The Spine Journal 19 (2019) 34–40

Clinical Study

Responsiveness of the Patient-Reported Outcomes Measurement Information System (PROMIS), Neck Disability Index (NDI) and Oswestry Disability Index (ODI) instruments in patients with spinal disorders

Man Hung, PhD*, Charles L. Saltzman, MD, Maren W. Voss, ScD, Jerry Bounsanga, BS, Richard Kendall, DO, Ryan Spiker, MD, Brandon Lawrence, MD, Darrel Brodke, MD

School of Medicine, University of Utah, 590 Wakara Way, Salt Lake City, UT 84108, United States

Received 17 April 2018; revised 19 June 2018; accepted 22 June 2018

Abstract

BACKGROUND CONTEXT: The Patient-Reported Outcomes Information System (PROMIS) instruments are an important advancement in the use of PROs, but need to be evaluated with longitudinal data to determine whether they are responsive to change in specific clinical populations.

PURPOSE: The purpose of this study was to assess the responsiveness of the PROMIS Physical Function (PF), PROMIS Pain Interference (PI), Neck Disability Index (NDI), and the Oswestry Disability Index (ODI).

STUDY DESIGN/SETTING: This study entailed prospective data collection from consecutive patients aged 18 and older, visiting a university-based orthopaedic spine clinic between October 2013 and January 2017.

PATIENT SAMPLE: A total of 763 participants in the sample had a mean age of 58 (SD = 15) years and the sample was 50.2% male and 92.8% Caucasian.

OUTCOME MEASURES: The PROMIS PF and PROMIS PI Computerized Adaptive Tests along with either the NDI or ODI instruments were administered on tablet computers before clinic visits. Global rating of change questions relating to pain and function levels was also administered.

METHODS: Baseline scores were compared with follow-up scores at four different time-points from 3-months to 6-months and beyond. Patient demographics, mean scores, paired-sample *t* tests, Standardized Response Mean (SRM), and Effect Size (ES) were analyzed to determine instrument responsiveness. This project was funded by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health under award number U01AR067138 and the authors have no conflicts of interest to disclose.

RESULTS: The PROMIS instruments were strongly correlated with each other as well as with the NDI and ODI. Responsiveness was significant on all four instruments at every time-point assessed (paired sample *t* tests ranged from $p < .001$ to $p = .049$). SRM's were large and over 0.94 for every instrument at every time-point. Cohen's d ES were large and over 0.96 for all at all time-points, except for the NDI which had ES ranging from 0.74 to 0.83. This study showed large effect sizes and responsiveness of the PROMIS PF, PROMIS PI, NDI and ODI in a population of orthopaedic patients with spine pathologies.

FDA device/drug status: Not applicable.

Author disclosure: ***MH***: Nothing to disclose. ***CLS***: Nothing to disclose. ***MWV***: Nothing to disclose. ***JB***: Nothing to disclose. ***RK***: Nothing to disclose. ***RS***: Nothing to disclose. ***BL***: Nothing to disclose. ***DB***: Nothing to disclose. Level of support is level G.

* Corresponding author. University of Utah Department of Orthopaedics, 590 Wakara Way, Salt Lake City, UT. 84108, United States. Tel.: 801-587-5372; Fax: 801-587-5411.

E-mail address: Man.Hung@hsc.utah.edu (M. Hung).

https://doi.org/10.1016/j.spinee.2018.06.355

CONCLUSION: This study demonstrates strong responsiveness of the PROMIS PF and PROMIS PI in a spine clinic population.

Keywords: NDI; ODI; Orthopaedics; Patient-reported outcomes; PROMIS; Responsiveness; Spine.

Introduction

Patient-reported outcome (PRO) instruments are recognized as an important aspect of high quality medical care. These instruments can inform medical personnel of the patient perspective, can provide reliable and valid clinical information, and can potentially improve the patient experience if conducted with a minimum of respondent burden [1]. To enhance the effectiveness of PRO measures, the Patient-Reported Outcomes Measurement Information System (PROMIS) developers conducted extensive reviews of existing PROs in clinical practice, analyzed the test items, and revised items when appropriate to create improved instruments [2]. The PROMIS instruments were developed using item-response theory (IRT), where individual test items are calibrated and validities and reliabilities are assessed. One advantage of this method is that IRT allows algorithms to be efficiently used in a computerized adaptive test (CAT) format [3]. This method limits redundancy, as early item responses inform later item selection, minimizing test burden while simultaneously maintaining the precision of the information [4–7]. The PROMIS instruments with CAT administration are an important advancement in the use of PROs in clinical practice [8].

One important aspect of PRO development is the ability to detect treatment related changes that occur over time, referred to as the responsiveness of the instrument [9]. Determining responsiveness requires longitudinal data with repeated measures at multiple time-points on the same individuals [10]. Responsiveness can be assessed in two ways, using either internal or external methods. Internal methods evaluate the level of change based on the magnitude of the difference in scores [9]. External methods, on the other hand, provide information on whether the level of change is meaningful by anchoring the change score with some other measure of treatment response [11]. Both methods can be useful in determining the ability of a measure to detect changes in patient outcomes.

The PROMIS instruments are satisfactory in terms of reliabilities and validities [4,5]. They have demonstrated sound psychometric properties specifically in a spine population [12,13]. Given the recent development of the PROMIS measures, and the time it takes to gather longitudinal data with repeated measures, research is only beginning to address the responsiveness of PROMIS instruments in specific patient populations. Responsiveness studies on the PROMIS measures in a general population have shown that they have excellent sensitivity to change [14], with up to four times the sensitivity of similar instruments [15]. However, responsiveness analysis of the PROMIS measures has not yet been conducted for orthopaedic spine patients.

The newly developed PROMIS instruments may not fully replace the use of other condition or region-specific PROs in clinical practice, thus it is useful to evaluate responsiveness of new and previously used instruments side-by-side. The Neck Disability Index (NDI) and the Oswestry Disability Index (ODI) are commonly used instruments in orthopaedics [12,16]. The NDI has overall shown questionable psychometric properties, even though it is the most widely used PRO for neck disorders [17–19]. The responsiveness to change of the NDI has been questioned, as studies over longer intervals suggest that the episodic nature of neck pain may limit the tools ability to accurately measure treatment effects [19]. The ODI has shown good to fair psychometric properties when validated both with classical test theory [20–22] and with the modern IRT approach [23]. Given the common use of these instruments in orthopaedic spine populations, there is value in comparing the responsiveness of these instruments with the newer PROMIS measures in the same orthopaedic patient sample.

The purpose of the present study was to examine the responsiveness to change for two PROMIS instruments, the ODI, and the NDI in an orthopaedic spinal population.

Materials and methods

Sample

All patients aged 18 and older seeking orthopaedic care for spinal conditions at a university clinic between November 2013 and January 2017 were enrolled if they were seen for follow-up care greater than 3 months after their initial visit. The PRO measures were administered on handheld tablet computers before the clinic visit both at baseline and follow-up visits. Follow-up time periods were categorized into four groupings including 3-month follow-up (80 to 100 days after initial assessment), greater than 3-month follow-up (90 days or more after initial assessment), 6-month follow-up (170 to 190 days after initial assessment), and greater than 6-month follow-up (180 days or more after initial assessment). The 3- and 6-month follow-up were selected as common treatment time-points in orthopaedic practice [24–31]. The greater than 3- and 6-month follow-up periods are sometimes used in research to measure longer-term outcomes [32–34]. As patients sought follow-up at different time-points depending on their individual treatment programs, the analysis at each time-point necessarily included different patient groupings. Demographics were reported for the entire patient sample. Institutional review board approval was obtained before the start of the study.

PRO instruments

Patients took three instruments including the PROMIS Physical Function (PF) v1.2, the PROMIS Pain Interference (PI) v1.1, and the NDI or the ODI, depending on patient condition. The PROMIS instruments were delivered through a web-based portal using CAT administration, with standard algorithms applied to select item administration until the standard error of measurement (SEM) was less than 0.33. The CAT administration adjusts item selection based on prior responses to minimize patient burden. For example, a patient who had answered 'unable to do' to the question 'Are you able to peel fruit?' would not then be asked about more strenuous tasks. Patients were electronically administered the full set of questions in the NDI and ODI.

The PROMIS PF v1.2 CAT was drawn from the 121-item full test bank which contains items in four domains including lower extremity, upper extremity, central (back and neck), and activities of daily living. Lower scores on the PROMIS PF are indicative of lower patient functioning. The PROMIS PI v1.1 has a 40-item bank with items related to how pain interferes with daily activities. Lower scores on the PROMIS PI are indicative of better patient functioning, with less pain interference. Both PROMIS instruments are standardized in *T* scores (Mean = 50; SD = 10) and are calibrated in a general population augmented with oversampling of some diagnostic groups [21]. The NDI and the ODI both consist of 10-items related to the neck and low back function, respectively. Scores on the ODI and NDI range from 0 to 100, with higher scores representing lower functioning levels and increased disability. The instruments were administered at the first clinic visit (ie, either within seven days before the clinic visit of a new spinal condition or on the day of the first clinic visit) and at each follow-up clinic visit.

Analyses

Change scores were calculated as the absolute difference between the baseline score and the follow-up score. Meaningful change was defined from the patient perspective as improvement in condition so that the analysis of responsiveness to change can be interpreted as meaningful levels of change [35]. The anchor question to measure physical function responsiveness was: 'Compared to your FIRST EVALUATION at the xxx: how would you describe your physical function now?' (much worse, worse, slightly worse, no change, slightly improved, improved, and much improved). Similarly, the anchor question used to measure pain interference responsiveness was: 'Compared to your FIRST EVALUATION at the xxx: how would you describe your episodes of PAIN now?' Global ratings of change (GRC) scales such as this are commonly used in orthopaedics [36]. This type of scale relies on retrospective reflection, which may be only weakly correlated with treatment effect [37], thus responsiveness analysis was comprehensive including both internal and external methods to address the limitations of any one analytic approach. The GRC anchors change scores to the patient perception of improvement, so that responsiveness can be assessed in terms of meaningful change from the patients' perspective.

Paired sample *t* tests were run on each instrument at each follow-up visit to evaluate the hypothesis that there was no-change between baseline and follow-up. Significance was determined a priori at p-value = .05, two sided. Correlations between each measure were calculated with Pearson's Product Moment Correlation using SPSS 24.0. For effect size (ES) we used the standardized Cohen's d, which takes into consideration the variability that exists in scores [9]. Cohen's d removes dependence on sample size and is normalized using the crosssectional standard deviation (SD) of scores. We calculated Cohen's d using the score difference between the baseline score and the follow-up score, divided by the baseline score's SD. Cohen's d can be interpreted as d = 0.20 indicating a small effect, d = 0.50 as a medium effect, and d = 0.80 as a large effect. A 0.80 effect represents a change where the difference is as least as great as 4 per 5 of a SD in scores [9].

We also calculated the standardized response mean (SRM) as an indicator of ES. SRM removes the dependence on sample size from the equation that is seen as a factor in the *t* test calculations [9]. The SRM is normalized based on the SD of the change score. SRM is calculated as the mean difference between baseline and follow-up scores divided by the SD of the difference score, reflecting individual changes in scores. The SRM values can be interpreted in the same way that we interpret Cohen's d, with values of 0.20, 0.50, and 0.80 for small, medium, and large effect, respectively [9].

Analyses were performed using SPSS 24.0 (IBM SPSS Statistics for Windows, Armonk, NY: IBM Corp.) [38], and R 3.30 (R Development Core Team, Vienna, AT: R Foundation for Statistical Computing) [39].

Results

The sample included 763 patients with an average age of 58.26 (SD = 14.72; Range = 18 to 89). It had 50.2% male (n = 383) and majority were White (n = 708; 92.8%), with 2.9% (n = 22) reporting Hispanic ethnicity (see Table 1). Patients were treated for multiple procedures including vertebral process or body fractures and removal procedures on the musculoskeletal system, among others. There was insufficient sample in each procedure or diagnostic code for meaningful stratification by condition.

Paired samples t *test*

The responsiveness of all four instruments was significant at every time-point as measured by paired sample *t* tests (see Table 2). For the PROMIS PF the statistical significance ranged from $p < .001$ at greater than 3-month and greater than 6-month follow-ups to p = .049 at 3-months. For the PROMIS PI the significance was $p < .001$ at every time-point. For the NDI significance values ranged from

Table 1
Patient demographics of included subjects (n = 763).

Variables	n (%)	Minimum/ Maximum	Mean (SD)
Age (years)		18/89	58.26 (14.72)
Gender			
Male	383 (50.2)		
Female	380 (49.8)		
Race			
White or Caucasian	708 (92.8)		
Asian	8 (1.0)		
American Indian and Alaska Native	1 (0.1)		
Native Hawaiian/Other Pacific Islander	3 (0.4)		
Black or African American	10 (1.3)		
Other	28 (3.7)		
Unknown/Missing	5 (0.7)		
Ethnicity			
Hispanic	22 (2.9)		
NonHispanic	733 (96.1)		
Missing	8 (1.0)		

$p < .001$ at greater than 3- and greater than 6-month follow-ups to $p = .042$ at 6-months. For the ODI significance values ranged from $p < .001$ at greater than 3- and greater than 6-month follow-ups to $p = .009$ at 3-months. Sample sizes ranged from 10 to 590 in the analysis of responsiveness in paired comparisons, depending on the instrument and the time-point evaluated.

Effect size and SRM

The effect sizes of all four instruments were large at every follow-up time-point as measured by both SRM and Cohen's d ES. For the PROMIS PF the SRM ranged from 0.97 (n = 34) at 6-months to 1.31 (n = 87) at 3-months and the ES ranged from 0.97 (n = 565) at greater than 3-months to 1.11 (n = 34) at 6-months (see Table 2). For the PROMIS PI, the SRM ranged from 0.94 (n = 44) at 6-months to 1.31 (n = 590) at greater than 3-months and the ES ranged from 1.12 (n = 390) at greater than 6-months to 1.39 (n = 93) at 3-months. For the NDI, the SRM ranged from 1.12 (n = 10) at 6-months to 1.27 (n = 105) at greater than 3-months and the ES ranged from 0.74 (n = 10) 6-months to 0.83 (n = 105) at greater than 3-months. For the ODI the SRM ranged from 1.16 (n = 69) at 3-months to 1.33 (n = 29) at 6-months and the ES ranged from 0.96 (n = 323) at greater than 6-months to 1.08 (n = 29) at greater than 6-months.

Mean change and correlations

The mean baseline score for the PROMIS PF was 35.88 (SD = 6.67; n = 87), and was 65.87 (SD = 6.12; n = 93) for the PROMIS PI, 45.64 (SD = 19.66; n = 22) for the NDI, and 41.85 (SD = 17.04; n = 69) for the ODI (see Appendix A). Correlations between the PROMIS PF and PROMIS PI were significant at every time-point and ranged from −0.56 to −0.72 (see Table 3). Because higher scores on the PROMIS PF indicate higher function, whereas higher scores on the PROMIS PI, NDI, and ODI indicate more disability, strong negative correlations indicate the same

Table 2
Responsiveness of the PROMIS instruments, NDI, and ODI.

Measurement	Follow-up period	n	SRM	ES	Paired-*t* test p-value
	3-month follow-up*				
PROMIS PF		87	1.31	0.98	0.049
PROMIS PI		93	1.16	1.39	<0.001
NDI		22	1.18	0.76	0.008
ODI		69	1.16	1.03	0.009
	>3-month follow-up†				
PROMIS PF		565	1.07	0.97	<0.001
PROMIS PI		590	1.31	1.19	<0.001
NDI		105	1.27	0.83	<0.001
ODI		472	1.26	1.00	< 0.001
	6-month follow-up‡				
PROMIS PF		34	0.97	1.11	0.004
PROMIS PI		44	0.94	1.29	0.001
NDI		10	1.12	0.74	0.042
ODI		29	1.33	1.08	0.002
	>6-month follow-up§				
PROMIS PF		385	1.03	0.98	<0.001
PROMIS PI		390	1.12	1.12	<0.001
NDI		67	1.18	0.82	<0.001
ODI		323	1.30	0.96	<0.001

PF = Physical Function; PI = Pain Interference; NDI = Neck Disability Index; ODI = Oswestry Disability Index.
* 90 ± 10 days.
† ≥ 90 days.
‡ 180 ± 10 days.
§ ≥180 days.

Table 3
Correlations between PROMIS instruments, NDI, and ODI scores.

Baseline				
	PROMIS PF	PROMIS PI	NDI	ODI
PROMIS PF	–	−0.56*	−0.60*	−0.66*
PROMIS PI	−0.56*	–	0.71*	0.59*
NDI	−0.60*	0.71*	–	–
ODI	−0.66*	0.59*	–	–
80–100 days follow-up				
	PROMIS PF	PROMIS PI	NDI	ODI
PROMIS PF	–	−0.72*	−0.62*	−0.76*
PROMIS PI	−0.72*	–	0.81*	0.79*
NDI	−0.62*	0.81*	–	–
ODI	−0.76*	0.79*	–	–
≥ 90 days follow-up				
	PROMIS PF	PROMIS PI	NDI	ODI
PROMIS PF	–	−0.66*	−0.67*	−0.74*
PROMIS PI	−0.66*	–	0.75*	0.72*
NDI	−0.67*	0.75*	–	–
ODI	−0.74*	0.72*	–	–
170–190 days follow-up				
	PROMIS PF	PROMIS PI	NDI	ODI
PROMIS PF	–	−0.70*	−0.72*	−0.80*
PROMIS PI	−0.70*	–	0.74*	0.83*
NDI	−0.72*	0.74*	–	–
ODI	−0.80*	0.83*	–	–
≥180 days follow-up				
	PROMIS PF	PROMIS PI	NDI	ODI
PROMIS PF	–	−0.66*	−0.65*	−0.80*
PROMIS PI	−0.66*	–	0.73*,*	0.73*
NDI	−0.65*	0.73*	–	–
ODI	−0.80*	0.73*	–	–

PF = Physical Function; lower score = lower patient functioning; PI = Pain Interference; higher score = lower patient functioning; NDI = Neck Disability Index; higher score = lower patient functioning; ODI = Oswestry Disability Index; higher score = lower patient functioning.

Shaded cells are "not applicable" or "redundant.".

* Pearson correlation is significant at the 0.01 level (2-tailed).

clinical interpretation between measures. Correlations between PROMIS PF and the NDI and ODI were significant at every time-point and ranged from −0.60 to −0.72 for the NDI and from −0.66 to −0.80 for the ODI. Correlations between the PROMIS PI and the NDI and ODI were significant at every time-point and ranged from 0.71 to 0.81 for the NDI and 0.59 to 0.83 for the ODI. The correlation between the NDI and ODI was not analyzed as patients did not typically complete both instruments during the same office visit (see Table 3).

Discussion

This study demonstrated strong responsiveness of the PROMIS PF CAT, PROMIS PI CAT, NDI and ODI in a population of patients visiting spine specialists at a university clinic. Past research using a reliable change index calculated from the scale's standard error of measurement found that the PROMIS PF and PI short forms had adequate sensitivity to change [40]. Yet the responsiveness of the PROMIS Pain Interference (PI) short-form was lower than for other pain instruments in a musculoskeletal pain group [41]. The current study shows a strong effect for the PROMIS PI CAT in detecting change in spine patients, though it was not compared directly to other pain instruments. The responsiveness of the NDI has been called into question in past research [19] and the ODI has shown good responsiveness and fair psychometric properties in the past [20–23], yet both showed strong responsiveness in this spine clinic population. Additionally, this study is the first to demonstrate the responsiveness of the PROMIS PF CAT and PROMIS PI CAT in a spine patient population.

There are factors related to the nature of medical visit follow-up that are potential limitations, given that patients who return for follow-up care may have needs and conditions that differ from the full population of spine patients. The sample analyzed might not be representative of the US population demographics as a whole, and might not be generalizable beyond the sample characteristics. Despite limitations in sample characteristics and follow-up conditions, this sample is typical of patients seen in spine clinics in the region and can be generalizable to similar clinical practices. Future research should examine the responsiveness of these instruments across other patient populations. Additionally, because we were not able to stratify the results by age, procedure code or diagnosis due to limited sample size per procedure, we plan to return to this topic and examine the responsiveness of these instruments for specific diagnostic conditions in the future when data are available.

Conclusions

Each analytical method to examine responsiveness has different strengths in measuring the sensitivity to change of the PRO instruments. The present analysis of change was based on the response to a GRC anchor question. The GRC rely on patient retrospection and can be subject to recall bias, even though it represents the patients' perspective on meaningful change. The calculation of mean change scores reflect patient reports of pain and function both at baseline and follow-up visits, and are less subject to recall bias. This study included distribution-based internal responsiveness analysis as well as the external anchor-based responsiveness to provide a comprehensive assessment of instrument responsiveness in this population. The PROMIS PF, PROMIS PI, NDI, and ODI all showed high responsiveness to change, large effect sizes, and significant differences in mean change scores. The current study included *t* tests, SRM and ES as three different methods of testing for responsiveness and found substantial agreement across these methods and across time-points used for follow-up. These findings provide confidence in using these four measures to evaluate treatment related changes in patient condition for spine practices.

Acknowledgment

This project was funded by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the

National Institutes of Health under award number U01AR067138. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Supplementary material

Supplementary material related to this article can be found at http://dx.doi.org/10.1016/j.spinee.2018.06.355.

References

[1] Deutsch L, Smith L, Gage B, Kelleher C, Garfinkel D. Patient-reported outcomes in performance measurement: commissioned paper on PRO-based performance measures for healthcare accountable entities. Paper presented at: Washington, DC: National Quality Forum 2012.

[2] DeWalt DA, Rothrock N, Yount S, Stone AA. Evaluation of item candidates: the PROMIS qualitative item review. Med Care 2007;45(5 Suppl 1):S12–21.

[3] Brodke DJ, Hung M, Bozic KJ. Item response theory and computerized adaptive testing for orthopaedic outcomes measures. J Am Acad Orthop Surg 2016;24(11):750–4.

[4] Hung M, Franklin JD, Hon SD, Cheng C, Conrad J, Saltzman CL. Time for a paradigm shift with computerized adaptive testing of general physical function outcomes measurements. Foot Ankle Int 2014;35(1):1–7.

[5] 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 2014;28(8):439–43.

[6] Choi SW. Firestar: computerized adaptive testing simulation program for polytomous item response theory models. Appl Psychol Meas 2009;33(8):644–5.

[7] Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. Health Technol Assess 1998;2(14):i–iv. 1-74.

[8] Cella D, Riley W, Stone A, 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(11):1179–94.

[9] Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 2000;53(5):459–68.

[10] 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;4:70.

[11] Wyrwich K, Norquist J, Lenderking W, Acaster S. Research IACoISfQoL. Methods for interpreting change over time in patient-reported outcome measures. Qual Life Res 2013;22(3):475–83.

[12] Hung M, Hon SD, Franklin JD, et al. Psychometric properties of the PROMIS physical function item bank in patients with spinal disorders. Spine (Phila Pa 1976) 2014;39(2):158–63.

[13] Brodke DS, Goz V, Voss MW, Lawrence BD, Spiker WR, Man H. PROMIS(R) PF CAT outperforms the ODI and SF-36 physical function domain in spine patients. Spine (Phila Pa 1976) 2016;42 (12):921–9.

[14] 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.

[15] Fries J, Rose M, Krishnan E. The PROMIS of better outcome assessment: responsiveness, floor and ceiling effects, and Internet administration. J Rheumatol 2011;38(8):1759–64.

[16] McCormick JD, Werner BC, Shimer AL. Patient-reported outcome measures in spine surgery. J Am Acad Orthop Surg 2013;21(2): 99–107.

[17] Hung M, Cheng C, Hon SD, et al. Challenging the norm: further psychometric investigation of the neck disability index. Spine J 2015;15 (11):2440–5.

[18] Young IA, Cleland JA, Michener LA, Brown C. Reliability, construct validity, and responsiveness of the neck disability index, patient-specific functional scale, and numeric pain rating scale in patients with cervical radiculopathy. Am J Phys Med Rehabil 2010;89(10):831–9.

[19] MacDermid JC, Walton DM, Avery S, et al. Measurement properties of the neck disability index: a systematic review. J Orthop Sports Phys Ther 2009;39(5):400–17.

[20] Davidson M, Keating JL. A comparison of five low back disability questionnaires: reliability and responsiveness. Phys Ther 2002;82 (1):8–24.

[21] Roland M, Fairbank J. The Roland−Morris disability questionnaire and the Oswestry disability questionnaire. Spine (Phila Pa 1976) 2000;25(24):3115–24.

[22] Mannion AF, Junge A, Grob D, Dvorak J, Fairbank JC. Development of a German version of the Oswestry Disability Index. Part 2: sensitivity to change after spinal surgery. Eur Spine J 2006;15 (1):66–73.

[23] Brodke DS, Goz V, Lawrence BD, Spiker WR, Neese A, Hung M. Oswestry Disability Index: a psychometric analysis with 1,610 patients. Spine J 2017;17(3):321–7.

[24] Paatelma M, Kilpikoski S, Simonen R, Heinonen A, Alen M, Videman T. Orthopaedic manual therapy, McKenzie method or advice only for low back pain in working adults: a randomized controlled trial with one year follow-up. J Rehabil Med 2008;40(10):858–63.

[25] Uchiyama S, Imaeda T, Toh S, et al. Comparison of responsiveness of the Japanese Society for Surgery of the Hand version of the carpal tunnel syndrome instrument to surgical treatment with DASH, SF-36, and physical findings. J Orthop Sci 2007;12(3):249–53.

[26] Carmont MR, Silbernagel KG, Nilsson-Helander K, Mei-Dan O, Karlsson J, Maffulli N. Cross cultural adaptation of the Achilles tendon Total Rupture Score with reliability, validity and responsiveness evaluation. Knee Surg Sports Traumatol Arthrosc 2013;21(6):1356–60.

[27] Landauer F, Wimmer C, Behensky H. Estimating the final outcome of brace treatment for idiopathic thoracic scoliosis at 6-month follow-up. Pediatr Rehabil 2003;6(3−4):201–7.

[28] Little DG, MacDonald D. The use of the percentage change in Oswestry Disability Index score as an outcome measure in lumbar spinal surgery. Spine (Phila Pa 1976) 1994;19(19):2139–43.

[29] Cornell CN, Levine D, O'Doherty J, Lyden J. Unipolar versus bipolar hemiarthroplasty for the treatment of femoral neck fractures in the elderly. Clin Orthop Relat Res 1998(348):67–71.

[30] Kotsis SV, Chung KC. Responsiveness of the Michigan hand outcomes questionnaire and the disabilities of the arm, shoulder and hand questionnaire in carpal tunnel surgery. J Hand Surg Am 2005;30(1):81–6.

[31] MacDermid JC, Richards RS, Donner A, Bellamy N, Roth JH. Responsiveness of the short form-36, disability of the arm, shoulder, and hand questionnaire, patient-rated wrist evaluation, and physical impairment measurements in evaluating recovery after a distal radius fracture. J Hand Surg Am 2000;25(2):330–40.

[32] Ibrahim T, Beiri A, Azzabi M, Best AJ, Taylor GJ, Menon DK. Reliability and validity of the subjective component of the American orthopaedic foot and ankle society clinical rating scales. J Foot Ankle Surg 2007;46(2):65–74.

[33] Segal NA, Glass NA, Teran-Yengle P, Singh B, Wallace RB, Yack HJ. Intensive gait training for older adults with symptomatic knee

osteoarthritis. Am J Phys Med Rehabilit/Assoc Acad Physiatr 2015;94(10 Suppl 1):848–58.
[34] Gregory J, Harwood D, Gochanour E, Sherman S, Romeo A. Clinical outcomes of revision biceps tenodesis. Int J Shoulder Surg 2012;6 (2):45.
[35] Gummesson C, Atroshi I, Ekdahl C. The disabilities of the arm, shoulder and hand (DASH) outcome questionnaire: longitudinal construct validity and measuring self-rated health change after surgery. BMC Musculoskelet Disord 2003;4:11.
[36] Kamper SJ, Maher CG, Mackay G. Global rating of change scales: a review of strengths and weaknesses and considerations for design. J Man Manip Ther 2009;17(3):163–70.
[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(8):869–79.
[38] Armonk, NY: IBM Corp.; 2015.
[39] Vienna, Austria: R Foundation for Statistical Computing; 2010.
[40] Shahgholi L, Yost KJ, Kallmes DF. Correlation of the National Institutes of Health patient reported outcomes measurement information system scales and standard pain and functional outcomes in spine augmentation. AJNR Am J Neuroradiol 2012;33(11):2186–90.
[41] Kean J, Monahan PO, Kroenke K, et al. Comparative responsiveness of the PROMIS pain interference short forms, brief pain inventory, PEG, and SF-36 bodily pain subscale. Med Care 2016;54(4):414–21.