Published in final edited form as:

J Neurosurg Spine. 2018 March ; 28(3): 268–279. doi:10.3171/2017.7.SPINE17661.

Validation of Patient-Reported Outcomes Measurement Information System (PROMIS) computerized adaptive tests in cervical spine surgery

Barrett S. Boody, MD¹, Surabhi Bhatt, BS¹, Aditya S. Mazmudar, BA¹, Wellington K. Hsu, MD¹, Nan E. Rothrock, PhD², Alpesh A. Patel, MD¹

¹Department of Orthopaedic Surgery, Feinberg School of Medicine, Chicago, Illinois

²Department of Medical Social Sciences, Feinberg School of Medicine, Chicago, Illinois

Abstract

OBJECTIVE—The Patient-Reported Outcomes Measurement Information System (PROMIS), which is funded by the National Institutes of Health, is a set of adaptive, responsive assessment tools that measures patient-reported health status. PROMIS measures have not been validated for surgical patients with cervical spine disorders. The objective of this project is to evaluate the validity (e.g., convergent validity, known-groups validity, responsiveness to change) of PROMIS computer adaptive tests (CATs) for pain behavior, pain interference, and physical function in patients undergoing cervical spine surgery.

METHODS—The legacy outcome measures Neck Disability Index (NDI) and SF-12 were used as comparisons with PROMIS measures. PROMIS CATs, NDI-10, and SF-12 measures were administered prospectively to 59 consecutive tertiary hospital patients who were treated surgically for degenerative cervical spine disorders. A subscore of NDI-5 was calculated from NDI-10 by eliminating the lifting, headaches, pain intensity, reading, and driving sections and multiplying the final score by 4. Assessments were administered preoperatively (baseline) and postoperatively at 6 weeks and 3 months. Patients presenting for revision surgery, tumor, infection, or trauma were excluded. Participants completed the measures in Assessment Center, an online data collection tool accessed by using a secure login and password on a tablet computer. Subgroup analysis was also performed based on a primary diagnosis of either cervical radiculopathy or cervical myelopathy.

RESULTS—Convergent validity for PROMIS CATs was supported with multiple statistically significant correlations with the existing legacy measures, NDI and SF-12, at baseline.

Correspondence: Barrett S. Boody: Feinberg School of Medicine, Chicago, IL. barrett.s.boody@gmail.com.

Author Contributions

Conception and design: Boody, Hsu, Patel. Acquisition of data: Bhatt. Analysis and interpretation of data: all authors. Drafting the article: all authors. Critically revising the article: Boody, Bhatt, Hsu, Rothrock, Patel. Reviewed submitted version of manuscript: Boody, Bhatt, Rothrock, Patel. Approved the final version of the manuscript on behalf of all authors: Boody. Statistical analysis: Boody, Mazmudar, Rothrock. Administrative/technical/material support: Bhatt, Rothrock. Study supervision: Hsu, Patel.

Disclosures

Dr. Hsu: consultant for Stryker, Medtronic, Mirus, Allosource, NuVasive, AONA, Xtant, and Bioventus.

Dr. Rothrock: served as a co-investigator on PROMIS grants from the NIH and provided guidance on PROMIS administration, score interpretation, and general educational information about measurement development and validation but was not involved in data collection, data management, analytic design, or the statistical analyses.

Furthermore, PROMIS CATs demonstrated known-group validity and identified clinically significant improvements in all measures after surgical intervention. In the cervical radiculopathy and myelopathic cohorts, the PROMIS measures demonstrated similar responsiveness to the SF-12 and NDI scores in the patients who self-identified as having postoperative clinical improvement. PROMIS CATs required a mean total of 3.2 minutes for PROMIS pain behavior (mean $\pm$ SD 0.9 ± 0.5 minutes), pain interference (1.2 ± 1.9 minutes), and physical function (1.1 ± 1.4 minutes) and compared favorably with 3.4 minutes for NDI and 4.1 minutes for SF-12.

CONCLUSIONS—This study verifies that PROMIS CATs demonstrate convergent and known-groups validity and comparable responsiveness to change as existing legacy measures. The PROMIS measures required less time for completion than legacy measures. The validity and efficiency of the PROMIS measures in surgical patients with cervical spine disorders suggest an improvement over legacy measures and an opportunity for incorporation into clinical practice.

Keywords

PROMIS; patient-reported outcomes; cervical spine

TO be able to more comprehensively evaluate the impact of surgical treatment on patients, there is a need to have reliable, valid, and efficient outcome measures.^{9,14–16} Because the US health care system places an increasing focus on the value of delivered care, clinicians require improved patient outcome metrics to both demonstrate the value and justify the costs for our clinical interventions.²⁰

Historically, the success of surgical management for patients undergoing spine surgery has relied exclusively on the clinician's physical examination, radiological interpretation, and overall perception of the patient's well-being. These have been demonstrated to be insufficient proxies for patient-centered outcomes.²⁴ Multiple patient-reported outcome (PRO) measures have been used for research in spine pathology and range from general measures of disability and function such as the Oswestry Disability Index (ODI) and visual analog scale pain scores to more specific outcome measures focused on the impact of spine pathology, including the Neck Disability Index (NDI), the Swiss Spinal Stenosis Questionnaire, the Oxford Spinal Stenosis Questionnaire, and the Maine-Seattle Back Questionnaires.²⁴ Historically, PROs are time consuming, demonstrate disease bias, and may display inaccuracies when testing patients with either severe functional disability or extreme functional ability (i.e., floor and ceiling effects, respectively).¹⁹ Floor and ceiling effects describe the failure of PROs that are below or above, respectively, certain scores to accurately measure data variance. Floor effects are seen when the patient's baseline functioning falls below of the measurable realm for a PRO, as the given score will be the "floor" for that measurement metric. In these scenarios, any change identified on posttesting assessment will not account for the discrepancy between the patient's true baseline function and the floor measurement score. Similarly, "ceiling" effects can be seen when posttest functioning exceeds the measurable realm for PROs, thereby obscuring the difference between the true improvement in function and the reported PRO ceiling value.

Spine-specific outcome measures such as NDI are commonly used to quantify the effect of pathology and/or treatment on spine functioning by asking questions focused on how the

neck and/or back limit common daily activities or how much pain is related to neck and/or back pathology. Conversely, generalized outcome measures, such as SF-12, ask questions without specifying the etiology of disability. This allows general outcome measures to understand the impact of treatment for spine pathology on the patient's overall quality of life and global level of functioning. Global measures capture concomitant pathologies that contribute to disability as well as the downstream effects of spine treatments on other organ systems (e.g., musculoskeletal, cardiovascular, pulmonary), thereby allowing assessment of the fractional contribution of spine pathology and/or treatment on overall functioning and pain behaviors.

The goal of the Patient-Reported Outcomes Measurement Information System (PROMIS) is to develop a psychometrically validated system of PRO measures in respondents with a wide range of chronic diseases and demographic characteristics.¹⁹ The PROMIS PROs assess subjective experiences including symptom frequency and severity, emotional and social well-being, and perceived level of health and functional ability.⁷ To facilitate the timely completion of PROs for improved use in the clinical setting, the PROMIS group developed computer adaptive tests (CATs).

CATs draw from a collection of questions that measure a similar construct (e.g., physical function [PF]). When administered, CATs use responses from previous questions to select relevant follow-up questions in order to sharpen the estimate of a respondent's score on the domain being measured and minimize the overall question burden for the respondent. This process is repeated until a person is asked the maximum number of items allowed for the CAT (generally 12 items for PROMIS domains) or until the score is adequately precise (standard error less than 3.0 on the T-score metric). An example of a PROMIS item for PF is "Are you able to push open a heavy door?" Examples of PROMIS CATs can be found at <https://www.assessmentcenter.net/ac1/assessments/catdemo>. The value of CATs is the high level of measurement precision with fewer required questions, thereby reducing the time required to complete the survey and potentially increasing their suitability for clinical settings.^{8,11,13,14,24,28,30} A recent study by Brodke et al. analyzed the floor and ceiling effects for PROs administered to patients with spine pathology and reported improved ceiling and floor effects for PROMIS PF CAT (0.81% and 3.86%, respectively) compared with ODI (6.91% and 44.24%, respectively) and SF-36 PF (5.97% and 23.65%, respectively).⁵

The utility and validity of PROMIS CATs have been demonstrated in a variety of medical and surgical fields, and the CATs display reliability, validity, flexibility, and inclusiveness under conditions such as depression, cancer, chronic obstructive pulmonary disease, heart failure, and other pathologies.^{1,12,17,22,23} The goal of our study is to determine the validity of PROMIS through comparison with widely accepted legacy measures in patients undergoing cervical spine surgery.

Methods

Design

Project Structure—After obtaining institutional review board approval, we invited all patients undergoing cervical spine surgery for the treatment of cervical radiculopathy or myelopathy secondary to primary degenerative disease of the cervical spine to participate. Inclusion criteria were age between 18 and 95 years and the ability to read and speak English. Patients who opted to participate in the study were consecutively enrolled. Any patients who presented for revision surgery or with osseous tumors, trauma, or infection were excluded from the study. Each patient was directed to complete the PROMIS and legacy measures with our wireless Internet-enabled iPad. Assessment Center, a free, online data collection tool, was used for data collection (www.assessmentcenter.net). An a priori sample size of 57 participants was found to demonstrate 80% and > 99% power for correlations of 0.432 and 0.700, respectively, using a Type I error rate < 5%. All power calculations were performed using PASS 2008 (NCSS).

Each assessment was administered prior to surgery (baseline; visit 1), as well as 6 weeks (visit 2) and 3 months (visit 3) postoperatively. Participants completed their baseline assessment at the clinic, while all postoperative assessments were completed over the phone or via an email link to Assessment Center. The order of PROs administered to patients was randomized to minimize effects from response fatigue. Patients unable to use the iPad were given the option to have the study coordinator assist them by reading the questions out loud and subsequently entering the patient's responses. Patients were not coached about how to respond to the questionnaires nor were interpretations of the items or other feedback given to the participant. The treating surgeon was not involved with or present for the collection of the patient outcomes.

Measures

The assessment included legacy measures with demonstrated validity within this patient population (Neck Disability Index [NDI] and SF-12¹⁶) and PROMIS CATs for PF, pain interference (PI), and pain behavior (PB), and impactful comorbid conditions.

NDI—NDI reports self-rated disability due to neck pain, as well as concentration and ability to perform usual activities (e.g., working, driving), using a scale from 0 to 50 with increasing scores indicating worsening disability. NDI-5 can be calculated from the 10-item NDI by eliminating the lifting, headaches, pain intensity, reading, and driving sections and multiplying the final score by 4. Previous concerns with NDI-10 include insufficient unidimensionality and a large floor effect.¹⁸ The abridged NDI-5 shows comparable psychometric properties as the NDI-10 and may be a more practical clinical tool due to the reduced number of questions and focus on functional evaluation.²⁹

SF-12—SF-12 assesses physical, social, and mental function over 12 domains. It is summarized into a composite physical component and mental component of a patient's health status. The SF-12 scale uses a population mean of 50 with a standard deviation of 10. Higher scores indicate better health.

PROMIS PF CAT—The PROMIS PF CAT measures self-reported capability rather than the actual performance of physical activities. This includes the functioning of one's upper extremities (dexterity), lower extremities (walking or mobility), and central regions (neck and back), as well as instrumental activities of daily living such as running errands. PROMIS PF CAT is appropriate for the general adult population and adults with chronic health conditions. PROMIS PF CAT assesses current function rather than function over a specified time period.

PROMIS PF CAT v1.2 is administered using a bank of 121 potential items. All PROMIS scores are reported on a T-score metric, with a score of 50 points aligning with the general population mean and a standard deviation of 10. Higher scores indicate better physical functioning.

PROMIS PI CAT—PROMIS PI CAT measures the self-reported consequences of pain on relevant aspects of one's life. This includes the extent to which pain hinders engagement with social, cognitive, emotional, physical, and recreational activities. PROMIS PI CAT assesses PI over the past 7 days. PROMIS PI CAT v1.0 consists of a bank of 41 potential items. Higher scores indicate more difficulty performing activities because of pain.

PROMIS PB CAT—PROMIS PB CAT measures self-reported external manifestations of pain such as behaviors that typically indicate to others that an individual is experiencing pain. These actions or reactions can be verbal, nonverbal, involuntary, or deliberate. They include observable displays (sighing and crying), pain severity behaviors (resting, guarding, facial expressions, and asking for help), and verbal reports of pain. PROMIS PB CAT assesses pain behavior over the past 7 days. PROMIS PB CAT v1.0 consists of a bank of 39 potential items. Higher scores indicate more external manifestations of experiencing pain (e.g., moving stiffly).

Impactful Comorbid Conditions—The impactful comorbid conditions assessment addresses the impact of other health conditions on physical function and pain. It includes the question "Are your answers to today's questions being affected by any conditions (e.g., arthritis, knee pain, heart disease, lung disease) other than what you are being seen for today?" and is answered as "yes" or "no."

Global Rating of Change—The global rating of change question assesses one's perception of change between assessments ("How is your neck or back condition since your last visit with us?"). Responses were "much better," "slightly better," "about the same," "slightly worse," and "much worse." This question was used to evaluate responsiveness.

Statistical Analysis

For all time points, PROMIS CAT T-scores were exported directly from Assessment Center. NDI-5 scores were calculated by summing all responses from NDI-10 and multiplying by 4. SF-12 Physical and Mental Component Summaries (PCS and MCS, respectively) were calculated using QualityMetric Health Outcomes Scoring Software 4.5.

All analysis was conducted using SAS (SAS Institute) or SPSS (IBM) statistical software. Descriptive statistics were generated to establish the demographic profile of the patient sample and were calculated for all PRO measures at baseline.

Convergent validity explores the strength of the relationship between theoretically related constructs. Convergent validity was evaluated using Pearson correlation coefficients among PROMIS CATs, NDI, and SF-12 at baseline. Changes in scores were calculated between each assessment point for all measures, and statistical significance was evaluated using single Student t-tests. Pearson correlation coefficients were also calculated using the change scores to evaluate responsiveness over time. Correlation values of 0.0–0.19, 0.20–0.39, 0.40–0.59, 0.60–0.79, and 0.80–1.0 are described as very weak, weak, moderate, strong, and very strong, respectively.

Known-groups validity asks if a measure is able to distinguish between groups of patients that are known to be different. To test discriminant (known groups) validity, the PROMIS CATs, NDI-5, and SF-12 questionnaire scores of the patients' baseline status were compared across groups with varying levels of limitations in the kind of work or activity (Item 3a on SF-12) using single Student t-tests.

The measure of construct responsiveness describes the ability of an outcome metric to appropriately identify anticipated change. For this study, the subgroup of patients who self-reported feeling "much better" at visit 3 compared with visit 1 were used to measure responsiveness in outcome scores. To determine the effect size, the standardized response mean (change in the mean across a given time point divided by the standard deviation) was reported. Thresholds used for interpreting the responsiveness of the standardized response means were > 0.8 (large), $0.5–0.8$ (moderate), and < 0.5 (small).

This study was not designed to evaluate whether the degree of change over time was clinically significant but instead was designed to evaluate whether the PROMIS measures were able to capture that change in comparison with legacy PRO measures. However, to provide clinical context for reported responsiveness (effect sizes), PROMIS CATs and legacy measures were followed over time to observe similarities in reaching minimally important clinical difference (MCID) thresholds. While there are a few publications on the MCIDs for the PROMIS PB, PI, and PF measures, no validated MCID for PROMIS measures for spine pathology have been published. An acceptable estimate that is currently used is 50% of the reported standard deviation.^{3,25} Amtmann et al. recently reported that a MCID of 3.5 to 5.5 points on PROMIS PI may be considered meaningful in the low-back pain patient population.¹ The commonly reported MCID ranges for NDI-10 varies from 4.8 to 13.4 but has been reported to be as low as 2.7.^{2,6,10} However, because the MCID for NDI-5 is not well established, NDI-10 data are used as a surrogate. The MCIDs for PCS and MCS of SF-12 that were previously reported for cervical spine pathology were 8.1 and 4.7, respectively.²⁷ Due to variability in deriving and reporting MCID thresholds, clinicians should interpret reaching MCID thresholds in isolation with caution.¹⁰

Finally, we evaluated subgroups of patients with cervical myelopathy and cervical radiculopathy to evaluate whether PROMIS CATs demonstrated known-groups validity and

responsiveness for 2 differing but commonly encountered clinical presentations. The subgroup analyses followed a similar structure to the overall group analysis to elucidate potential differences focused on diagnostic group responsiveness. Patients with cervical myelopathy were expected to have stable PF over time. Patients with cervical radiculopathy were expected to report improved PF and decreased PB and PI after surgery.

Results

Of the 59 patients enrolled in this study, 90% completed the baseline, 6-week, and 3-month assessments. The group of 59 patients (mean $\pm$ SD age 55.7 ± 12.2 years) was 61.0% male and 67.8% Caucasian. At the baseline assessment, the mean PROMIS PB and PI scores were 8.8 and 11.9 points above the general population mean of 50, respectively, thereby indicating greater than average impairments secondary to pain. Furthermore, baseline PF was mean 9.9 points below the population mean of 50, thereby indicating worse baseline PF than the population average. Mean SF-12 T-scores at baseline were below the population mean of 50 (mean PCS and MCS of 35.2 and 45.7, respectively), with subsection scores ranging from 32.7 for role limitations to 47 for general health. The mean NDI score at baseline was 35.2.

Convergent validity for PROMIS CATs was supported with multiple correlations in the expected direction at baseline. PROMIS PF correlated as expected with the NDI-5 ($r = -0.47$) and SF-12 PCS ($r = 0.57$) (each $p < 0.05$). Similarly, PROMIS PI correlated strongly with NDI-5 ($r = 0.61$, $p < 0.05$) but demonstrated a weaker correlation with SF-12 PCS ($r = -0.34$). Unexpectedly, it also correlated with SF-12 MCS ($r = -0.44$, $p < 0.05$). While PROMIS PB demonstrated moderate correlations with NDI-5 ($r = 0.59$) at baseline, weaker correlations were seen with SF-12 MCS ($r = -0.44$, $p < 0.05$) and SF-12 PCS ($r = -0.15$, $p = 0.289$) (Table 1).

The baseline self-reported level of limitation (completely limited, limited some or most of time, or limited little or none of the time) suggested variability in functioning for the group overall. PROMIS PF and PI showed statistically significant differences between limitation groups ($p < 0.01$), while PB showed nonsignificant trends in variability between groups ($p = 0.13$) (Tables 1–3).

PF and pain improved following surgery across all measures as expected. The scores for PROMIS PB and PI decreased 4.4 and 6.6 points, respectively, at 3 months and reached MCID (each $p < 0.01$). PROMIS PF increased 2.5 points over the same time period ($p < 0.01$), but that change did not meet the threshold for MCID. NDI-5 scores decreased clinically and significantly by 11.6 points ($p < 0.01$) at 3 months. SF-12 MCS and PCS increased 3.6 ($p = 0.07$) and 3.8 points ($p = 0.03$), respectively, but failed to demonstrate MCID.

The sample was divided into subgroups based on self-reported changes (“much better” versus all others). Comparing baseline to the 6-week follow-up (visit 1 vs visit 2), the clinically improved group reported improvement in PROMIS PB, PI, and PF (mean -5.7 , -6.4 , and 5.6 , respectively) as well as legacy measures NDI-5, SF-12 MCS, and SF-12 PCS (mean -12.6 , 10.4 , and 4.05 , respectively), as expected. The standardized response means

for the “much better” subgroup from visit 1 versus visit 2 demonstrated comparable effect sizes among outcome measures (PB = -0.46; PI = -0.53; PF = 0.48; NDI-5 = -0.64; SF-12 MCS = 0.99; SF-12 PCS = 0.34), thereby suggesting the similar responsiveness of the PROMIS measure to legacy measures. The scores for the clinically improved group also improved between 6 weeks and 12 weeks (visit 2 vs visit 3) and largely in the expected direction (PB = -2.1; PI = -3.7; PF = -2.2; NDI-5 = -1.54; SF-12 MCS = 0.74; SF-12 PCS = 1.77) (Tables 4 and 5).

Diagnostic subgroup (radiculopathy and myelopathy) analyses were performed to evaluate PROMIS CAT responsiveness for 2 commonly seen pathologies. Comparison of diagnostic subgroups revealed that PROMIS CAT performed as expected for the cervical myelopathy and radiculopathy subgroups and demonstrated responsiveness consistent with legacy PROs (Figs. 1–4).

In the cervical myelopathy subgroup (n = 25), PROMIS PB and PI decreased 3.9 (p = 0.113) and 5.3 points (p = 0.043), respectively, from baseline to the 6-week mark, with outcomes largely plateauing by the 3-month assessment. PROMIS PF T-scores increased an average of 1.8 points (p = 0.184) between visits 1 and 3. Similar to the overall analysis, patients who reported feeling “much better” between visits 1 and 2 displayed more robust improvements in outcome measures, including an increase in the PROMIS PF score of 5.1 points and an decrease in NDI-5 of 8.6 points, suggesting the responsiveness of the PROMIS and legacy outcome measures.

In the cervical radiculopathy subgroup (n = 28), PROMIS PB and PI decreased 5.2 (p = 0.001) and 6.5 points (p = 0.001), respectively, and the PROMIS PF T-scores increased 2.5 points (p = 0.103) by the 3-month follow up. For patients who reported feeling “much better,” PROMIS CAT showed improved responsiveness similar to the legacy outcome measures. Between visits 1 and 2, PROMIS PB and PI decreased 8.8 and 10.5 points, respectively, PROMIS PF increased 10.6 points, NDI-5 decreased 16.8 points, and SF-12 MCS and PCS increased 15.4 and 3.1, respectively. Between visits 2 and 3, both the PROMIS and legacy measures demonstrated minimal interval changes.

PROMIS CATs required a mean total of 3.2 minutes to administer all 3 measures. Mean $\pm$ SD individual CAT completion times were 0.9 ± 0.5 minutes for PB, 1.2 ± 1.9 minutes for PI, and 1.1 ± 1.4 minutes for PF. This compares favorably with the observed completion times for NDI and SF-12 form (3.4 and 4.1 minutes, respectively). The individual completion times for the items used in the modified NDI-5 were not available (Table 6).

In addition to the improved efficiency of outcomes measurement, the cervical spine PROMIS outcome measures demonstrated minimal floor and ceiling effects^{4,21} (Fig. 5–7) despite a substantial number of individuals experiencing severe disability (17 of 59 patients) who identified as crippled or severely disabled.

Discussion

This study demonstrates an improvement in outcome measurement using PROMIS CATs when compared with legacy measures in surgical patients with cervical spine disorders.

PROMIS CATs demonstrated convergent validity, responsiveness, and known-groups validity with correlations and responsiveness similar to legacy measures. PROMIS CATs demonstrated minimal floor effects despite a high degree of disability in this population. Additionally, the 3 PROMIS measures were completed faster than the other measures while asking fewer questions to patients.

The PROMIS PF item bank has been previously evaluated for suitability for spinal pathology. Hung et al. administered the full set of 124 PF questions to 438 patients with spinal pathology. Review of the PF scores displayed minimal floor and ceiling effects, high interitem reliability, and minimal item bias.²⁰ This provides further evidence for the validity of these item banks in patients with spine problems. Our study extends this evidence to demonstrate the validity of PROMIS CATs in patients undergoing surgical intervention for cervical spine disease.

PROMIS CATs reduced the time required to capture outcomes data compared with conventional legacy measures. For our study, PROMIS CATs required a mean total of 3.1 minutes to administer PROMIS PB (mean $\pm$ SD 0.9 ± 0.5 minutes), PI (1.2 ± 1.9 minutes), and PF (1.1 ± 1.4 minutes) and compared favorably with 3.4 minutes for NDI and 4.1 minutes for SF-12. Because the adaptive testing format allows for a lower volume of highly relevant questions, we believe PROMIS CATs completion times will be less disruptive to clinical workflow than other longer PROs. Additionally, real-time scoring with PROMIS CATs enables the clinician to use the score at the time of the clinical encounter for patient education and feedback.

Due to the increasing demands placed on health care practitioners, automated and effective outcome tools are highly desirable. Utilizing a Web-based data collection model for PROMIS instruments allows the tracking of completion times, time and date stamps on responses, immediate scoring, and automated tracking of missing data. These features make PROMIS ideal for clinical settings by facilitating data tracking and collection for research and quality control purposes, as well as immediate point-of-care data analysis and reviewing results with patients.

We selected widely used measures to assess the same constructs as the PROMIS measures to test convergent validity. While the overall comparisons demonstrated variable strength in the correlation between PROMIS and legacy PROs, it is important to note that stronger correlations were seen among measures when grouped by function (PROMIS PF and SF-12 PCS, $r = 0.57$, $p < 0.05$) and pain (PROMIS PI and NDI, $r = 0.61$; PROMIS PB and NDI-5, $r = 0.59$; PROMIS PB and SF-12 MCS, $r = -0.44$; $p < 0.05$ for all). While correlations between function and pain measures can be seen, strong correlations would not be expected. When the NDI items and the PROMIS items are compared, they demonstrate overlapping but not redundant constructs. NDI-5 includes 3 items (personal care, work, and sleep) that have responses about the amount of disturbance or disability a patient experiences in these arenas. This closely maps to the PROMIS definition of PF. However, several NDI items (concentration and recreation) are outside of the realm of what PROMIS PF measures. Consequently, a very strong correlation is not expected because the construct of PF is overlapping but not entirely redundant to disability. Additionally, the failure of the initial

groups to reach MCID may be in part due to variable amounts of clinical improvement seen at early follow-up. In the myelopathy group, the lack of clinical improvement may still be deemed a successful outcome given the unfavorable natural history and expected clinical deterioration of untreated disease.

While our study supports the use of PROMIS CATs in surgical patients with cervical spine disease, several limitations are noted. The study population, while meeting the a priori determined size, is small and limits subgroup analysis. While the 3-month follow-up is sufficient to assess the validity and responsiveness of the outcome measures, it is not enough to assess the effectiveness of the surgical treatment rendered. Parker et al. suggested 12-month follow-up, as they found the MCID of 3-month ODI for lumbar surgery predicted 12-month MCID thresholds with only 62.6% specificity and 86.8% sensitivity.²⁶ As such, the results of this study cannot be used to support the long-term effectiveness of surgical treatment. Other limitations are based on the study design and objectives of this study. Updates of PROMIS CAT are currently being developed to improve its ability to accurately identify and capture functional outcomes data for spine conditions, but these changes may alter calculated correlations. The cost of implementing PROMIS measures in routine clinical care was not explored. Cost is a significant factor and should be addressed in other research studies. Furthermore, some postoperative assessments were completed at home. While this improves data capture by minimizing dropout, it is possible that the different locations affected responses. Lastly, our study was not designed to evaluate test-retest reliability.

Future directions include increasing the sample size, adding a longer follow-up, and potentially integrating updated PROMIS CAT tools for spinal pathology. Despite potential limitations, we believe that PROMIS CATs are an improvement over legacy outcomes measures with strong evidence of convergent validity, responsiveness, known-groups validity, and efficiency. This supports the use of PROMIS CATs in the evaluation of surgical patients with cervical spine disease. Future studies should also assess the effect of real-time scoring and interpretation on physician decision making as well as patient satisfaction.

Conclusions

In a health care climate focused on the value of surgical interventions, physicians require metrics that are easy to administer and capture relevant, meaningful, patient-centered data to better evaluate the outcomes of care. This study demonstrates that PROMIS CATs are valid measures for PF, PI, and PB in patients undergoing surgery for cervical spine disease. Furthermore, PROMIS CATs are an improvement over legacy measures by reducing capture time while also minimizing floor and ceiling effects that have limited the use of patient-reported outcome measures in clinical practice.

ABBREVIATIONS

CAT	computer adaptive test
MCID	minimally important clinical difference
MCS	mental component score

NDI	Neck Disability Index
ODI	Oswestry Disability Index
PB	pain behavior
PCS	physical component score
PF	physical function
PI	pain interference
PRO	patient-reported outcome
PROMIS	Patient-Reported Outcomes Measurement Information System

References

1. Amtmann D, Kim J, Chung H, Bamer AM, Askew RL, Wu S, et al.: Comparing CESD-10, PHQ-9, and PROMIS depression instruments in individuals with multiple sclerosis. *Rehabil Psychol* 59:220–229, 2014 [PubMed: 24661030]
2. Auffinger B, Lam S, Shen J, Roitberg BZ: Measuring surgical outcomes in subaxial degenerative cervical spine disease patients: minimum clinically important difference as a tool for determining meaningful clinical improvement. *Neurosurgery* 74:206–214, 2014 [PubMed: 24220005]
3. Beaton DE: Simple as possible? Or too simple? Possible limits to the universality of the one half standard deviation. *Med Care* 41:593–596, 2003 [PubMed: 12719682]
4. Beckmann JT, Hung M, Bounsanga J, Wylie JD, Granger EK, Tashjian RZ: Psychometric evaluation of the PROMIS Physical Function Computerized Adaptive Test in comparison to the American Shoulder and Elbow Surgeons score and Simple Shoulder Test in patients with rotator cuff disease. *J Shoulder Elbow Surg* 24:1961–1967, 2015 [PubMed: 26321484]
5. Brodke DS, Goz V, Voss MW, Lawrence BD, Spiker WR, Man H: PROMIS® PF CAT outperforms the ODI and SF-36 physical function domain in spine patients. *Spine (Phila Pa 1976)* 42:921–929, 2017 [PubMed: 27792105]
6. Carreon LY, Glassman SD, Campbell MJ, Anderson PA: Neck Disability Index, short form-36 physical component summary, and pain scales for neck and arm pain: the minimum clinically important difference and substantial clinical benefit after cervical spine fusion. *Spine J* 10:469–474, 2010 [PubMed: 20359958]
7. 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* 45 (5 Suppl 1):S3–S11, 2007
8. Choi SW: Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Appl Psychol Meas* 33:644–645, 2009
9. Chotai S, Parker SL, Sivaganesan A, Godil SS, McGirt MJ, Devin CJ: Quality of life and general health after elective surgery for cervical spine pathologies: determining a valid and responsive metric of health state utility. *Neurosurgery* 77:553–560, 2015 [PubMed: 26191977]
10. Copay AG, Martin MM, Subach BR, Carreon LY, Glassman SD, Schuler TC, et al.: Assessment of spine surgery outcomes: inconsistency of change amongst outcome measurements. *Spine J* 10:291–296, 2010 [PubMed: 20171937]
11. Fitzpatrick R, Davey C, Buxton MJ, Jones DR: Evaluating patient-based outcome measures for use in clinical trials. *Health Technol Assess* 2:i–iv, 1–74, 1998
12. Flynn KE, Dew MA, Lin L, Fawzy M, Graham FL, Hahn EA, et al.: Reliability and construct validity of PROMIS® measures for patients with heart failure who undergo heart transplant. *Qual Life Res* 24:2591–2599, 2015 [PubMed: 26038213]
13. Fries JF, Bruce B, Cella D: The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. *Clin Exp Rheumatol* 23 (5 Suppl 39):S53–S57, 2005

14. Godil SS, Parker SL, Zuckerman SL, Mendenhall SK, Devin CJ, Asher AL, et al.: Determining the quality and effectiveness of surgical spine care: patient satisfaction is not a valid proxy. *Spine J* 13:1006–1012, 2013 [PubMed: 23685216]
15. Godil SS, Parker SL, Zuckerman SL, Mendenhall SK, Glassman SD, McGirt MJ: Accurately measuring the quality and effectiveness of lumbar surgery in registry efforts: determining the most valid and responsive instruments. *Spine J* 14:2885–2891, 2014 [PubMed: 24768731]
16. Godil SS, Parker SL, Zuckerman SL, Mendenhall SK, McGirt MJ: Accurately measuring the quality and effectiveness of cervical spine surgery in registry efforts: determining the most valid and responsive instruments. *Spine J* 15:1203–1209, 2015 [PubMed: 24076442]
17. Hung M, Baumhauer JF, Latt LD, Saltzman CL, SooHoo NF, Hunt KJ: Validation of PROMIS® Physical Function computerized adaptive tests for orthopaedic foot and ankle outcome research. *Clin Orthop Relat Res* (471):3466–3474, 2013 [PubMed: 23749433]
18. Hung M, Cheng C, Hon SD, Franklin JD, Lawrence BD, Neese A, et al.: Challenging the norm: further psychometric investigation of the neck disability index. *Spine J* 15:2440–2445, 2015 [PubMed: 24662211]
19. Hung M, Clegg DO, Greene T, Saltzman CL: Evaluation of the PROMIS physical function item bank in orthopaedic patients. *J Orthop Res* 29:947–953, 2011 [PubMed: 21437962]
20. Hung M, Hon SD, Franklin JD, Kendall RW, Lawrence BD, Neese A, et al.: Psychometric properties of the PROMIS physical function item bank in patients with spinal disorders. *Spine (Phila Pa 1976)* 39:158–163, 2014 [PubMed: 24173018]
21. Hung M, Stuart AR, Higgins TF, Saltzman CL, Kubiak EN: Computerized adaptive testing using the PROMIS physical function item bank reduces test burden with less ceiling effects compared with the short musculoskeletal function assessment in orthopaedic trauma patients. *J Orthop Trauma* 28:439–443, 2014 [PubMed: 24378399]
22. Irwin DE, Atwood CA Jr, Hays RD, Spritzer K, Liu H, Donohue JF, et al.: Correlation of PROMIS scales and clinical measures among chronic obstructive pulmonary disease patients with and without exacerbations. *Qual Life Res* 24:999–1009, 2015 [PubMed: 25307510]
23. Jensen RE, Potosky AL, Reeve BB, Hahn E, Cella D, Fries J, et al.: Validation of the PROMIS physical function measures in a diverse US population-based cohort of cancer patients. *Qual Life Res* 24:2333–2344, 2015 [PubMed: 25935353]
24. McCormick JD, Werner BC, Shimer AL: Patient-reported outcome measures in spine surgery. *J Am Acad Orthop Surg* 21:99–107, 2013 [PubMed: 23378373]
25. Norman GR, Sloan JA, Wyrwich KW: Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care* 41:582–592, 2003 [PubMed: 12719681]
26. Parker SL, Asher AL, Godil SS, Devin CJ, McGirt MJ: Patient-reported outcomes 3 months after spine surgery: is it an accurate predictor of 12-month outcome in real-world registry platforms? *Neurosurg Focus* 39(6):E17, 2015
27. Parker SL, Godil SS, Shau DN, Mendenhall SK, McGirt MJ: Assessment of the minimum clinically important difference in pain, disability, and quality of life after anterior cervical discectomy and fusion: clinical article. *J Neurosurg Spine* 18:154–160, 2013 [PubMed: 23176164]
28. Revicki DA, Cella DF: Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. *Qual Life Res* 6:595–600, 1997 [PubMed: 9330558]
29. Walton DM, MacDermid JC: A brief 5-item version of the Neck Disability Index shows good psychometric properties. *Health Qual Life Outcomes* 11:108, 2013 [PubMed: 23816395]
30. Weiss DJ: Computerized adaptive testing for effective and efficient measurement in counseling and education. *Meas Eval Couns Dev* 37:70–84, 2004

FIG. 1. Overall group analysis showing the PROMIS T-scores at baseline (general population mean 50 points).

FIG. 2.
PROMIS T-scores by work and activity limitations.

FIG. 3.
SF-12 scores by work and activity limitations.

FIG. 4.

Change in PROMIS T-scores over the assessment period by subgroup. **A:** Effect sizes for the PROMIS scores versus NDI by subgroup. **B:** Effect sizes for the PROMIS scores versus NDI for the myelopathy subgroup. **C:** Effect sizes for the PROMIS scores versus NDI for the radiculopathy subgroup.

FIG. 5.
Floor and ceiling effects of the PROMIS PF scores.

FIG. 6.
Floor and ceiling effects of the PROMIS PI scores.

FIG. 7.
Floor and ceiling effects of the PROMIS PB scores.

TABLE 1.

Overall group analysis of changes in scores between visits

Assessment	No. of Patients	Mean $\pm$ SD	Median	p Value
Visit 2 vs 1				
PB	51	-3.66 ± 8.16	-2.4	<0.01
PI	50	-4.52 ± 8.86	-4.25	<0.01
PF	50	1.17 ± 8.68	1.25	0.34
NDI-5	45	-6.4 ± 19.49	-4	0.03
SF-12 MCS	44	3.34 ± 13.99	5.72	0.12
SF-12 PCS	44	2.63 ± 11.24	2.65	0.13
Visit 3 vs 2				
PB	50	-0.92 ± 5.12	0	0.21
PI	50	-2.27 ± 6.44	-1.75	0.02
PF	50	1.3 ± 6.55	1.1	0.17
NDI-5	46	-4.61 ± 13.88	-2	0.03
SF-12 MCS	45	2.79 ± 10.91	2.37	0.09
SF-12 PCS	45	1.3 ± 8.02	0.39	0.28
Visit 3 vs 1				
PB	52	-4.35 ± 8.02	-2.85	<0.01
PI	52	-6.62 ± 8.13	-6.15	<0.01
PF	52	2.53 ± 6.71	2.9	<0.01
NDI-5	49	-11.59 ± 15.75	-12	<0.01
SF-12 MCS	49	3.61 ± 13.61	3.08	0.07
SF-12 PCS	49	3.78 ± 11.47	3.27	0.03

TABLE 2.

Overall group analysis of questionnaire scores by limitation in work or activity (Item 3a on SF-12)

Measure	Completely Limited	Limited Most or Some of the Time	Limited Little or None of the Time	P Value
No. of patients	17	32	9	
PB	59.7 ± 7	59.4 ± 4.5	54.8 ± 8.5	0.13
PI	64.8 ± 8.8	62.2 ± 4.9	55 ± 7.5	<0.01
PF	35.6 ± 6.3	40.3 ± 5.8	48.5 ± 10.8	<0.01
NDI-5	43.1 ± 18.6	34.3 ± 14.2	21.5 ± 8.5	<0.01
SF-12 MCS	44.7 ± 13	45 ± 11.7	50.9 ± 10.9	0.43
SF-12 PCS	25.7 ± 6.9	36.6 ± 8.9	49.7 ± 6.9	<0.01

All values are shown as the mean ± SD unless indicated otherwise.

TABLE 3.

Overall group analysis of changes in scores from baseline to 3 months by whether or not other conditions impacted response

Measure	Yes	No	p Value
PB	-4.7 ± 10.7	-6 ± 7.1	0.64
PI	-4.1 ± 7.4	-9 ± 7.7	0.06
PF	0.8 ± 5.9	4.7 ± 7.3	0.09
NDI-5	-5.9 ± 17.8	-17.1 ± 14.2	0.04
SF-12 MCS	0.4 ± 16.6	9.4 ± 9.6	0.06
SF-12 PCS	3.7 ± 15.2	3.6 ± 9.4	0.98

All values are shown as the mean $\pm$ SD unless indicated otherwise.

TABLE 4.

Overall group analysis of changes in scores between visits 1 and 2 by patient-rated change category

Measure	Much Better		Slightly Better or Much Worse	
	Mean $\pm$ SD	SRM	Mean $\pm$ SD	SRM
PB	-5.65 $\pm$ 12.22	-0.46	-1.63 $\pm$ 6	-0.27
PI	-6.44 $\pm$ 12.23	-0.53	-1.52 $\pm$ 6.42	-0.24
PF	5.56 $\pm$ 11.57	0.48	-0.55 $\pm$ 7.87	-0.07
NDI-5	-12.62 $\pm$ 19.72	-0.64	-0.86 $\pm$ 22.53	-0.04
SF-12 MCS	10.42 $\pm$ 10.53	0.99	5.13 $\pm$ 14.49	0.35
SF-12 PCS	4.05 $\pm$ 11.84	0.34	1.02 $\pm$ 12.91	0.08

SRM = standardized response mean.

TABLE 5.

Overall group analysis of changes in scores between visits 2 and 3 by patient-rated change category

Measure	Much Better		Slightly Better or Much Worse	
	Mean $\pm$ SD	SRM	Mean $\pm$ SD	SRM
PB	-2.05 $\pm$ 7.63	-0.27	-1.6 $\pm$ 3.68	-0.43
PI	-3.68 $\pm$ 7.65	-0.48	-2.66 $\pm$ 5.59	-0.48
PF	-2.23 $\pm$ 5.14	-0.43	2.67 $\pm$ 6.03	0.44
NDI-5	-1.54 $\pm$ 12.17	-0.13	-8.21 $\pm$ 15.11	-0.54
SF-12 MCS	0.74 $\pm$ 13.11	0.06	3.17 $\pm$ 8.6	0.37
SF-12 PCS	1.77 $\pm$ 9.18	0.19	0.03 $\pm$ 7.67	0.00

TABLE 6.

Time to completion of outcomes measures

Measure	No. of Patients	Mean $\pm$ SD*	Median*
PB	57	0.9 $\pm$ 0.5	0.8
PI	56	1.2 $\pm$ 1.9	0.7
PF	57	1.1 $\pm$ 1.4	0.8
NDI	58	3.4 $\pm$ 2.6	2.5
SF-12	58	4.1 $\pm$ 3.9	2.7

* Values are shown in minutes.