INTERVENTIONAL PAIN & SPINE MEDICINE SECTION

Support for the Reliability and Validity of the National Institutes of Health Impact Stratification Score in a Sample of Active-Duty U.S. Military Personnel with Low Back Pain

Ron D. Hays [iD], PhD,* Maria Orlando Edelen, PhD,[†,‡] Anthony Rodriguez, PhD,[†] and Patricia Herman [iD], PhD[§]

*Division of General Internal Medicine & Health Services Research, UCLA Department of Medicine, Los Angeles, California; [†]RAND Corporation, Boston, Massachusetts; [‡]Patient Reported Outcomes, Value and Experience (PROVE) Center, Department of Surgery, Brigham and Women's Hospital, Boston, Massachusetts; and [§]RAND Corporation, Santa Monica, California, USA

Correspondence to: Ron D. Hays, PhD, Division of General Internal Medicine & Health Services Research, UCLA Department of Medicine, 1100 Glendon Avenue, Los Angeles, CA 90024, USA. E-mail: drhays@ucla.edu.

Disclosures and conflicts of interest: There are no conflicts of interest to report.

Abstract

Objective. Evaluate the Impact Stratification Score (ISS) measure of low back pain impact that assesses physical function, pain interference, and pain intensity. **Design**. Secondary analyses of a prospective comparative effectiveness trial of active-duty military personnel with low back pain. **Setting**. A Naval hospital at a military training site (Pensacola, Florida) and two military medical centers: 1) Walter Reed National Military Medical Center (Bethesda, Maryland); and 2) San Diego Naval Medical Center. **Subjects**. The 749 active-duty military personnel had an average age of 31 years, 76% were male, and 67% were white. **Methods**. Participants completed questionnaires at baseline, 6 weeks later, and 12 weeks later. Measures included the ISS, Roland-Morris Disability Questionnaire (RMDQ), PROMIS-29 v1.0 satisfaction with social role participation scale, and single-item ratings of average pain, feeling bothered by low back pain in the past week, and a rating of change in low back pain. **Results**. Internal consistency reliability for the ISS was 0.92–0.93 at the three time points. The ISS correlated 0.75 to 0.84 with the RMDQ, 0.51 to 0.78 with the single-item ratings, and −0.64 to −0.71 with satisfaction with social role participation. The ISS was responsive to change on the three single items. The area under the curve for the ISS predicting improvement on the rating of change from baseline to 6 weeks later was 0.83. **Conclusions**. This study provides support for the reliability and validity of the ISS as a patient-reported summary measure for acute, subacute, and chronic low back pain. The ISS is a useful indicator of low back impact.

Key Words: Back Pain; Measurement; Quality of Life; Research

Introduction

The National Institutes of Health (NIH) Pain Consortium steering committee convened a research task force that defined pain impact in terms of pain intensity, interference with activities, and physical function. The task force proposed an Impact Stratification Score (ISS) for chronic low back pain [1, 2] consisting of nine

Patient-Reported Outcomes Measurement Information System (PROMIS®)-29 items [3]. The ISS represents the physical health dimension of the PROMIS-29 [4] with 4 physical function items, 4 pain interference items and 1 pain intensity item. Physical function (*without any difficulty* = 1 to *unable to do* = 5) and pain interference (*not at all* = 1 to *very much* = 5) each contribute from 4 to 20

points, and pain intensity (0–10 rating) contributes from 0 to 10 points. The ISS has a possible range of 8 (least pain impact) to 50 (greatest pain impact).

Some psychometric support for the ISS has been published, but the samples have been limited to chronic pain and the number of patients included in studies has been small. The PROMIS measure that is the basis of the ISS was designed to be applicable to the general population and patients with different conditions and severity, not limited to those with chronic low back pain. The research task force conducted a "preliminary assessment of its validity" (1, p. 2037) in a sample of 218 patients that received epidural steroid injections for chronic low back pain. Spearman correlations of 0.81 and 0.66, respectively, were reported between the ISS and the Roland-Morris Disability Questionnaire (RMDQ) and the Oswestry Disability Index. In addition, the ISS was more responsive to change than the RMDQ (0.75 vs 0.41 standardized response mean) to epidural steroid injections. In a study of 198 patients with chronic musculoskeletal pain, internal consistency reliability (standardized coefficient alpha) of 0.91 and an intraclass test-retest correlation coefficient of 0.73 over 3 months (among 91 patients who reported that their pain was "about the same" now compared to 3 months before) were reported for the ISS [5]. The ISS was monotonically associated with patient reports of how much worse their pain was at 3 months compared to baseline. Moreover, the ISS was significantly higher for those on worker's compensation and individuals who had fallen in the prior 3 months.

Despite the encouraging preliminary work, further assessment of the reliability and validity of the ISS is needed. This paper assesses the reliability and construct validity of the ISS in a prospective, multisite, parallel-group comparison effectiveness clinical trial of 749 active-duty US military personnel [6, 7]. This data set enables an assessment of the ISS for patients with acute and subacute as well as chronic low back pain. In addition, we estimate responsiveness to change of the ISS based on multiple external indicators (anchors) of change.

Material and Methods

Setting
This is a secondary analyses of data collected at one small hospital at a military training site (Naval Hospital in Pensacola, Florida) and two large military medical centers in major metropolitan areas: 1) Walter Reed National Military Medical Center in Bethesda, Maryland; and 2) Naval Medical Center in San Diego, California.

Sample
The sample characteristics were reported in Goertz et al. [7]. Average age of the 749 study participants was 31 years, 76% were male, and 67% were white. Most of the participants reported low back pain for more than

3 months (chronic low back pain, 51%), but the sample also included those with acute (38%) and subacute (11%) low back pain.

Measures
Questionnaires were administered at baseline, 6-weeks later, and 12-weeks later. At baseline, study participants were asked to report the recency of the start of the current episode of low back pain to identify acute (<1 month), subacute (1–3 months), and chronic (>3 months) low back pain. At each of the three time points, the sample completed the ISS, RMDQ, and the PROMIS®-29 v1.0 satisfaction with social role participation scale. As noted above, two prior studies [1, 5] provided support for the reliability and validity of the ISS. The RMDQ is a 24-item measure assessing physical disability due to lower back pain [8]. The items are dichotomous and reflect physical function content directly relevant to lower back pain. The RMDQ yields a total summary score ranging from 0 (no disability) to 24 (maximum disability). Adults with chronic conditions have been shown to report less satisfaction with participation on the PROMIS scale [9], and the scale is responsive to change among patients with back pain [10].

Questions adapted from existing measures [11, 12] were administered at all three time points to assess being bothered by low back pain in the past week, average low back pain in the past week, and worst low back pain in the past 24 hours: 1) During the past week, how bothersome has low back pain been? (*not at all bothersome*; *slightly bothersome*; *moderately bothersome*; *very bothersome*; *extremely bothersome*); 2) Select the number that best describes your average low back pain during the past week with a 0 (no pain) to 10 (worst possible pain) numerical rating scale; and 3) Select the number that best describes your low back pain at its worst during the past 24 hours using the same 0–10 numerical rating scale. A retrospective rating of change in pain was administered at the 6-week post-baseline assessment: Compared to your first visit, your low back pain is: *much worse, a little worse, about the same, a little better, moderately better, much better,* or *completely gone.*

Analysis Plan
Reliability. We estimate internal consistency reliability for the ISS in the sample overall and test-retest reliability from baseline to 6 weeks later for the subset of subjects reporting they are "about the same" on the self-reported change in low back pain item that was administered at the 6-week assessment. We also report reliability within pain duration subgroups (acute, subacute, chronic).

Validity. We estimate product-moment correlations between the ISS and other variables hypothesized to be associated with it in the overall sample and for the chronic low back pain subgroup: RMDQ, bothered by low back pain in the past week, average low back pain in

the past week, worst low back pain in the past 24 hours), the PROMIS-29 satisfaction with social role participation scale, and age. We hypothesized that the ISS would have large positive correlations with the RMDQ and pain items, a large negative correlation with satisfaction with social role participation, and a small positive correlation with age. Magnitude of correlations was defined by 0.10, 0.30 and 0.50 as small, medium and large, respectively [13].

Level of Pain Impact and Responsiveness to Change. For the three data collection points, we report the degree of pain impact using the ISS cut points suggested by the NIH task force: 8–27 (mild), 28–34 (moderate), and 35–50 (severe) in the overall sample and for the pain duration subgroups. We estimate the effect size for change and responsiveness to change of the ISS from baseline to 6 weeks later for the overall sample using multiple anchors: 1) Retrospective rating of change in low back pain; 2) How bothersome was your low back pain; and 3) Average low back pain. We estimate Spearman correlations between change in the ISS and the anchors and assess whether the amount of change on the ISS is monotonically related to change implied by the levels of change suggested by the anchors. We also report F-statistics from general linear models with the ISS as the dependent variable and the anchors as independent variables.

We estimate the area under the curves with the ISS predicting improvement on the retrospective rating of change in low back pain item coded as *moderately better, much better* or *completely gone* (all other categories coded as not improved). We also evaluate different cut points for change in the ISS associated with improvement in the retrospective rating of change item. We identify cut points using the Youden [14] index: (sensitivity + specificity)-1.

All analyses were conducted using SAS. Area under the curve analyses were performed using SAS PROC Logistic, SAS PROC Gplot, and the SAS rocplot macro.

Results

Internal consistency reliability of the ISS was 0.92, 0.93, and 0.93 at baseline, 6 weeks post-baseline and 12 weeks post-baseline, respectively, for the overall sample. Internal consistency reliability estimates within pain duration subgroups (acute, subacute, chronic) ranged from 0.90 to 0.93 at baseline, 0.92 to 0.95 at 6 weeks post-baseline, and 0.92 to 0.94 at 12 weeks post-baseline. Test-retest reliability of the ISS from baseline to 6 weeks later was 0.77 (0.82, 0.86, and 0.74 in the acute, subacute, and chronic subgroups, respectively).

The ISS correlation with the RMDQ ranged from 0.75 (baseline) to 0.84 (6 weeks later). The increase over time in the size of correlations is related to an increase in variance. The standard deviation of the ISS was 8.4, 8.8 and 9.1 at baseline, 6 weeks post-baseline, and 12 weeks post-baseline, respectively. Similarly, the standard

deviation of the RMDQ was 5.5, 6.1, and 6.0 at these three timepoints. Correlations were similar for those with chronic pain and for the overall sample (Table 1). Correlations were also noteworthy with being bothered by back pain in the past week, worst back pain in the past 24 hours, average back pain in the last week, and satisfaction with participation in social roles. Correlations for the overall sample versus those with chronic back pain (shown at baseline) with these variables were similar (Table 1).

Table 2 provides ISS severity categories based on the NIH research task force suggested categories by duration of low back pain (acute, subacute, chronic) and for the overall sample. Sixty-five percent of the overall sample had mild pain impact, 21% had moderate impact and 14% severe impact. Mild impact was more common among those with chronic low back pain and severe impact were more common among those with acute low back pain. The percent of study participants in the mild, moderate, and severe ISS categories, respectively, were 82%, 12%, and 6% at 6 weeks, and 84%, 9%, and 6% at 12 weeks. As shown in Appendix Table 1, for each of the three low back pain duration subgroups the majority of those with mild or moderate impact at baseline were mild post-baseline (6 weeks and 12 weeks). For those with severe impact at baseline the majority of those with acute back pain were mild at follow-up, while the majority of those who were subacute or chronic were moderate or severe at follow-up.

Effect size for change in the ISS from baseline to 6-weeks later was small (0.21). Spearman correlations between change on the ISS and the anchors were 0.63 (retrospective ratings of change in low back pain), 0.68 (bothered by low back pain), and 0.57 (average low back pain), exceeding the 0.37 threshold for an acceptable anchor [15]. The second rows in Tables 3–5 provide the mean ISS change (6 weeks-baseline) for each anchor level. Means that do not differ significantly (Duncan's multiple range test) share a superscript. Change in ISS scores were monotonically associated with the retrospective rating of change item (Table 3). ISS change did not differ significantly between those who reported on the retrospective change item 6 weeks post-baseline that they were *about the same* or *a little better* compared to baseline. The overall F-statistic for the change in ISS was F (6, 626) = 75.36, P < .0001.

Changes in ISS scores were monotonically associated with the change in bothered by low back pain item except for one person that changed from *not at all* to *extremely* bothered (Table 4). ISS change did not differ significantly between those whose change on the bother item was *about the same* or *a little better* or declined by one, two, or three categories. Those who stayed the same did not differ significantly from those who declined one or two categories or improved by one category on the bother item. The overall F-statistic for the change in ISS was F (8, 625) = 91.10, P < .0001.

Table 1. Product-moment correlations of impact stratification score with other variables

	Baseline (chronic back pain only)	6 weeks later	12 weeks later
RMDQ	0.75 (0.76)	0.84 (0.82)	0.83 (0.83)
BTHLBP	0.65 (0.66)	0.78 (0.76)	0.78 (0.79)
PRSAvg	0.51 (0.56)	0.74 (0.71)	0.75 (0.76)
PRSWorst	0.56 (0.56)	0.77 (0.75)	0.77 (0.76)
SRTScore	−0.64 (−0.66)	−0.68 (−0.69)	−0.71 (−0.72)
Age	0.02 (0.09)	0.15 (0.09)	0.16 (0.11)

RMDQ = Roland-Morris Disability Questionnaire; BHTLBP = bothered by low back pain in the past week; PRSAvg = average low back pain in the past week; PRSWorst = worst low back pain in the past 24 hours; SRTScore = PROMIS-29 v1.0 satisfaction with social role participation scale. Age = Age in years at baseline.

Table 2. Impact stratification score severity at baseline by duration of low back pain

Severity	Acute (n = 286)	Subacute (n = 79)	Chronic (n = 384)	Overall (n = 749)
Mild	56%	61%	72%	65%
Moderate	23%	29%	18%	21%
Severe	21%	10%	10%	14%
Column Percent	38%	11%	51%	100%

Table 3. Mean change in ISS by retrospective rating of low back pain

Much worse	A little worse	About the same	A little better	Moderately better	Much better	Completely gone
8[a]	3[b]	−2[c]	−3[c]	−6[d]	−12[e]	−17[f]
n = 16	n = 64	n = 178	n = 104	n = 89	n = 146	n = 36

Compared to your first visit, your low back pain is: *much worse, a little worse, about the same, a little better, moderately better, much better or completely gone.*

Note: F (6, 626) = 75.36, P < .0001. Cells that share a superscript do not differ significantly from one another.

Changes in ISS scores were generally monotonically associated with the change in the average low back pain item except for two cases in the sample that declined three categories (Table 5). ISS change was similar for those who stayed the same or declined on the average low back pain item. The F-statistic for the change in ISS was F (6, 627) = 58.39, P < .0001.

The area under the curve estimate for the improvement in the retrospective rating of change item predicted by change in ISS from baseline to 6-weeks later was 0.83 (Figure 1). The optimal cut-point was 7 according to the Youden index (Figure 2): sensitivity at this cut point = 66%, specificity = 85%, negative predictive value = 77%, and positive predictive value = 76%.

Discussion

This study supports the psychometric properties of the ISS. Internal consistency reliability and test-retest reliability estimates in this study were similar to those of Deyo et al. [5]. We also found similar magnitude of associations of the ISS with the RMDQ to those previously reported by Deyo et al. [1]. In addition, we found new information about the construct validity of the ISS based on its significant associations with being bothered by back pain, worst back pain and less satisfaction with participation in social roles and activities. Although the ISS was proposed as a measure for chronic low back pain, we found that these associations were similar for the overall sample (acute, subacute, and chronic low back pain) and when looking only at those with chronic low back pain.

Change in ISS scores in this study were largely monotonically associated with multiple independent anchors of change (retrospective rating of change, change in bother by low back pain, change in average low back pain). Area under the curve analysis demonstrated that the ISS was able to capture change defined by those who reported at the 6-week assessment that they were *moderately better, much better,* or their low back pain was *completely gone*. The optimal cut-point on the ISS for identifying this definition of change was 7 points—a 7-point decrease on the ISS represents those who feel their low back pain has improved substantially and might be considered responders to treatment. This estimate is similar to the change of 7.5 points on the ISS among those who reported that they were much improved or completely improved at 12 months follow-up in a study of 223 patients at a Dutch spine clinic with low back and/or leg pain [16].

Study limitations are worth noting. As is true with all studies of low back pain of musculoskeletal origin, the specific diagnosis was difficult to determine or confirm. In addition, caution is warranted in generalizing from a sample of active-duty members of the US military. Furthermore, a challenge of conducting research in the military is following patients who are transient, especially in times of war. When this study was designed, the active-duty population of interest was likely to be

Table 4. Mean change in ISS by low back pain bother in the past week

From not at all to Extremely bothersome	Declined three categories	Declined two categories	Declined one category	Stayed the same	Improved one category	Improved two categories	Improved three categories	From Extremely to Not at all bothersome
−9[e, d]	10[a]	8[a, b]	2[a, b, c]	0[b, c, d]	−5[c, d, e]	−11[e, f]	−19[f]	−29[g]
n = 1	n = 1	n = 8	n = 62	n = 217	n = 176	n = 114	n = 45	n = 10

How bothersome was your low back pain in the past week? *Not at all bothersome, Slightly bothersome, Moderately bothersome, Very bothersome, Extremely bothersome* [6 weeks-baseline].

Note: $F_{(8, 625)} = 91.10$, $P < .0001$. Cells that share a superscript do not differ significantly from one another.

Table 5. Mean change in ISS by average low back pain during the past week.

Declined 3 categories	Declined 2 categories	Declined 1 category	Stayed the same	Improved 1 category	Improved 2 categories	Improved 3 categories
−6[b, c]	6[a]	1[a, b]	−2[b, c]	−8[c]	−16[d]	−24[e]
n = 2	n = 10	n = 85	n = 273	n = 203	n = 51	n = 10

Select the number that best describes your average low back pain during the past week. 0 = No pain; 10 = Worst possible pain (recoded so that 10 = 5, 7–9 = 4, 4–6 = 3, 1–3 = 2, and 0 = 1; based on Sheehan Disability Scale and the Flushing Questionnaire) (6 weeks-baseline).

Note: $F_{(6, 627)} = 58.39$, $P < .0001$. Cells that share a superscript do not differ significantly from one another.
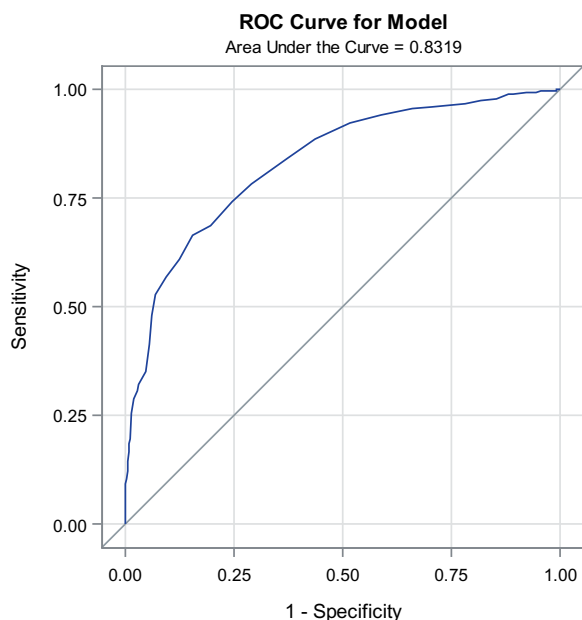

Figure 1. Area under the curve predicting improvement on retrospective change by change in ISS (baseline to 6 weeks later).

Figure 2. Optimal cut points for change in ISS (baseline to 6 weeks later).

deployed. Participants were excluded if they were scheduled to leave the country within the 12-week study period. Thus, the relatively short follow-up is a limitation of the study. Finally, the ISS severity categories we used were suggested by the NIH Pain Consortium research task force but collapsing a continuous measure into three categories discards information. Moreover, comparison of the NIH taskforce cut points with other alternatives is needed in future studies.

Despite the limitations, the study contributes to the sparse literature on the ISS and provides additional support for its re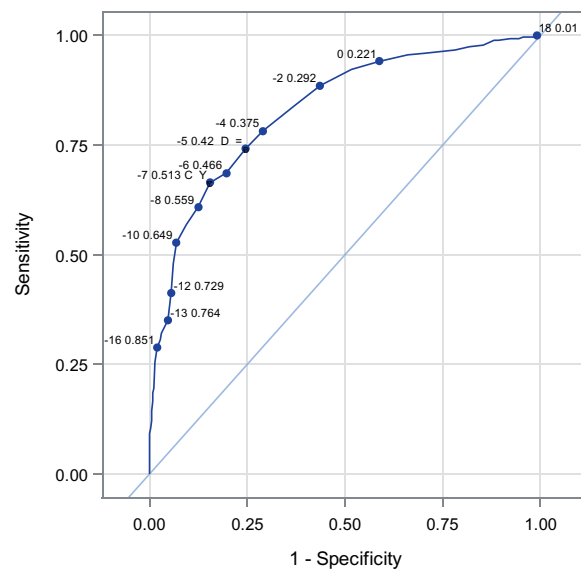liability and validity. Administering the entire PROMIS-29 v2.1 (3) rather than just the 9 ISS items has the advantage of more comprehensive assessment of health-related quality of life by virtue of 20 more items: 4-item fatigue, sleep disturbance, depression, anxiety, and ability to participate in social roles and activities scales. But the ISS represents the core physical health outcomes associated with low back pain with a 69% reduction in the number of items. The ISS is a parsimonious measure for use in research and potentially in clinical practice. Further evaluation of the psychometric properties of the ISS in other datasets and the tentative severity classification levels proposed by the NIH Pain Consortium research task force is needed.

Appendix

Table 1. Impact stratification score severity at baseline compared to post-baseline

	6 Weeks Post-baseline			12 Weeks Post-baseline		
	Mild	Moderate	Severe	Mild	Moderate	Severe
Baseline Acute						
Mild	96%	2%	2%	97%	3%	0%
Moderate	82%	12%	6%	91%	5%	5%
Severe	68%	21%	11%	81%	7%	12%
Baseline Subacute						
Mild	87%	10%	9%	88%	12%	0%
Moderate	61%	33%	6%	93%	7%	0%
Severe	25%	38%	38%	43%	57%	0%
Baseline Chronic						
Mild	91%	7%	19%	90%	6%	4%
Moderate	61%	27%	12%	52%	31%	17%
Severe	34%	26%	40%	42%	19%	39%
Baseline Overall						
Mild	92%	6%	1%	92%	6%	2%
Moderate	70%	22%	9%	73%	17%	10%
Severe	51%	24%	24%	63%	16%	21%

Note: Row percent is shown within the post-baseline follow-up assessments.

References

1. Deyo RA, Dworkin SF, Amtmann D, et al. Focus article: Report of the NIH Task Force on Research Standards for Chronic Low Back Pain. Eur Spine J 2014;23(10):2028–45.
2. Deyo RA, Dworkin SF, Amtmann D, et al. Report of the NIH Task Force on research standards for chronic low back pain. Pain Med 2014;15(8):1249–67.
3. Cella D, Choi SW, Condon DM, et al. PROMIS® adult health profiles: Efficient short-form measures of seven health domains. Value in Health 2019;22(5):537–44.
4. Hays RD, Spritzer KL, Schalet BD, Cella D. PROMS®-29 v2.0 profile physical and mental health summary scores. Qual Life Res 2018;27(7):1885–91.
5. Deyo RA, Ramsey K, Buckley DI, et al. Performance of a Patient Reported Outcomes Measurement Information System (PROMIS) short form in older adults with chronic musculoskeletal pain. Pain Med 2016;17:314–24.
6. Goertz CM, Long CR, Vining RD, et al. Assessment of chiropractic treatment for active duty, U.S. military personnel with low back pain: study protocol for a randomized controlled trial. Trials 2016;17:70.
7. Goertz CM, Long CR, Vining RD, Pohlman KA, Walter J, Coulter I. Effect of usual medical care plus chiropractic care vs usual medical care alone on pain and disability among US service members with low back pain: A comparative effectiveness clinical trial. JAMA Netw Open 2018;1(1):e180105.
8. Roland M, Morris R. A study of the natural history of back pain. Part I: Development of a reliable and sensitive measure of disability in low-back pain. Spine 1983;8(2):141–4.
9. Wilson R, Bocell F, Bamer AM, Salem R, Amtmann D. Satisfaction with social role participation in adults living with chronic conditions: Comparison to a US general population sample. Cogent Psychol 2019;6(1):1588696.
10. Hahn EA, Beaumont JL, Pilkonis PA, et al. The PROMIS satisfaction with social participation measures demonstrated responsiveness in diverse clinical populations. J Clin Epidemiol 2016;73:135–41.
11. Patrick DL, Deyo RA, Atlas SJ, Singer DE, Chapin A, Keller RB. Assessing health-related quality of life in patients with sciatica. Spine 1995;20(17):1899–995.
12. Childs JD, Piva SR, Fritz JM. Responsiveness of the numeric pain rating scale in patients with low back pain. Spine 2005;30(11):1331–4.
13. Cohen J. Statistical power analysis for the behavioral sciences, 2nd edition. Routledge; 1988.
14. Youden WJ. Index for rating diagnostic tests. Cancer 1950;3(1):32–5.
15. Hays RD, Farivar SS, Liu H. Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. COPD 2005;2(1):63–7.
16. Dutmer AL, Reneman MF, Preuper HRS, Wolff AP, Speijer BL, Soer R. The NIH minimal dataset for chronic low back pain. Spine 2019;44(20):E1211–8.