Journal of Clinical Neuroscience 143 (2026) 111717

Contents lists available at ScienceDirect

Journal of Clinical Neuroscience

journal homepage: www.journals.elsevier.com/journal-of-clinical-neuroscience

Clinimetric analysis of the numeric pain rating scale, Oswestry disability index, and the Roland-Morris disability questionnaire in patients with lumbar spinal stenosis treated with conservative interventions

Ian Young [a,b,*], James Dunning [b,c], James Escaloni [b], Filippo Maselli [d,e], Joshua Prall [b,f], Firas Mourad [g], Nathan Hutting [h], César Fernández-de-las-Peñas [i]

[a] *Tybee Wellness & Osteopractic, Tybee Island, GA, USA*
[b] *American Academy of Manipulative Therapy Fellowship in Orthopaedic Manual Physical Therapy, Montgomery, AL, USA*
[c] *Montgomery Osteopractic Physical Therapy & Acupuncture, Montgomery, AL, USA*
[d] *Wellward Regenerative Medicine, Lexington, KY, USA*
[e] *Department of Human Neurosciences, Sapienza University of Rome, Rome, Italy*
[f] *Lebanon Valley College, Department of Physical Therapy, Annville, PA, USA*
[g] *Luxembourg Health & Sport Sciences Research Institute, Differdange, Luxembourg*
[h] *Department of Occupation and Health, School of Organisation and Development, HAN University of Applied Sciences, Nijmegen, the Netherlands*
[i] *Department of Physical Therapy, Occupational Therapy, Rehabilitation and Physical Medicine, Universidad Rey Juan Carlos, Alcorcón, Madrid, Spain*

ARTICLE INFO

Keywords:
Low back pain
Spinal stenosis
Reliability
Validity
MCID
MDC
Responsiveness
Pain
Disability

ABSTRACT

Background context: The numeric pain rating scale (NPRS), Oswestry disability index (ODI) and the Roland-Morris disability questionnaire (RMDQ) are commonly used patient-reported outcome measures (PROMs) in patients with low back pain. However, there is a paucity of evidence supporting the reliability, validity, and responsiveness of these PROMs in patients with lumbar spinal stenosis (LSS) treated with conservative interventions including spinal manipulation, electrical dry needling, joint mobilization, and exercise.
Purpose: To establish the reliability, construct validity, responsiveness, minimal detectable change (MDC), and minimum clinically important difference (MCID) of the NPRS, ODI, and RMDQ in patients with LSS.
Study design/setting: Clinimetric analysis of a prior randomized clinical trial.
Patient sample: One hundred twenty-eight (n = 128) patients with LSS treated with outpatient physical therapy, including manual therapy, exercise and electrical dry needling.
Outcome measures: PROMs included the NPRS, ODI, RMDQ, and the global rating of change scale (GROC).
Methods: The MDC (90 % and 95 % confidence interval) and MCID was established for "improved" (GROC: +3 to + 7) and "much-improved" (GROC: +5 to + 7) patients at 3-months follow-up. The intraclass correlation coefficient ($ICC_{2,1}$) was analyzed at 2 weeks, 6 weeks and 3-months. Pearson's correlation (r) was also calculated.
Results: The NPRS ($ICC_{2,1} = 0.55$; [95 % CI: 0.19–0.79]) exhibited moderate reliability, while the ODI ($ICC_{2,1} = 0.86$; [95 % CI: 0.70–0.94]) and RMDQ ($ICC_{2,1} = 85$; [95 % CI: 0.64–0.94]) exhibited good reliability at the 3-month follow-up. All PROMs exhibited acceptable responsiveness (area under the curve range 0.73 to 0.92; $p < 0.001$) and strong construct validity (Pearsons *r*: range 0.51 to 0.72; $p < 0.001$). The MDC_{95} was 2.3 points for the NPRS, 8.5 points for the ODI, and 6.1 points for the RMDQ. The MCID exceeded measurement error in the "much improved" patients for the NPRS (2.8 points) and the ODI (8.5 points), but did not for the RMDQ (4.5 points).
Conclusions: All three PROMs demonstrated suitable clinimetric properties in patients with LSS. Based on this analysis, patients who are "much improved" should have at least a 2.8-point reduction on the NPRS, and an 8.5-point reduction on the ODI to be considered clinically meaningful, and to exceed the measurement error. The MCID of the RMDQ did not exceed measurement error in either improvement category.

* Corresponding author at: Tybee Wellness & Osteopractic, 8 Horsepen Point Drive, Tybee Island, GA 31328, USA.
E-mail addresses: tybeewellness@gmail.com (I. Young), drjamesdunning@gmail.com (J. Dunning), james.escaloni@spinalmanipulation.org (J. Escaloni), masellifilippo76@gmail.com (F. Maselli), prall@lvc.edu (J. Prall), firasmourad@me.com (F. Mourad), Nathan.Hutting@han.nl (N. Hutting), cesar.fernandez@urjc.es (C. Fernández-de-las-Peñas).

https://doi.org/10.1016/j.jocn.2025.111717
Received 11 August 2025; Accepted 25 October 2025

1. Introduction

The annual per population direct cost of LBP continues be a financial burden in both high income (~3.6 billion USD) and lower income countries (~$2.2 billion USD) [1,2]. The numeric pain rating scale (NPRS), Oswestry disability index (ODI) and the Roland-Morris disability questionnaire (RMDQ) are well established and commonly used patient-reported outcome measures (PROMs) in patients with low back pain (LBP) [3–6]. However, there are gaps in the evidence supporting the reliability, validity, and responsiveness of these PROMs in patients with specific sub-categories of LBP, such as lumbar spinal stenosis (LSS). To date, there is only one high-quality clinimetric analysis on patients with LSS after treatment with outpatient physical therapy/conservative intervention. In 2012, Cleland et al. [7] reported the reliability (intraclass correlation coefficient) of the NPRS (0.49 to 0.52), ODI (0.84) and the minimum clinically important difference (MCID) of the NPRS (3.4 to 3.7) and ODI (5.5) in a small sample (n = 55) of patients with LSS. Since then, there have been no further analyses in this patient population following conservative interventions. Furthermore, comprehensive clinimetric data of the RMDQ has not yet been established in patients with LSS. Comparatively, the MCID of these PROMs have been examined in patients with lumbar disc herniation after 4 weeks of physical therapy intervention (n = 92; MCID: ODI = 6.5 points; RMDQ = 5.5 points) [8], and at 7 days after a single conservative traditional Chinese medicine intervention (n = 329; MCID: NPRS = 1.7, ODI = 5.9, RMDQ = 1.7) [9]. With a relatively high prevalence (11 % to 38 %) of clinically diagnosed LSS [10], acquiring recent clinimetric data on larger samples, updated treatment paradigms and longer follow-up time frames seems to be of paramount importance. Therefore, the purpose of this study was to examine the reliability, validity, responsiveness of the NPRS, ODI, and RMDQ in patients with LSS (n = 128) that were treated with conservative/non-surgical interventions.

2. Methods

This study is a secondary analysis of a large multi-center randomized clinical trial that investigated the effects of two different standardized physical therapy interventions in patients with LSS [11]. Consecutive individuals with LSS (n = 128) were randomized to receive spinal manipulation, electrical dry needling and conventional physical therapy (joint mobilization and exercise) or conventional physical therapy alone [11]. The inclusion and exclusion criteria are described in detail in the previous trial. [11] For patients to be eligible, they had to be ≥ 50 years-old and meet the following criteria: (1) Symptoms of neurogenic claudication (pain in the buttock, thigh, or leg during ambulation that improves with rest) or radicular leg symptoms with associated neurological deficits on physical examination for at least 12 weeks and (2) confirmatory imaging (ie, magnetic resonance imaging, computed tomography, myelography, ultrasound, or radiographic images of either central [ie, spinal canal] or lateral [ie, foraminal] lumbar stenosis at one or more levels in the lumbar spine). The NPRS, ODI, and RMDQ were collected in all patients at baseline, 2 weeks, 6 weeks, and 3 months. Perceived recovery using the Global Rating of Change Scale (GROC) [12] was collected at all follow-up points. To investigate the clinimetric properties of the outcome measures, both groups completing the trial were collapsed into a single cohort using the 3-month follow-up time frame.

2.1. Outcome measures

The NPRS is an 11-point scale ranging from 0 (“no pain”) to 10 (“worst pain imaginable”) [13]. It has been shown to be a valid and reliable instrument to assess pain intensity [14,15] and adequately allows specific responses [16] in spine pain patients. [17–21] In the current study, overall low back, buttock and leg pain intensity was assessed by the NPRS. Given the MCID has been shown to be 1.74 in patients with different chronic pain conditions, [14] a 2-point change or 30 % decrease in pain compared to baseline may be considered the MCID for chronic musculoskeletal pain [14,22].

The ODI is a 10-question standardized assessment test that is widely used to measure disability related to LBP [23]. In the current study, the raw score from 0 to 50 was used for data analysis. The test has been found to be both reliable and valid for patients with chronic LBP, [24,25] and it has an excellent test–retest reliability [26] with a 0.89 intraclass correlation coefficient for patients with LSS [27]. The MCID on the ODI ranges from 4 to 10.5 [23,26,28–30].

The RMDQ is a 24-item self-report questionnaire of how LBP affects functional activities [31]. While the ODI is considered to have better re-test reliability and less error than the RMDQ, the RMDI is thought to have superior construct validity for measuring functional impairment [32]. Like the ODI, the RMDQ has been found to be both valid and reliable for measuring function in patients with chronic LBP [33]. In general, changes of at least 2–3 points on the RMDQ should be considered the MCID [31]. However, other investigators suggest that a 30 % reduction in baseline score more accurately represents the MCID on the RMDQ [34,35].

The GROC described by Jaeschke et al. [12] is a 15-point questionnaire assessing the patients perceived recovery. This scale ranges from −7 (a very great deal worse) to zero (about the same) to + 7 (a very great deal better). Intermittent descriptors of worsening or improving are assigned values from −1 to −6 and + 1 to + 6, respectively. The MCID for the GROC has not been specifically reported, however, scores of + 4 and + 5 have typically been indicative of moderate changes in patient status [36]. The GROC has been used in previous studies that investigated the conservative treatment of LSS [7,37,38]. For LSS, ratings >+3 and < -3 are indicative of an improvement and worsening of the patient’s condition, respectively [7]. The GROC was collected at 2 weeks, 6 weeks and 3 months following the initial treatment session.

2.2. Data analysis

Data analysis was performed using SPSS v30.0. We categorized patients into 3 mutually exclusive groups at the 3-month follow-up based on their GROC scores. Those scoring between −2 to + 2 were considered clinically “stable’’ (minimal to no change); those scoring + 3 to + 7 were considered clinically “improved” (at least somewhat better), and those scoring + 5 to + 7 were considered clinically “much improved” (at least a good deal better). Similar categorization has been used in previous clinimetric studies [15,39,40].

Reliability: Test-retest reliability was examined for the NPRS, ODI and RMDQ using “stable” patients. Intraclass correlation coefficient ($ICC_{2,1}$) for the NPRS, ODI and RMDQ were calculated for the group of patients who were classified as being “stable” (GROC: −2 to + 2, n = 22) by comparing scores at the initial examination with those at the 3-month follow-up. The $ICC_{2,1}$ was calculated according to procedures described by Shrout and Fleiss [41,42]. Values < 0.50 indicate poor reliability, while values between 0.50 and 0.75, between 0.75 and 0.90, and > 0.90 denote moderate, good, and excellent agreement, respectively [42]. Reliability ($ICC_{2,1}$) was also examined at the 2-week and 6-week follow-up, for comparative analysis.

Construct Validity: Construct validity of the NPRS, ODI, and RMDQ were examined by comparing the change in outcome scores for the “stable’’, “improved”, and “much improved” groups using separate, two-way analyses of variance (ANOVA) for repeated measures at baseline and reevaluation. We hypothesized that “stable’’ patients in each group would have NPRS, ODI, and RMDQ intake scores that had very little change, whereas patients classified in the improved categories would demonstrate a significant change in values. This would be represented by a significant group x time interaction. Pearson’s correlation (r) was also calculated to examine the linear association between the outcome measures and the criterion measure used for perceived improvement (GROC). Values between 0.30 and 0.49 indicate a moderate correlation, while values between 0.50 and 1.0 indicate a strong correlation [43]. Further, secondary to a small sample, bootstrapping (1000 samples; 95 %CI) was also performed on the “stable” group (n = 22) at 3-months to examine the significance of correlations found between all outcomes and the GROC at 3 months.

Responsiveness: The responsiveness and interpretability of the NPRS, ODI and RMDQ were assessed using the clinically “stable”, and “improved” groups at the 3-month follow-up point. Receiver operator characteristic (ROC) curves [44,45] were constructed by plotting sensitivity values (true-positive rate) on the y axis and 1-specificity values (false-positive rate) on the x axis for each level of change score. Separate ROC curves were constructed for improved and much-improved groups for all three PROMs. The area under the curve (AUC) and the 95 % CI were obtained as a method for determining the ability of each measure to distinguish improved patients from stable patients in each category. An AUC of 0.50 indicates that the measure has no diagnostic accuracy beyond chance, whereas a value of 1 suggests perfect accuracy [44,45].

MCID: MCID, the smallest difference that patients perceive as beneficial, was calculated by identifying the point on the ROC curve nearest to the upper left-hand corner (maximizing sensitivity and specificity values), which is considered to be the best cutoff score for distinguishing “improved” and “much improved” patients [44–46]. Sensitivity and specificity values for the selected cutoff scores were also calculated.

MDC: The MDC, the amount of change that must be observed before the change can be considered to have exceeded measurement error, was calculated by determining the standard error of measurement (SEM) for the NPRS, ODI and RMDQ in the “stable” group (n = 13) [46]. The SEM was estimated using the formula SD*$\sqrt{\,}$(1- reliability coefficient) where SD is the pooled standard deviation in the “stable” group. The SEM was multiplied by 1.96 to determine the 95 % CI (MDC$_{95}$) [47]. This value was multiplied by the $\sqrt{2}$ to account for the errors taken with repeated measurements [47].

3. Results

One hundred twenty-eight patients met the inclusion and exclusion criteria, completed the study, and were included in data analysis. Baseline characteristics of the participants are reported in Table 1. The mean GROC score for all patients included in the analysis at the 3-month follow-up was + 4.4 (SD + 2.4). The mean GROC score for the “stable”, “improved” and “much-improved” groups was + 0.5 (SD 1.3), +5.3 (SD 1.7), and + 6.1 (SD + 0.86), respectively. At the 3-month follow-up, 105 (82 %) patients were classified as “improved”, 74 (58 %) were “much-improved”, and 22 (17 %) remained “stable”. There was a significant difference (p < 0.001) in mean change scores between “stable” vs. “improved” and “stable” vs. “much-improved” groups in all three PROMs at the 3-month follow-up (Table 2).

3.1. Reliability and MDC

The ICC$_{2,1}$ and MDC$_{90/95}$ values are reported in Table 3. At the 3-month follow-up, the NPRS (ICC 0.55, 95 %CI 0.19–0.79), had moderate reliability, while the ODI (ICC 0.86, 95 %CI 0.70–0.94), and RMDQ (ICC 0.85, 95 %CI 0.64–0.94), exhibited good reliability. The reliability at each follow-up point, for each PROM is noted in Table 3. The MDC$_{90/95}$ was 2.0/2.3 points for the NPRS, 7.2/8.5 points for the ODI, and 5.1/6.1 points for the RMDQ (Table 3).

Table 1
Baseline characteristics.

	N = 128 Mean (SD)
Gender: Male/Female	57 / 71
Age: yrs.	66.7 (±10.2)
Weight: kg	84.6 (±17.5)
Years with Symptoms	4.2 (±4.9)
Numeric Pain Rating Scale	5.9 (±1.9)
Oswestry –Total (0–50)	20.9 (±7.5)
Roland Morris Disability Questionnaire (0–24)	10.5 (±4.9)
Average Number of Treatment Visits	11.2 (±1.5)
Medication Intake: 1-2x per day (%)	53 (41.4)

Table 2
Difference between change scores from baseline to 3-month follow-up on self-report outcomes.

	Improved GROC (+3 to + 7) N = 105 Mean (SD)	Stable GROC (−2 to + 2) N = 22 Mean (SD)	Mean Difference (95 % CI)	*P*
NPRS	3.53 (2.2)	0.45 (1.8)	3.08 (2.1; 4.1)	*P < 0.0001*
OSW	11.53 (6.7)	0.86 (4.2)	10.67 (7.7; 13.6)	*P < 0.0001*
RMDQ	6.49 (4.6)	1.27 (2.6)	5.22 (3.2; 7.2)	*P < 0.0001*

NPRS = numeric pain rating scale (0–10), ODI = oswestry disability index (total score 0–50), RMDQ = roland morris disability questionnaire (0–24), GROC = global rating of change scale (−7 to + 7), AUC = area under the curve, CI = confidence interval.

3.2. Responsiveness, MCID & construct validity

The NPRS, ODI and RMDQ demonstrated very good responsiveness in the “improved” group (AUC range 0.85 to 0.92; Table 3) and at least good responsiveness in the “much improved” group (AUC range 0.73 to 0.83; Table 3). The MCID and the sensitivity/specificity associated with each cutoff score are also reported in Table 3. For the NPRS, the MCID was 1.3 points for the “improved” group and 2.8 points for the “much-improved” group. For the ODI, the MCID was a 5.5- and 8.5-point change for the “improved” and “much-improved” groups, respectively. For the RMDQ, the MCID was 3.5 points for the “improved” group and 4.5 points for the “much-improved” group. All PROMs also exhibited acceptable construct validity at 3 months (Pearsons *r*: range 0.51 to 0.72; *p* < 0.001; Table 4). Notably, the RMDQ had weaker correlations with the NPRS (Pearsons *r* = 0.51; 95 %CI 0.37 to 0.63) and GROC (Pearsons *r* = 0.51; 95 %CI 0.37 to 0.63). Finally, in support of using the small “stable” group for multiple analyses, bootstrapping supported significant correlations (Pearsons r: p < 0.005) between the NPRS/GROC (95 %CI 0.14 to 0.82), ODI/GROC (95 %CI 0.01 to 0.81), RMDQ/GROC (95 %CI 0.34 to 0.78) at 3 months.

4. Discussion

This study examined the clinimetric properties of three commonly used PROMs used in patients with LBP in a large sample of patients with LSS treated with conservative interventions. All outcome measures exhibited moderate to good reliability, appropriate construct validity, and acceptable responsiveness (Table 2-4). Overall, our analysis suggests that these commonly used PROMs of pain and disability have suitable clinimetric properties for use in patients with LSS (Table 2-4, Figs. 1-3).

4.1. Reliability

The NPRS (ICC$_{2,1}$ 0.55) demonstrated moderate reliability at the 3-month follow-up, while the ODI and RMDQ exhibited good reliability (ICC$_{2,1}$: ODI 0.86, RMDQ 0.85) at the 3-month follow-up (Table 3). In comparison, Cleland et al. [7] reported moderate reliability of the NPRS for back/buttock region (ICC$_{2,1}$ 0.52), poor reliability for thigh/leg region (ICC$_{2,1}$ 0.48), and good reliability for the ODI (ICC$_{2,1}$ 0.84) in a small sample of LSS patients at the 6-week follow-up. To date, the

Table 3
Clinimetric properties of patient reported outcomes for lumbar spinal stenosis: 3-month follow-up.

	AUC 95 % CI Improved N = 105	MCID Sn; Sp Improved N = 105	AUC 95 % CI Much Improved N = 74	MCID Sn; Sp Much Improved N = 74	$ICC_{2,1}$ 95 % CI Stable N = 22	SEM	MDC_{90}	MDC_{95}
NPRS	0.85 0.77; 0.93	1.3 0.84; 0.70	0.83 0.76; 0.91	2.8 0.80; 0.77	0.55 0.19, 0.79	0.84	2.0	2.3
ODI	0.92 0.87; 0.97	5.5 0.81; 0.91	0.83 0.76; 0.90	8.5 0.74; 0.77	0.86 0.70, 0.94	3.1	7.2	8.5
RMDQ	0.85 0.77; 0.94	3.5 0.78; 0.76	0.73 0.64; 0.82	4.5 0.68; 0.66	0.85 0.64, 0.94	2.2	5.1	6.1

NPRS = numeric pain rating scale (0–10), ODI = Oswestry disability index (total score 0–50), RMDQ = Roland Morris disability questionnaire (0–24), improved = global rating of change scale (+3 to + 7), much improved = global rating of change scale (+5 to + 7), stable = global rating of change scale (−2 to + 2), AUC = area under the curve, CI = confidence interval, MCID = minimum clinically important difference, Sn = sensitivity, Sp = specificity, $ICC_{2,1}$ = intraclass correlation coefficient, SEM = standard error of measure, MDC = minimal detectable change (90 % and 95 % confidence interval).

Table 4
Pearson's Correlation Coefficient (r): 3 months follow-up.

Outcome Measures	ODI r (95 % CI)	RMDQ r (95 % CI)	GROC
NPRS	0.64 (0.52; 0.73) *P<0.001*	0.51 (0.37; 0.63) *P<0.001*	0.62 (0.50; 0.72) *P<0.001*
ODI	———	0.72 (0.63; 0.78) *P<0.001*	0.67 (0.56; 0.76) *P<0.001*
RMDQ	———	———	0.51 (0.37; 0.63) *P<0.001*

NPRS=numeric pain rating scale (0-10), ODI=Oswestry disability index (total score 0-50), RMDQ= Roland-Morris disability questionnaire (0-24), GROC= global rating of change scale (-7 to +7).

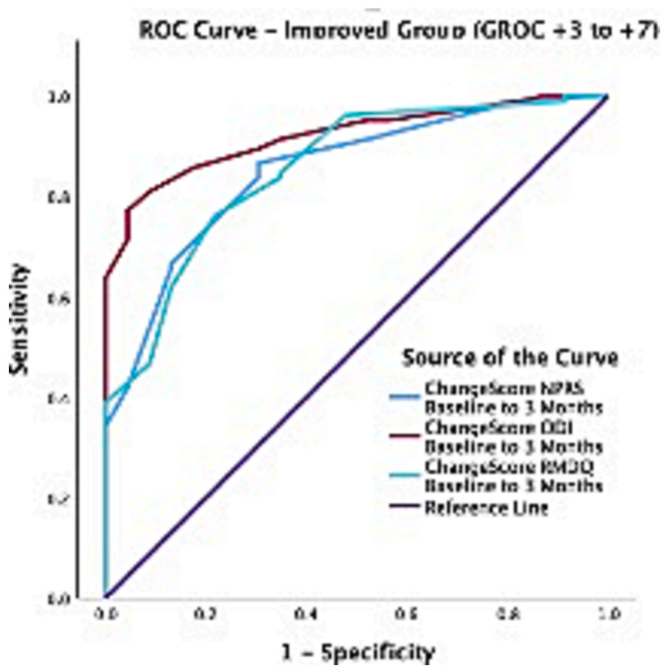

Fig. 1. Receiver Operating Curve: Improved Patients.

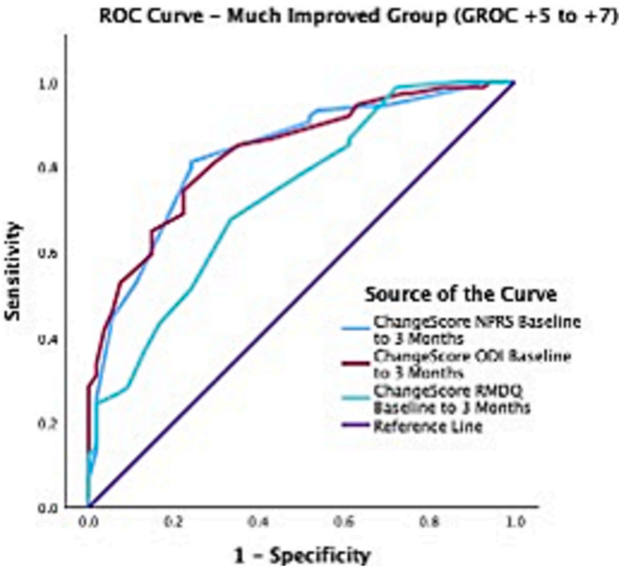

Fig. 2. Receiver Operating Curve: Much-Improved Patients.

reliability of the RMDQ has not been established for patients with LSS treated with outpatient physical therapy. Interestingly, similar findings have been reported in patients with lumbar disc herniation following conservative intervention, showing good reliability for the ODI ($ICC_{2,1}$ = 0.87) and RMDQ ($ICC_{2,1}$ = 0.86) at 7 days follow-up, with the exception of excellent reliability of the NPRS ($ICC_{2,1}$ 0.99) [9]. Our analysis across multiple follow-up time points may aid in the explanation of the discrepancy in the NPRS reliability reported in the aforementioned studies (Fig. 3). In the current sample, our results suggest that over longer periods of follow-up, the NPRS has progressive limitations in reliability ($ICC_{2,1}$: 2 weeks = 0.77, 6 weeks = 0.66, 3 months = 0.55). This compromised reliability of pain rating scales at the 3-month follow-up seems to be consistent with clinimetric analyses of other musculoskeletal diagnoses [15,40,48] and prior recommendations for optimal ICC analysis at 2 week intervals [49]. Nonetheless, we propose that longer term analyses are essential to establish reliability of PROMs for use in mid and long term randomized clinical trials.

4.2. Responsiveness

The NPRS and ODI demonstrated very good responsiveness (AUC) in the "improved" (AUC: NPRS 0.85, ODI 0.92) and "much improved" (AUC: NPRS 0.83, ODI 0.83) groups (Table 3), while the RMDQ demonstrated very good responsiveness in the "improved" group (AUC 0.85) and good responsiveness in the "much improved" group (AUC 0.73). Cleland et al [7] reported much lower AUC values for the NPRS

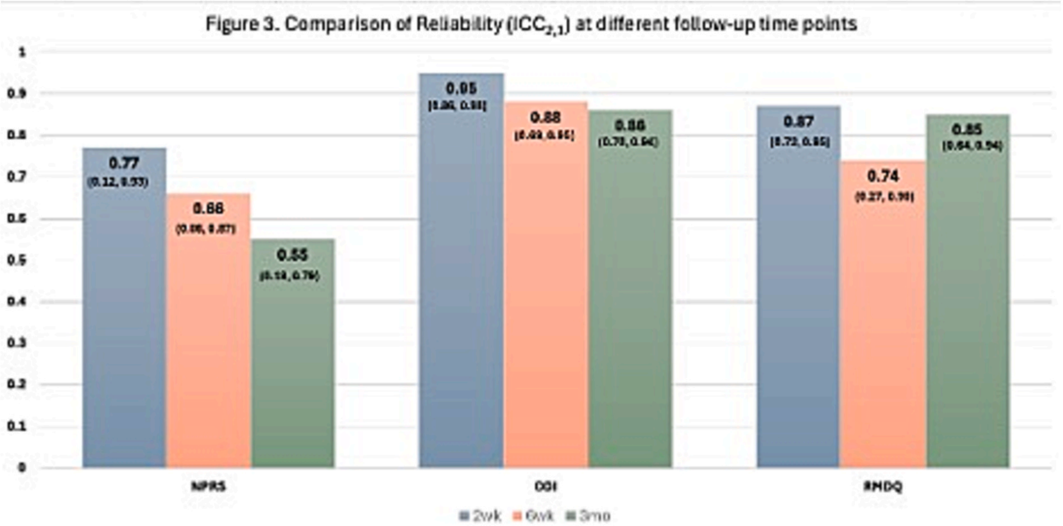

Fig. 3. Comparison of reliability ($ICC_{2,1}$) at different follow-up time points. Reliability = $ICC_{2,1}$ (95 % confidence interval), NPRS = numeric pain rating scale, ODI = Oswestry disability index, RMDQ = Roland Morris disability questionnaire.

(back/buttock, AUC 0.67; thigh/leg, AUC 0.65) with a much smaller sample (n = 35) of "improved" patients" compared to the current study (n = 105). To date, this is the first study to examine the responsiveness of the RMDQ in patients with LSS treated conservatively with outpatient physical therapy. The RMDQ has been reported to have a comparable responsiveness in post-surgical decompression patients with LSS (AUC 0.82) [50], and a wider range of responsiveness in patients with lumbar disc herniation (AUC range: 0.66 to 0.87) [8,9]. However, short term follow-up [9], the use of a non-validated global perceived effect scale [9,50,51] and a non-conservative treatment approach (surgery) [50] should be noted in those analyses. Overall, the NPRS, ODI and RMDQ demonstrate suitable responsiveness for their clinical application in patients with LSS.

4.3. MDC & MCID

In the current analysis, the MCID of the NPRS for the "improved" group (1.3 points) is consistent with findings reported by Cleland et al [7] (NPRS: back/buttock 1.3 points, thigh/leg 1.5 points) in "improved" patients with LSS. However, the MCID did not exceed measurement error ($MDC_{90/95}$) in the "improved" group (Table 3). [7] Nevertheless, in the current study, the MCID in patients that were "much improved" (2.8 points) exceeded measurement error (MDC_{90} 2.0 points MDC_{95} 2.3 points). Regarding the MCID of the ODI, the cut-scores did not exceed measurement error in the "improved" group in the Cleland et al study (MDC_{90} 13.1 points; MCID 5.0 points) or the current study (MCID 5.5 points; MDC_{90} 7.2 points). In the current study, like the NPRS, the MCID of patients in the "much improved" group (8.5 points) exceeded measurement error at both the 90 % and 95 % confidence intervals (MDC_{90} 7.2 points, MDC_{95} 8.5 points). Nevertheless, for the RMDQ, the MCID in both the "improved" (3.5 points) and "much" improved groups (4.5 points) did not exceed measurement error (MDC_{90} 5.1 points, MDC_{95} 6.1 points). The MCID of the RMDQ should be interpreted with caution in this patient population, and further research is warranted.

In the present analysis, choosing a more robust anchor/level of improvement, i.e., –"much-improved" (GROC: +5 to + 7), ensured the MCID values exceeded and are technically free from random measurement error. [52,53] Similar findings have been reported in other musculoskeletal diagnoses and their associated PROMs. [15,40,48] Standardized statistical methodology as well as a more comprehensive analysis (i.e., multiple improvement categories) may be essential to avoid an over/under estimation, and misinterpretation of these clinimetric measures in the research and clinical setting. Furthermore, as highly advanced/skilled interventions demonstrate greater efficacy, the change scores for the MCID should proportionally increase. [52–54].

4.4. Study strengths & limitations

The strengths of this study include: 1) comparative analysis of $ICC_{2,1}$ at multiple follow-up time frames, and 2) comparative analysis of multiple anchor-based improvement categories for future clinical/research applications. The are also several limitations that need to be outlined in this secondary analysis. First, the NPRS included "overall low back, buttock and leg pain" instead of individual NPRS ratings for each region. This raises a potential inconsistency, as the ODI does not specifically capture leg pain-related disability and the RMDQ partially reflects lower-extremity symptoms. Such inconsistency may have led to the reduced reliability of the NPRS or weaker criterion validity when comparing NPRS with GROC. Second, although our analyses (Pearson's r and bootstrapping) confirmed significant correlation between all PROMs and the GROC at 3 months, in a small "stable" group, performing subgroup analyses based on an external criterion like the GROC may potentially introduce selection bias and external validity issues. Furthermore, selection bias may arise due to the specific inclusion/exclusion criteria and participant characteristics in the original trial, possibly limiting generalizability. In light of these potential issues, future research using independent external datasets is warranted. Finally, the MCID of the RMDQ should be interpreted with caution, as the values for both "improved" and "much improved" patients did not exceed measurement error.

5. Conclusion

The NPRS, ODI and RMDQ demonstrated statistically acceptable reliability, construct validity, and responsiveness. The ODI had the most consistent reliability at each follow-up point. Clinicians and researchers should expect a 2.8-point change on the NPRS and an 8.5-point change on the ODI in the "much improved" patients to be considered clinically meaningful. The MCID exceeded measurement error ($MDC_{90/95}$) in both the NPRS and ODI when selecting a more robust level of improvement ("much improved") on the GROC. The MCID of the RMDQ did not exceed measurement error in either improvement category. In light of the potential limitations of selection bias and external validity in a secondary analysis, the results of this study should be interpreted with caution.

CRediT authorship contribution statement

Ian Young: Writing – review & editing, Writing – original draft, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **James Dunning:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **James Escaloni:** Writing – review & editing, Conceptualization. **Filippo Maselli:** Writing – review & editing, Writing – original draft, Formal analysis. **Joshua Prall:** Writing – review & editing, Methodology, Conceptualization. **Firas Mourad:** Writing – review & editing, Methodology. **Nathan Hutting:** Writing – review & editing, Methodology, Conceptualization. **César Fernández-de-las-Peñas:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Fatoye F, Gebrye T, Mbada CE, Useh U. Clinical and economic burden of low back pain in low- and middle-income countries: a systematic review. *BMJ Open*. Apr 25 2023;13(4):e064119. doi:10.1136/bmjopen-2022-064119.

[2] Fatoye F, Gebrye T, Ryan CG, Useh U, Mbada C. Global and regional estimates of clinical and economic burden of low back pain in high-income countries: a systematic review and meta-analysis. Front Public Health 2023;11:1098100. https://doi.org/10.3389/fpubh.2023.1098100.

[3] Ostelo RW, de Vet HC. Clinically important outcomes in low back pain. Best Pract Res Clin Rheumatol Aug 2005;19(4):593–607. https://doi.org/10.1016/j.berh.2005.03.003.

[4] Smeets R, Köke A, Lin CW, Ferreira M, Demoulin C. Measures of function in low back pain/disorders: low Back pain Rating Scale (LBPRS), Oswestry Disability Index (ODI), Progressive Isoinertial Lifting Evaluation (PILE), Quebec Back Pain Disability Scale (QBPDS), and Roland-Morris Disability Questionnaire (RDQ). Arthritis Care Res (Hoboken) Nov 2011;63(Suppl 11):S158–73. https://doi.org/10.1002/acr.20542.

[5] Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L, Grunnet-Nilsson N. Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. *BMC Musculoskelet Disord*. Oct 25 2006;7:82. doi:10.1186/1471-2474-7-82.

[6] Childs JD, Piva SR, Fritz JM. Responsiveness of the numeric pain rating scale in patients with low back pain. *Spine (Phila Pa 1976)*. Jun 1 2005;30(11):1331-4. doi:10.1097/01.brs.0000164099.92112.29.

[7] Cleland J, Whitman J, Houser J, Wainner R, Childs J. Psychometric properties of selected tests in patients with lumbar spinal stenosis. Spine J: Off J North Am Spine Soc 2012;12(10).

[8] Ghaderi Niri H, Ghanavati T, Mostafaee N, et al. Oswestry Disability Index, Roland-Morris Disability Questionnaire, and Quebec Back Pain Disability Scale: Responsiveness and Minimal Clinically Important changes in iranian people with Lumbar Disc Herniation following Physiotherapy. Arch Bone Jt Surg 2024;12(1):58–65. https://doi.org/10.22038/abjs.2023.72246.3366.

[9] Yao M, Xu B-p, Li Z-j, et al. A comparison between the low back pain scales for patients with lumbar disc herniation: validity, reliability, and responsiveness. Health Qual Life Outcomes 2020;18(1):175. https://doi.org/10.1186/s12955-020-01403-2.

[10] Jensen RK, Jensen TS, Koes B, Hartvigsen J. Prevalence of lumbar spinal stenosis in general and clinical populations: a systematic review and meta-analysis. Eur Spine J Sep 2020;29(9):2143–63. https://doi.org/10.1007/s00586-020-06339-1.

[11] Young I, Dunning J, Butts R, et al. Spinal manipulation and electrical dry needling as an adjunct to conventional physical therapy in patients with lumbar spinal stenosis: a multi-center randomized clinical trial. Spine J. Dec 14 2023;doi:10.1016/j.spinee.2023.12.002.

[12] Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Clinical Trial Research Support, Non-U.S. Gov't. Control Clin Trial. 1989;10(4):407-15.

[13] Jensen MP, Karoly P, Braver S. The measurement of clinical pain intensity: a comparison of six methods. Pain Oct 1986;27(1):117–26.

[14] Farrar JT, Young JP, Jr., LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Clinical Trial Controlled Clinical Trial Multicenter Study. *Pain*. Nov 2001;94(2):149-58.

[15] Young IA, Dunning J, Butts R, Cleland JA, Fernandez-de-Las-Penas C. Psychometric properties of the Numeric Pain Rating Scale and Neck Disability Index in patients with cervicogenic headache. Cephalalgia Jan 2019;39(1):44–51. https://doi.org/10.1177/0333102418772584.

[16] Clement RC, Welander A, Stowell C, et al. A proposed set of metrics for standardized outcome reporting in the management of low back pain. Acta Orthop 2015;86(5):523–33. https://doi.org/10.3109/17453674.2015.1036696.

[17] Goldsmith ES, Taylor BC, Greer N, et al. Focused evidence Review: Psychometric Properties of Patient-Reported Outcome measures for Chronic Musculoskeletal Pain. J Gen Intern Med May 2018;33(Suppl 1):61–70. https://doi.org/10.1007/s11606-018-4327-8.

[18] Jamison RN, Raymond SA, Slawsby EA, McHugo GJ, Baird JC. Pain assessment in patients with low back pain: comparison of weekly recall and momentary electronic data. J Pain Mar 2006;7(3):192–9. https://doi.org/10.1016/j.jpain.2005.10.006.

[19] Jensen MP, Mardekian J, Lakshminarayanan M, Boye ME. Validity of 24-h recall ratings of pain severity: biasing effects of "Peak" and "end" pain. Pain 2008;137(2):422–7. https://doi.org/10.1016/j.pain.2007.10.006.

[20] Bolton JE, Humphreys BK, van Hedel HJ. Validity of weekly recall ratings of average pain intensity in neck pain patients. J Manipulative Physiol Ther Oct 2010;33(8):612–7. https://doi.org/10.1016/j.jmpt.2010.08.009.

[21] Chow NW, Southerst D, Wong JJ, Kopansky-Giles D, Ammendolia C. Clinical Outcomes in Neurogenic Claudication using a Multimodal Program for Lumbar Spinal Stenosis: a Study of 49 patients with prospective long-term follow-up. J Manipulative Physiol Ther Mar - Apr 2019;42(3):203–9. https://doi.org/10.1016/j.jmpt.2018.11.004.

[22] Salaffi F, Stancati A, Silvestri CA, Ciapetti A, Grassi W. Minimal clinically important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. Eur J Pain Aug 2004;8(4):283–91. https://doi.org/10.1016/j.ejpain.2003.09.004.

[23] Fairbank JC, Pynsent PB. The Oswestry Disability Index. *Spine (Phila Pa 1976)*. Nov 15 2000;25(22):2940-52; discussion 2952. doi:10.1097/00007632-200011150-00017.

[24] Lee C, Fu T, Liu C, Hung C. Psychometric evaluation of the Oswestry Disability Index in patients with chronic low back pain: factor and Mokken analyses. Heat Qual Life Outcomes 2017;15(192).

[25] Brodke DS, Goz V, Lawrence BD, Spiker WR, Neese A, Hung M. Oswestry Disability Index: a psychometric analysis with 1,610 patients. Spine J Mar 2017;17(3):321–7. https://doi.org/10.1016/j.spinee.2016.09.020.

[26] Vianin M. Psychometric properties and clinical usefulness of the Oswestry Disability Index. J Chiropr Med Dec 2008;7(4):161–3. https://doi.org/10.1016/j.jcm.2008.07.001.

[27] Pratt RK, Fairbank JC, Virr A. The reliability of the Shuttle Walking Test, the Swiss Spinal Stenosis Questionnaire, the Oxford Spinal Stenosis Score, and the Oswestry Disability Index in the assessment of patients with lumbar spinal stenosis. *Spine (Phila Pa 1976)*. Jan 1 2002;27(1):84-91. doi:10.1097/00007632-200201010-00020.

[28] Davidson M, Keating JL. A comparison of five low back disability questionnaires: reliability and responsiveness. Phys Ther Jan 2002;82(1):8–24. https://doi.org/10.1093/ptj/82.1.8.

[29] Davidson M, Keating J. Oswestry Disability Questionnaire (ODQ). Aust J Physiother 2005;51(4):270. https://doi.org/10.1016/s0004-9514(05)70016-7.

[30] Ferreira ML, Ferreira PH, Latimer J, et al. Comparison of general exercise, motor control exercise and spinal manipulative therapy for chronic low back pain: a randomized trial. Pain Sep 2007;131(1–2):31–7. https://doi.org/10.1016/j.pain.2006.12.008.

[31] 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 (Phila Pa 1976)*. Mar 1983;8(2):141-4. doi:10.1097/00007632-198303000-00004.

[32] Chiarotto A, Maxwell LJ, Terwee CB, Wells GA, Tugwell P, Ostelo RW. Roland-Morris disability questionnaire and oswestry disability index: which has better measurement properties for measuring physical functioning in nonspecific low back pain? System Rev Meta-Anal Phys Ther Oct 2016;96(10):1620–37. https://doi.org/10.2522/ptj.20150420.

[33] Chapman JR, Norvell DC, Hermsmeyer JT, et al. Evaluating common outcomes for measuring treatment success for chronic low back pain. *Spine (Phila Pa 1976)*. Oct 1 2011;36(21 Suppl):S54-68. doi:10.1097/BRS.0b013e31822ef74d.

[34] Stratford PW, Binkley JM, Riddle DL, Guyatt GH. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. Phys Ther Nov 1998;78(11):1186–96. https://doi.org/10.1093/ptj/78.11.1186.

[35] Maughan EF, Lewis JS. Outcome measures in chronic low back pain. Eur Spine J Sep 2010;19(9):1484–94. https://doi.org/10.1007/s00586-010-1353-6.

[36] Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials*. Dec 1989;10(4):407-15.

[37] Marchand AA, Suitner M, O'Shaughnessy J, Chatillon CE, Cantin V, Descarreaux M. Feasibility of conducting an active exercise prehabilitation program in patients awaiting spinal stenosis surgery: a randomized pilot study. Sci Rep Aug 22 2019;9(1):12257. doi:10.1038/s41598-019-48736-7.

[38] Whitman JM, Flynn TW, Childs JD, et al. A comparison between two physical therapy treatment programs for patients with lumbar spinal stenosis: a randomized clinical trial. Spine (Phila Pa 1976). Oct 15 2006;31(22):2541-9. doi:10.1097/01.brs.0000241136.98159.8c.

[39] Thoomes-de Graaf M, Scholten-Peeters W, Duijn E, et al. The Responsiveness and interpretability of the shoulder pain and disability index. J Orthop Sports Phys Ther Apr 2017;47(4):278–86. https://doi.org/10.2519/jospt.2017.7079.

[40] Young I, Dunning J, Mourad F, Escaloni J, Bliton P, Fernández-de-Las-Peñas C. Clinimetric analysis of the visual analogue scale and pain free mouth opening in patients with muscular temporomandibular disorder. *Cranio*. Feb 10 2025:1-7. doi:10.1080/08869634.2025.2464227.

[41] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull Mar 1979;86(2):420–8. https://doi.org/10.1037//0033-2909.86.2.420.

[42] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med Jun 2016;15(2):155–63. https://doi.org/10.1016/j.jcm.2016.02.012.

[43] Schober P, Boer C, Schwarte LA. Correlation Coefficients: Appropriate Use and Interpretation. Anesth Analg May 2018;126(5):1763–8. https://doi.org/10.1213/ane.0000000000002864.

[44] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology Apr 1982;143(1):29–36. https://doi.org/10.1148/radiology.143.1.7063747.

[45] Nahm FS. Receiver operating characteristic curve: overview and practical use for clinicians. Korean J Anesthesiol Feb 2022;75(1):25–36. https://doi.org/10.4097/kja.21209.

[46] Beaton DE, Bombardier C, Katz JN, et al. Looking for important change/differences in studies of responsiveness. OMERACT MCID Working Group. Outcome Measures in Rheumatology. Minimal Clinically Important Difference. J Rheumatol Feb 2001; 28(2):400-5.

[47] Swets JA. Measuring the accuracy of diagnostic systems. Science 1988;240(4857): 1285–93. https://doi.org/10.1126/science.3287615.

[48] Young I, Dunning J, Mourad F, Escaloni J, Bliton P, Fernández-de-Las-Peñas C. Clinimetric analysis of the numeric pain rating scale, patient-rated tennis elbow evaluation, and tennis elbow function scale in patients with lateral elbow tendinopathy. *Physiother Theory Pract*. Jan 10 2025:1-9. doi:10.1080/09593985.2025.2450090.

[49] Streiner DL, Norman GR, Cairney J. Health Measurement Scales: a Practical Guide to their Development and Use. Oxford University Press; 2014.

[50] Vishwanathan K, Braithwaite I. Construct validity and responsiveness of commonly used patient reported outcome instruments in decompression for lumbar spinal stenosis. J Clin Orthop Trauma May 2021;16:125–31. https://doi.org/10.1016/j.jcot.2021.01.002.

[51] Fischer D, Stewart AL, Bloch DA, Lorig K, Laurent D, Holman H. Capturing the patient's view of change as a clinical outcome measure. J Am Med Assoc 1999;282 (12):1157–62. https://doi.org/10.1001/jama.282.12.1157.

[52] Grönkvist R, Vixner L, Äng B, Grimby-Ekman A. Measurement error, minimal detectable change, and minimal clinically important difference of the short Form-36 health survey, hospital anxiety and depression scale, and pain numeric rating scale in patients with chronic pain. J Pain 2024;25(9):104559. https://doi.org/10.1016/j.jpain.2024.104559.

[53] Copay AG, Chung AS, Eyberg B, Olmscheid N, Chutkan N, Spangehl MJ. Minimum clinically important difference: current trends in the orthopaedic literature, part i: upper extremity: a systematic review. JBJS Rev Sep 2018;6(9):e1.

[54] Draak THP, de Greef BTA, Faber CG, Merkies ISJ. The minimum clinically important difference: which direction to take. Eur J Neurol Jun 2019;26(6):850–5. https://doi.org/10.1111/ene.13941.