et al. (2014). Evaluation of the Patient-Reported Outcomes Measurement Information System in a large cohort of patients with inflammatory bowel diseases. *Clinical Gastroenterology and Hepatology, 12*(8), 1315–1323. **e2**.
8. Fries, J. F., Cella, D., Rose, M., Krishnan, E., & Bruce, B. (2009). Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. *Journal of Rheumatology, 36*(9), 2061–2066.
9. Hays, R. D., Spritzer, K. L., Fries, J. F., & Krishnan, E. (2013). 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. *Annals of Rheumatic Diseases,*. doi:10.1136/annrheumdis-2013-204053.
10. Broderick, J. E., Schneider, S., Junghaenel, D. U., Schwartz, J. E., & Stone, A. A. (2013). Validity and reliability of Patient-Reported Outcomes Measurement Information System instruments in osteoarthritis. *Arthritis Care & Research, 65*(10), 1625–1633.
11. Yost, K. J., Eton, D. T., Garcia, S. F., & Cella, D. (2011). Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. *Journal of Clinical Epidemiology, 64*(5), 507–516.
12. Hahn, E. A., Devellis, R. F., Bode, R. K., Garcia, S. F., Castel, L. D., Eisen, S. V., et al. (2010). Measuring social health in the Patient-Reported Outcomes Measurement Information System (PROMIS): Item bank development and testing. *Quality of Life Research, 19*(7), 1035–1044.
13. Pilkonis, P. A., Choi, S. W., Reise, S. P., Stover, A. M., Riley, W. T., Cella, D., & PROMIS Cooperative Group. (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.
14. 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.
15. McHorney, C. A., Ware, J. E, Jr, Lu, J. F., & Sherbourne, C. D. (1994). The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Medical Care, 32*(1), 40–66.
16. Ware, J. E, Jr, & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care, 30*(6), 473–483.
17. Kroenke, K., Spitzer, R. L., & Williams, J. B. (2003). The Patient Health Questionnaire-2: Validity of a two-item depression screener. *Medical Care, 41*(11), 1284–1292.
18. Balke, B. (1963). *A simple field test for the assessment of physical fitness*. Oklahoma City, OK: Civil Aeromedical Research Institute, Aeromedical Research Division, 1963 Apr. Report No.: 63-6.
19. ATS Committee on Proficiency Standards for Clinical Pulmonary Function Laboratories. (2002). ATS statement: Guidelines for the six-minute walk test. *American Journal of Respiratory and Critical Care Medicine, 166*(1), 111–117.
20. Grady, K. L., & Lanuza, D. M. (2005). Physical functional outcomes after cardiothoracic transplantation. *Journal of Cardiovascular Nursing, 20*(5 Suppl), S43–S50.
21. Molzahn, A. E., Burton, J. R., McCormick, P., Modry, D. L., Soetaert, P., & Taylor, P. (1997). Quality of life of candidates for and recipients of heart transplants. *Canadian Journal of Cardiology, 13*(2), 141–146.
22. Paris, W., & White-Williams, C. (2005). Social adaptation after cardiothoracic transplantation: a review of the literature. *Journal of Cardiovascular Nursing, 20*(5 Suppl), S67–S73.
23. Hays, R. D., & Reeve, B. B. (2010). Measurement and modeling of health-related quality of life. In J. Killewo, H. K. Heggenhougen, & S. R. Quah (Eds.), *Epidemiology and demography in public health* (pp. 195–205). San Diego: Academic Press.
24. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
25. Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
26. Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105.
27. Bjorner, J. B., Rose, M., Gandek, B., Stone, A. A., Junghaenel, D. U., & Ware, J. E, Jr. (2014). Difference in method of administration did not significantly impact item response: An IRT-based analysis from the Patient-Reported Outcomes Measurement Information System (PROMIS) initiative. *Quality of Life Research, 23*(1), 217–227.
28. Hahn, E. A., Rao, D., Cella, D., & Choi, S. W. (2008). Comparability of interview- and self-administration of the Functional Assessment of Cancer Therapy-General (FACT-G) in English- and Spanish-speaking ambulatory cancer patients. *Medical Care, 46*(4), 423–431.