Seminars in Arthritis and Rheumatism 78 (2026) 152954

Contents lists available at ScienceDirect

Seminars in Arthritis and Rheumatism

journal homepage: www.elsevier.com/locate/semarthrit

Suitability of the Numerical Pain Rating Scale for measuring pain in clinical trials evaluating interventions for people with shoulder disorders according to the OMERACT filter 2.2

Rochelle Furtado [a,*], Hana Marmura [b], Romi Haas [c], Sofia Ramiro [d], Arianne Verhagen [e], Samuel Whittle [c,f], Dorcas Beaton [g], Beverley Shea [h], Pamela Richards [i], Joel Gagnier [j], Rachelle Buchbinder [c]

[a] University Health Network - KITE, Toronto, Canada
[b] University of North Carolina at Chapel Hill, United States
[c] Musculoskeletal Health Unit and Wiser Health Care Group, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia
[d] Leiden University Medical Center, Leiden, and Zuyderland Medical Center, Heerlen, The Netherlands
[e] Graduate School of Health, University of Technology Sydney, Sydney, Australia
[f] The Queen Elizabeth Hospital, Adelaide, Australia
[g] Institute of Health & Work and the University of Toronto, Toronto, Canada
[h] Ottawa Hospital Research Institute, University of Ottawa, Canada
[i] Patient Research Partner, Bristol, UK
[j] Departments of Epidemiology & Biostatistics and Surgery, Western University, Canada

A B S T R A C T

Background: The Outcome Measures in Rheumatology (OMERACT) group recommends pain as a core domain in clinical trials for shoulder disorders. This study evaluated the suitability of the Numerical Pain Rating Scale (NPRS) for measuring pain using the OMERACT Filter 2.2.
Methods: Following the OMERACT Handbook for instrument selection, a systematic review assessed the NPRS's construct validity, reliability, longitudinal construct validity, clinical trial discrimination, and thresholds of meaning. Articles were independently screened, appraised with the COSMIN-OMERACT Good Methods Checklist, and rated as green (good), amber (caution), red (poor), or white (no evidence). Only green and amber studies were included. Findings were summarized in a Summary of Measurement Properties table and discussed at the 2025 OMERACT Shoulder Special Interest Group workshop.
Results: Twelve studies, including 18 pieces of evidence met eligibility criteria. Three studies examined construct validity, two test-retest reliability, four longitudinal construct validity, six clinical trials discrimination and three thresholds of meaning. Six (33%) components of evidence had good methods and were rated green and 12 were rated amber. There was significant heterogeneity in NPRS versions and formats evaluated across studies, questioning the validity of synthesizing the data together. Therefore, 75% of respondents at the 2025 OMERACT conference agreed that synthesizing results from studies using different NPRSs was inappropriate.
Conclusions: The NPRS cannot progress to the endorsement stage for the core domain of pain for shoulder conditions due to insufficient evidence for any single NPRS version. Current evidence is insufficient to support its use to measure pain in clinical trials of shoulder disorders.

Introduction

Since 2015, the Outcome Measures in Rheumatology (OMERACT) Shoulder Working Group has been working to generate both a core domain and core outcome measurement set specifically for shoulder disorders, in accordance with OMERACT methodology [1–3] A systematic review conducted by the group identified 409 trials investigating interventions for shoulder disorders, in which 32 outcome domains were assessed using 319 different measurement instruments [2] An international Delphi study including clinicians, researchers and people with lived experience of shoulder disorders from 13 different countries considered which outcome domains should be included in a core domain set [3,4] In 2016, a preliminary core domain set was presented at an OMERACT shoulder special interest group meeting,[2] and in 2018, OMERACT endorsed a core domain set for clinical trials of shoulder disorders [4] The core set includes four mandatory domains: pain, patient global-shoulder, physical function/activity and adverse events, including death.

Following the development of the core domain set, we commenced the instrument selection phase of the OMERACT process, beginning with

* Corresponding author at: University Health Network - KITE, Toronto, Canada.
E-mail address: rochelle.furtado@uhn.ca (R. Furtado).

https://doi.org/10.1016/j.semarthrit.2026.152954

Available online 20 February 2026

the pain domain. Pain was prioritised as the first domain based on patient partners' feedback that it is most important to them. Our systematic review identified 38 instruments used to measure pain in individuals with shoulder disorders, and the Numeric Pain Rating Scale (NPRS) was one of the most frequently used [4] The NPRS is a patient-reported outcome measure of pain intensity and it has been used in many contexts for different musculoskeletal disorders.

In accordance with the OMERACT process, candidate instruments are evaluated using the Filter 2.2, which considers evidence for truth (validity), discrimination (reliability and responsiveness) and feasibility [5] The purpose of this study was to evaluate the suitability of the NPRS to be included within a core outcome measurement set for evaluating shoulder pain in clinical trials of individuals with shoulder disorders using this filter.

Methods

This study followed the methods outlined in the OMERACT Handbook on Instrument Selection for Core Outcome Measurement Sets to determine the truth, discrimination, and feasibility of the NPRS [5] The selection process includes reviewing published evidence for candidate instruments (NPRS), determining whether the chosen instruments reflect the target domain (shoulder pain), determining their appropriateness for clinical use (truth, discrimination and feasibility filter) and reviewing evidence of instrument performance across key measurement properties (validity, reliability, responsiveness). The OMERACT handbook also details a traffic light rating at key decision points: Green indicates good to proceed, Amber indicates proceed with caution due to some concerns or limitations and Red indicates the need to stop (Fig. 1).

The NPRS is a unidimensional pain scale that measures pain intensity. It is usually presented as an 11-point scale with scores from 0 to 10 and anchors of 0 = no pain and 10 = worst possible pain [6] It has been used to assess shoulder pain intensity under various conditions, such as pain at rest, pain with normal activities, average pain and worst pain [6,7] It is easy to administer and score and is easily translated for use in other languages [6–8]

Domain Match (Truth) and Feasibility

The process of domain matching addresses whether the instrument appears to be a good match for the target domain and whether the instrument has appropriate content to represent the experience of the intended target population and study situation for the target domain. Pain, reflecting pain intensity, was defined as the extent of shoulder pain experienced, including pain at rest, during and after activity and at night [1,2] To assess whether the NPRS matched the domain definition of pain, OMERACT Shoulder Working Group members were asked to consider relevance, completeness, redundancy, clarity of the items, suitability of the response options and the appropriateness of the scoring method.

The feasibility assessment for an outcome measure is based on factors related to the practical use of a given instrument. To assess feasibility, OMERACT Shoulder Working Group members were asked to consider the overall ease of use, cost, burden (patient, responder), method of administration, equipment needs, potential copyright issues, and language or culture needs. Domain match and feasibility were each given a separate traffic light rating, and if an instrument was rated green or amber for both domain match (truth) and feasibility, it would move onto the next stage of the OMERACT process: systematic synthesis of the evidence to determine the measurement properties of the instrument.

Systematic review of measurement properties of the NPRS

We conducted a systematic review to synthesize the available evidence on the measurement properties of the NPRS. The study protocol was registered on PROSPERO CRD42023440148 and reported according to the PRISMA-COSMIN guidance for reporting systematic reviews of outcome measurement instruments [9]

Selection Criteria and Search Methods

Table 1 outlines the eligibility criteria for our search. Studies investigating measurement properties of the NPRS for individuals with shoulder conditions were included. Studies were excluded in >25% of

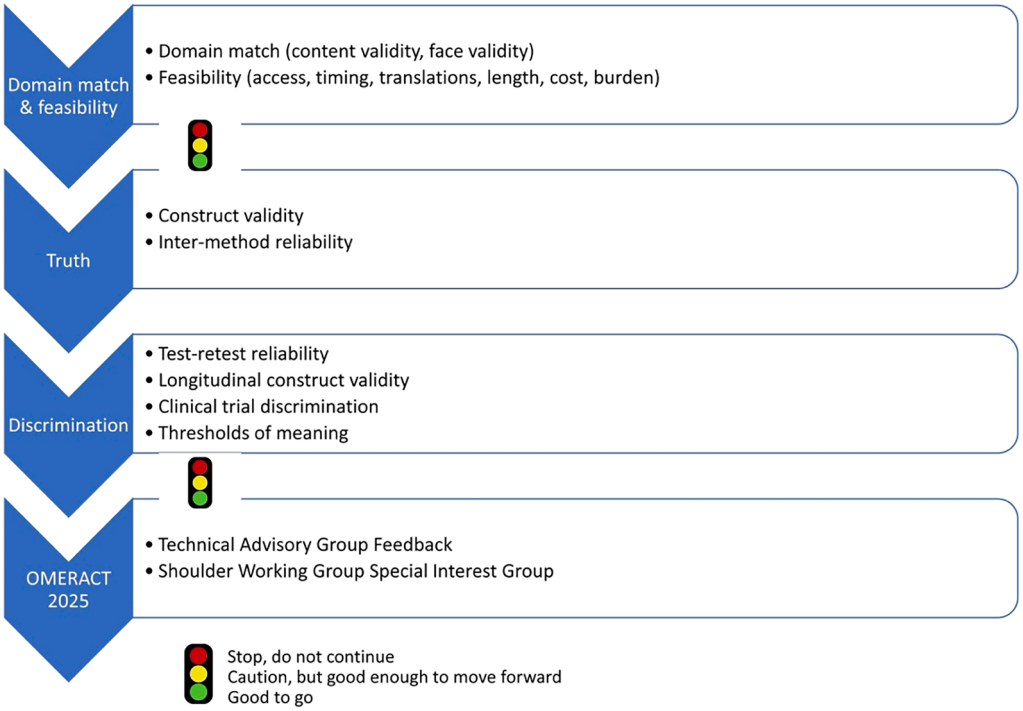

Fig. 1. Instrument selection process for assessing the suitability of the Numeric Pain Rating Scale to measure pain intensity in clinical trials of shoulder disorders.

Table 1
Summary of the eligibility criteria.

Criteria	Inclusion	Exclusion
Population	• Adults with rotator cuff related shoulder pain (tendinopathy, impingement, subacromial bursitis, partial and full thickness tears), adhesive capsulitis, instability, glenohumeral osteoarthritis, dislocation, proximal humeral or humeral head fractures or unspecified shoulder pain • ≥ 75% of the sample with an included shoulder condition • Mixed sample studies if results were reported separately for at least one eligible shoulder disorder	• Inflammatory conditions e. g., rheumatoid arthritis, axial spondyloarthritis, psoriatic arthritis • Paediatric population • Acromioclavicular conditions • Shoulder pain after breast cancer surgery • If ≥25% of the sample has an excluded condition
Instruments	• Numerical Pain Rating Scale (English version)	
Measurement properties	• Floor and ceiling effects* • Construct validity • Test-retest reliability • Inter-rater reliability • Responsiveness • Clinical trial discrimination • Thresholds of meaning	

* Studies assessing floor or ceiling effects were included in our systematic review because raw

Bottom of Form.

the population included any of the following: inflammatory conditions (rheumatoid arthritis, axial spondylarthritis, psoriatic arthritis), paediatric population, acromioclavicular conditions, or shoulder pain after breast cancer surgery.

We searched Medline via Ovid, Embase via Ovid, and Cumulative Index of Nursing and Allied Health (CINAHL) via EBSCO from inception to 5 June 2023. We used the Boolean operator "AND" to combine search terms relating to shoulder disorders, instruments and studies of measurement properties. A validated, sensitive search filter for identifying studies of measurement properties was used in Medline and adapted for the other databases (supplementary table 1). To identify randomised controlled trials using the NPRS as an outcome, we consulted a review of adult shoulder function measures[10] and conducted a targeted Google search using "NPRS" and "trial". Reference lists of included studies and relevant review articles were also screened.

Study Selection

Initial screening of titles and abstracts, followed by full-text review of potentially eligible articles, was independently conducted by pairs of authors (RH, DB, and/or AF) using Covidence systematic review software (Veritas Health Innovation, Melbourne, Australia) in June 2023. Discrepancies were resolved through discussion with additional members of the Working Group (SR, RB, JG, and/or AV) until consensus was reached. During full-text review, eligible studies were tagged according to the measurement properties they assessed, and each component of evidence was displayed in the Summary of Measurement Properties (SOMP) table.

Methodological Quality Appraisal

Pairs of reviewers (RF, HM, RH, DB) independently assessed the methodological quality of each study using the COSMIN-OMERACT Good Methods Checklist [5] For each study evaluating one or more components of measurement property evidence, reviewers completed the relevant checklist items and provided comments. Each reviewer then provided an overall judgement on whether to include or exclude each component of evidence using the traffic light rating. Discrepancies were discussed and resolved by consensus with additional members of the investigator team (SR, JG, AV, SW, and/or RB). See supplementary Tables 2 and 3 for detailed information.

Data extraction

Components of evidence rated as green or amber on the COSMIN-OMERACT Good Methods Checklists were included for data extraction and further review of performance adequacy. For each included component, data were extracted using the OMERACT data extraction template and included details on study design, population, methodology, and the reported results for each measurement property. Data extraction was performed independently by pairs of reviewers for two studies per measurement property (RF, HM, RH, DB). As discrepancies were minimal, the remaining data were extracted by one reviewer and independently verified by another. Any discrepancies were resolved through discussion.

Definitions, relevant statistics and interpretation thresholds for each measurement property were based on standards outlined in the OMERACT filter 2.2 and are shown in Supplementary Table 2 and 3. For construct validity and responsiveness (longitudinal construct validity), observed correlations between the NPRS and other instruments were compared with a priori hypotheses. For clinical trial discrimination, between-group effect sizes were evaluated against expected differences. Where there were no explicit a priori hypotheses, we made hypotheses, based on prior literature where possible.

Adequacy of performance

For each included study, pairs of reviewers (RF, HM, RH, DB) independently assessed the adequacy of performance for each measurement property. Discrepancies were resolved through consensus within the investigator team. Performance was rated as either adequate (+) if ≥75% of hypotheses were confirmed, equivocal (±) if 25-74% were confirmed, and inadequate (-) if <25% were confirmed for all measurement properties except test-retest reliability and thresholds of meaning. For test-retest reliability, performance was rated as adequate (+) if the intraclass correlation coefficient (ICC) was ≥0.75, and inadequate (-) if <0.75. For thresholds of meaning where no established standards for calculated thresholds exist, performance was considered adequate (+) if thresholds were based on at least one anchor reflecting meaningful within-patient change, and inadequate (-) if based solely on a distribution-based approach.

Summary of findings for each measurement property

The evidence supporting each measurement property was presented in a summary of measurement properties table (SOMP) [11] Given the heterogeneity in NPRS scales reported throughout the literature, we planned to synthesize results only across studies that used the same NPRS instrument. The OMERACT guide dictates that the synthesis ratings for each measurement property consider the quality (rated as Green/Amber or Red), quantity, adequacy of performance (rated as +, ± or -) and consistency of results across studies. A Green rating was assigned if there were at least two good quality (green) studies showing consistent findings with adequate performance (+). A Red rating was assigned if there were at least two good quality articles (green) with inconsistent findings or one good quality study (green) with inadequate performance (-). White was assigned if there no evidence was available. Amber was assigned to all other situations, see Supplementary Table 4 for more information.

OMERACT Technical Advisory Group (TAG) review and OMERACT 2025 Special Interest Group Workshop

All findings and overall ratings presented in the SOMP tables were reviewed by the OMERACT TAG prior to the OMERACT 2025 conference. Our findings and TAG feedback were presented at the OMERACT 2025 Shoulder Working Group Special Interest workshop. At the end of this workshop session, we planned online polling to address two specific questions: (a) whether participants agreed with our ratings for the synthesis of each measurement property? and (b) We synthesized all NPRS scales regalrdless of context. Do you think this was appropriate?

Results

Domain Match and Feasibility

The NPRS passed the Truth and Feasibility Filter, with a consensus amber rating for domain match (truth) and a green rating for feasibility. Noted limitations of the instrument leading to an amber rating for domain match included its limited breadth (i.e., only one question) and variations in format (i.e., differing question prompts, scales, anchors and types of pain evaluated).

Study identification and characteristics

The search identified 11,372 records. Of these, 12 studies met the eligibility criteria to be included for the assessment of measurement properties of the NRPS (Fig. 2) [12–23] A total of 18 components of evidence were identified within the 12 studies: three studies examined construct validity,[13,20,21] two test-retest reliability, [16,21] four responsiveness, [15–17,23] six clinical trial discrimination, [12,14–16, 18,19] and three thresholds of meaning [15,16,22]

Detailed characteristics of the 12 included studies are in Table 2. The included studies were published between 2001 and 2020. All 12 were published in the English language and most (n=8) were conducted within the USA [13–17,19–21] Four studies were RCTs (n=4),[12,14, 18,19] four studies were longitudinal observational studies, [12,15,16, 23] three were secondary analyses of RCTs (n=3), [17,20,22] and one was a cohort study [2] The included studies were mainly conducted in outpatient clinics and hospital or academic physical therapy departments, [13–18,20–23] one was conducted at a military medical center [19] and one was conducted at a chiropractic college [12] Five studies included participants with unspecified/generic shoulder pain, [15,16,19–21] two included participants with adhesive capsulitis, [18, 23] four included participants with rotator cuff disease, [12,14,17,22] and one study included participants described as having shoulder range of motion impairments [13].

The median (IQR) sample size of the studies was 80 (52.5-112), with sample sizes ranging from 10 to 503 participants, mean age was 31 to 59. 5 years, and the proportion of male participants was 30 to 72%. Follow-up duration ranged from 48 hours to 12 months.

Instrument heterogeneity

The type and format of the NPRS instrument assessed varied across the included studies with no two studies assessing the same measurement property used the exact same NPRS instrument (Table 3). The number of scales, format and test conditions of the NPRS, including its wording, anchors, description, scale and recall period varied across the included studies. While some studies described the exact phrasing of the

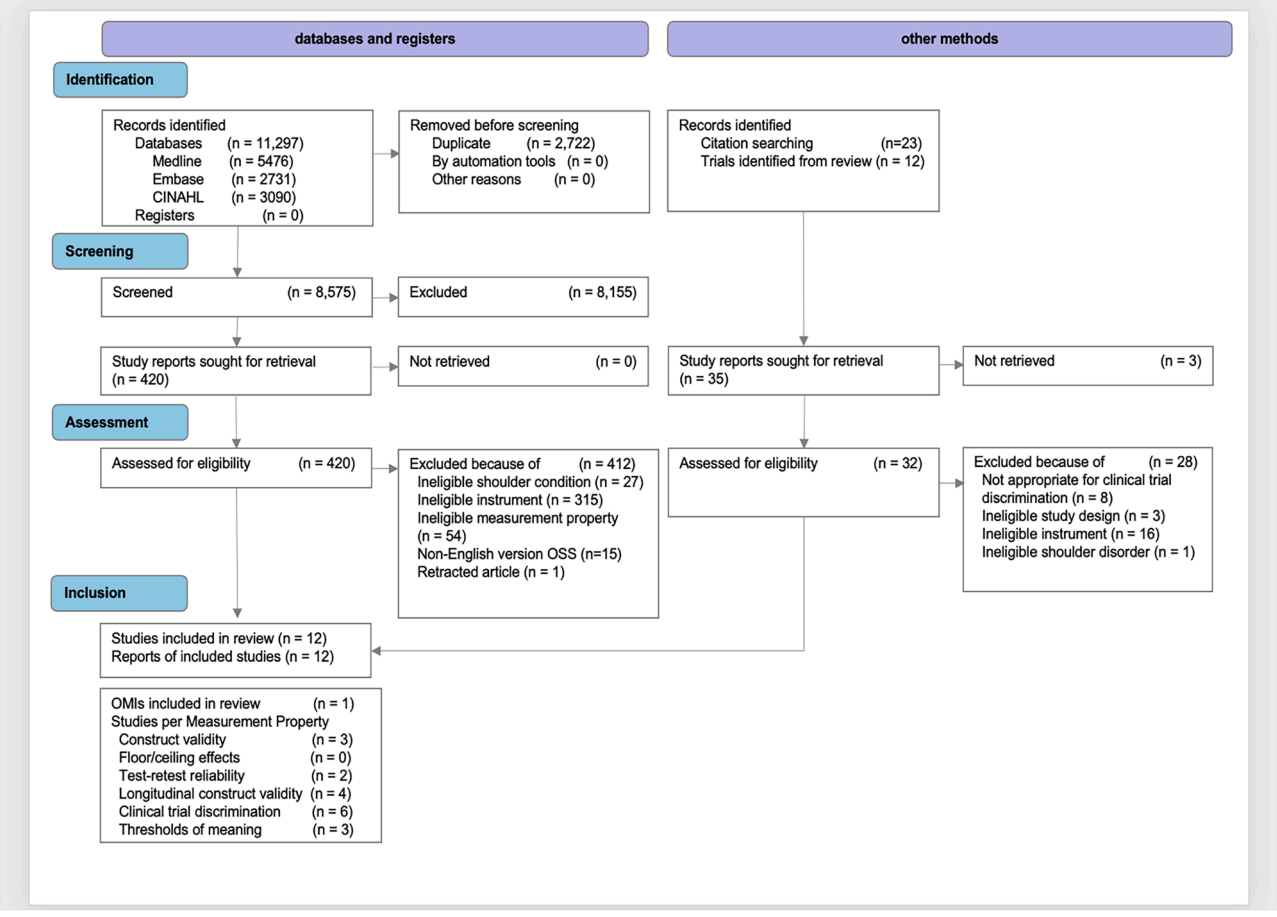

Fig. 2. PRISMA-COSMIN flow diagram of the systematic search conducted for articles evaluating measurement properties of the Numeric Pain Rating Scale (NPRS).

Table 2
Characteristics of included studies.

Observational studies

Author	Year	Country	Design	Population	Setting	Measurement properties	Sample size	Follow-up (final)	Age Mean (SD)	Male (%)	Baseline NPRS Mean (SD)	Additional Instruments
Hummel-Berry [13]	2001	USA	Longitudinal	Chronic impairments in shoulder range of motion	Hospital PT Dept	Construct Validity	55	12 weeks	59.5 (range 24-80)	50.9	Pain at rest: 3.4 (3.3) Pain with movement: 7.9 (2.1) 0-10	SST SF-36
vanMeeteren [23]	2006	Netherlands	Longitudinal	Unilateral adhesive capsulitis treated with intra-articular injection	Not reported	Longitudinal construct validity (responsiveness)	10	12 weeks	52.8 (range 35-77)	30	79.5 (21.1)	Shoulder Disability Questionnaire (Dutch Version)
Mintken[16]	2009	USA	Longitudinal	Shoulder pain	OPC	Test-retest reliability Longitudinal construct validity (responsiveness) Clinical trial discrimination Thresholds of meaning	101	27.2 (18.2) days	Stable patients: 44.4 (17.4) Improved patients: 39.1 (18.8)	Stable patients: 59 Improved patients: 66	4.4 (2.4)	Pain diagram QuickDASH GRC
Michener [15]	2011	USA	Longitudinal	Shoulder pain (mixed diagnoses, surgical and non-surgical conditions)	OPC	Longitudinal construct validity (responsiveness) Clinical trial discrimination Thresholds of meaning	136	3 to 4 weeks	51.7 (16.4)	48.5	5.0 (1.6)	PSS

Secondary Analyses of Cohort Studies or Randomised Controlled Trials (all patients analysed together to investigate validity, reliability, or responsiveness)

Author	**Year**	**Country**	**Design**	**Population**	**Setting**	**Intervention arms**	**Sample size**	**Follow-up (final)**	**Age** Mean (SD)	**Male (%)**	**Baseline NPRS**	**Instruments**
Tubach[22]	2006	France	Secondary analysis of 3-arm double blind double-dummy trial (all patients analysed together)	Acute painful flare with clinically diagnosed rotator cuff syndrome lasting 7 days	OPC	Thresholds of meaning	271	7 days	47.6 (10.4)	38.1	67.9 (14.7)	Neer rating to function subscale Response to NSAID treatment on GRC PASS
O'Halloran [17]	2013	USA	Secondary analysis of pragmatic 2-arm trial (all patients analysed together)	Shoulder impingement syndrome	OPC or academic PT clinic	Longitudinal construct validity (responsiveness)	68	56.1 (55.4) days (discharge from PT)	52.6 (14.1)	54.5	5.9 (2.1)	QuickDASH PASS RoR
Riley[20]	2015	USA	Secondary analysis of 2-arm trial (all patients analysed together)	Non-post-surgical unilateral shoulder pain	Academic OPC	Construct validity	92	Immediately post-treatment	48.7 (11.0)	38.1	Average pain: 4.3 (1.7) Present pain: 3.2 (2.3) Above shoulder pain: 4.3 (2.4) Hawkins-Kennedy pain: 4.8 (2.7) Neer pain: 4.8 (2.5)	SPADI

(*continued on next page*)

Table 2 (*continued*)

Observational studies												
Author	Year	Country	Design	Population	Setting	Measurement properties	Sample size	Follow-up (final)	Age Mean (SD)	Male (%)	Baseline NPRS Mean (SD)	Additional Instruments
Riley[21]	2019	USA	Secondary analysis of cohort study	Non-post-surgical unilateral shoulder pain	Hospital PT Dept	Construct validity Test-retest reliability	38	4 weeks	48.1 (15.6)	38.5	3.6 (2.5)	FABQ SPADI
Randomised Controlled Trials (Assessed for clinical trial discrimination)												
Author	**Year**	**Country**	**Design**	**Population**	**Setting**	**Intervention arms**	**Sample size**	**Follow-up (final)**	**Age** Mean (SD)	**Male (%)**	**Baseline NPRS**	**Instruments**
Atkinson[12]	2008	South Africa	2-arm	Rotator cuff tendinopathy	Chiropractic college	Shoulder manipulation (high velocity thrust, 3 sessions)	30	2 weeks	41.53 (range: 18 to 63)	73.3	43.2 (14.5)	-
						Sham laser therapy (6 sessions)	30		42 (range: 20 to 76)	70	40.7 (13.3)	
Rhon[19]	2014	USA	2-arm	Unilateral shoulder pain	Army medical center	Corticosteroid (up to 3 injections one month apart over one year)	52	24-48 hours	42 (12)	73	3.3 (95%CI 2.7 to 3.9)	SPADI GRC
						Manual physical therapy (twice a week for three weeks)	52		40 (12)	63	3.8 (95%CI 3.2 to 4.5)	
Kardouni [14]	2015	USA	2-arm	Unilateral subacromial impingement syndrome	Academic PT department	Thoracic spinal manipulative therapy	24	24-48 hours	31.1 (12.3)	42	3.5 (1.4)	PSS GRC
						Sham thoracic spinal manipulation	21		31.2 (12.1)	57	4.0 (1.4)	
Ranagan[18]	2020	UK	Pragmatic 3-arm multi-centre superiority	Unilateral adhesive capsulitis	Hospital PT department	Manipulation under anaesthesia + steroid injection + standard post-procedure PT	503	12 months	54.5 (7.7)	36		OSS QuickDASH EuroQOL 5D Perceived extent of recovery VAS
						Arthroscopic capsular release + manipulation under anaesthesia + standard post-procedure PT			53.9 (7.7)	38		
						Early structured PT + steroid injection			54.5 (7.8)	35		

Abbreviations.

USA: United States of America; OPC: outpatient clinic; PT: physical therapy; SD: standard deviation; SST: Simple Shoulder Test; SF-36: Medical Outcomes Study Short Form 36; PSS: Penn Shoulder Score; QuickDASH: Disabilities of the Arm, Shoulder and Hand Questionnaire Short Form; PASS: Patient Acceptable Symptom State; RoR: Rate of Recovery; SPADI: Shoulder Pain and Disability Index; FABQ: Fear-Avoidance Beliefs Questionnaire; GRC: Global Rating of Change; EQ-5D: EuroQOL 5-Dimension Questionnaire; VAS: Visual Analog Scale

Table 3
Description of Numerical Pain Rating Scales (NPRS) used in the included studies.

Author	Year	Number of NPRS Scales Used	Scale, Anchors	Type of Pain	Other Information
Observational Studies					
Hummel-Berry [13]	2001	2	0 to 10 0 = no pain 10 = most intense pain possible	1. Pain at rest 2. Pain during motion	
vanMeeteren [23]	2006	1	0 to 100 0 = no pain 100 = unbearable pain	Pain "at that moment" (current pain)	Labelled "NRS-101"
Mintken [16]	2009	1	0 to 10 0 = no pain 10 = worst imaginable pain	Not described	Study states: "Patients rate their current level of pain and their worst and least amount of pain in the last 24 hours. The average of the 3 ratings or any single rating may be used to represent the patient's level of pain." It is not clear which of these options are used in the study.
Michener [15]	2011	4	0 to 10 0 = no pain 10 = worst pain possible	1. Pain at rest 2. Pain with activities of daily living 3. Pain with strenuous activity 4. NPRS-average: mean of 3 scores above	The NPRS-average was the only score used. It was calculated by averaging the scores of the 3 pain items of the Penn Shoulder Score, which are each NPRS ratings of pain (#1-3 in type of pain column)
Secondary Analyses (of Cohort Studies or Randomised Controlled Trials)					
Tubach [22]	2006	1	0 to 10 Anchors not described	Pain on movement	0 to 10 score normalized and reported as 0 to 100.
O'Halloran [17]	2013	1	0 to 10 0 = no pain 5 = moderate pain 10 = worst possible pain	Not described	Reported as raw and percentage change scores.
Riley [20]	2015	5	0 to 10 0 = no pain 10 = worst pain imaginable	1. Average pain in the previous 24 hours 2. Present pain at rest 3. Pain with elevation above shoulder level 4. Pain with Hawkins-Kennedy test (average of 3 trials) 5. Pain with Neer test (average of 3 trials)	
Riley [21]	2019	1	0 to 10 0 = no pain 10 = pain as bad as it can be	Pain "at the time of testing" (current pain)	
Randomised Controlled Trials					
Atkinson [12]	2008	1	0 to 100 Anchors not described	Not described	Labelled "NRS-101"
Rhon [19]	2014	1	0 to 10 0 = no pain 10 = worst imaginable pain	Not described	"Used to assess pain intensity". It is not clear what type of pain is asked about.
Kardouni [14]	2015	2	0 to 10 0 = no pain at all 10 = pain as bad as it can be	1. "Please rate your shoulder pain at the present time" (current pain at baseline) "Now that you have had the manual therapy treatment to your thoracic spine, please rate your shoulder pain" (current pain, post-intervention/sham)	Pre- and post-intervention NPRS scales assessed as if they are the same scale, however the wording is different at each time point (see type of pain column)
Ranagan [18]	2020	1	0 to 10 Anchors not described	Not described	Anchors in study cited for use of NPRS uses anchors of 0 = no pain and 10 = worst possible pain.

question of the NPRS given to participants in a study, others simply referenced the NPRS as a well-known instrument to assess pain in the methods section and did not provide any specifics around how participants were asked to complete it. For example, the study by Mintken and colleagues reports the following in the methods section: "*Patients rate their current level of pain and their worst and least amount of pain in the last 24 hours. The average of the 3 ratings or any single rating may be used to represent the patient's level of pain.*" [16] However, only one NPRS rating is reported in the study, and it is not clear which condition these scores refer to (i.e., current, worst in the last 24 hours, an average of three scores or a single score).

The different versions of the NPRS instrument used in each study were grouped into four categories based on the type of pain evaluated: background pain (i.e., pain at rest), evoked pain (i.e., pain with movement or physical testing), average pain (i.e., current pain, average pain, or the average of multiple scores), or not specified (i.e., no description regarding what type of pain participants were asked about) (Tables 3 and 4). Using these categories, two studies evaluated background pain, [13,20] four studies evaluated evoked pain, [13,20,22,23] four studies evaluated averaged pain, [14,15,20,21]and five studies did not specify pain context [12,16–19]. See Table 5 for further details.

Methodological quality

Only six of the 18 components of evidence assessed were judged to have good methodological quality (green) based on the OMERACT Good Methods Checklist (Supplementary Table 5) [15,16] These comprised one assessment of construct validity, [21] one of test-retest reliability, [21] two of longitudinal construct validity, [16,21] and two of thresholds of meaning [15,16]. The remaining components of evidence were rated as amber [15,16]. Two observational cohort studies were rated as amber for methodological quality within the clinical trial discrimination

Table 4
Summary of Findings of Measurement Properties (SOMP) for the Numeric Pain Rating Scale (NPRS). (References ([12–23]) is cited in Table body part).

	Instrument studied (& version): Numeric Pain Rating Scale (NPRS) **Background and target use**									
Target Domain:	Pain intensity						**Dates & initial assessment**			
	Definition: How much a person's shoulder hurts, reflecting the overall magnitude of the pain experience (i.e., at rest, during and after activity, at night)						**Working Group signoff:**		**Review Dates:**	
							Truth (domain match)	May 2018	**Start**	**End**
							Feasibility	May 2018	June 2023	February 2025
Target Population:			**Intervention:**				**Control:**		**Desired use:**	
Shoulder complaints			Any				Any		Change within clinical trials	
	Review findings*									
Source				Truth			**Discrimination**			
Author	**Year**	**NPRS brief description**	**NPRS Pain Category**	**Construct Validity**	**Inter-method reliability**	**Test- retest reliability**	**Longitudinal construct validity†**	**Clinical trial discrimination**	**Thresholds of meaning**	
Observational Studies										
Hummel-Berry[14]	2001	Pain at rest, Pain during motion	Background Evoked	±						
vanMeeteren[24]	2006	Pain on movement	Evoked				+			
Mintken[17]	2009	Not described	NS			±	-	-	+	
Michener[16]	2011	Pain at rest, Pain with ADLs, Pain with strenuous activity, NPRS average of above 3 items	Averaged (only averaged pain reported/used in analyses)				+	+	+	
Secondary Analyses (of Cohort Studies or Randomised Controlled Trials)										
Tubach[23]	2006	Pain on movement	Evoked						±	
O'Halloran[18]	2013	Not described	NS				±			
Riley[21]	2015	Average pain in past 24h, Present pain at rest, Pain with shoulder elevation, Pain with Hawkins-Kennedy test, Pain with Neer test	Averaged Background Evoked	-						
Riley[22]	2019	Pain at time of testing	Averaged	±		+				
Randomized Controlled Trials										
Atkinson[13]	2008	Not described	NS					-		
Rhon[20]	2014	Not described	NS					+		
Kardouni[15]	2015	Current pain (different wording pre and post intervention)	Averaged					+		
Ranagan[19]	2020	Not described	NS					±		
	Synthesis									
Studies per property (n)				3	0	2	4	4	3	
Synthesis statement per property‡	Synthesis pre property not possible due to heterogeneity in NPRS instruments									
We do not plan to seek OMERACT endorsement for this instrument for the pain domain.										

†Also called responsiveness or sensitivity to change.

Table 5
Variation of NPRS instruments across 12 included studies.

Type of Pain Assessed by Numeric Pain rating Scale (NPRS)	Number of Studies
Background Pain (pain at rest)	2
Evoked pain (pain with movement/testing)	4
Averaged pain (current or average pain, average of multiple scores)	4
Not specified	5

Note: some studies used multiple NPRS formats, therefore the number of studies > 12

component of the SOMP because they were not trials. The quality assessments for each study's methods for each individual measurement property are indicated by the colours in the Summary of Measurement Properties (SOMP), Table 4, which also summarises the adequacy of performance for each component of evidence.

Measurement Properties

Construct Validity

Of the three studies evaluating construct validity of the NPRS, [13,20,21] two reported equivocal findings (i.e., inconsistent evidence regarding adequacy of performance for a given measurement property) [13,21], one with an amber rating for quality [13] and the other with a green rating [21]. These studies investigated the correlation between NPRS instruments and the simple shoulder test (SST), the Medical Outcomes Study Short Form 36 (SF-36) Bodily Pain subscale, and the

Fear Avoidance Beliefs Questionnaire (FABQ) physical activity and work subscales. The remaining study published in 2015 was rated amber with inadequate performance (i.e., below the threshold of adequate performance) based on study results indicating weaker correlations between the NPRS and Shoulder Pain and Disability Index (SPADI) pain subscale than expected [20] Detailed information can be found in Supplementary Table 6 and 7.

Test-retest Reliability

Two studies evaluated test-retest reliability (Supplementary Table 8) [16,21] One study that did not specify the exact conditions under which pain was measured was rated amber and reported equivocal performance [16] The other measuring pain at the time of testing was rated green with adequate performance (i.e., above the threshold of adequate performance with intraclass correlation coefficients (ICCs) of $\geq$0.75) [21]

Longitudinal construct validity (Responsiveness)

Four studies evaluated longitudinal construct validity (Supplementary Table 9 and 10) [15–17,23] Of the four studies, two were rated green [15,16] and two were rated amber [17,23] The two studies rated green reported equivocal and inadequate findings of longitudinal construct validity respectively, [15,16] while the two studies rated amber had equivocal and adequate findings, respectively [17,23] The studies included for longitudinal construct validity assessed responsiveness using different methods including area under the curve calculations (n=2)[15,16], standardized response means over time (n=1) [23], and correlations between NPRS change scores and change scores for other instruments (n=3) including the Disabilities of the Arm, Shoulder, and Hand Questionnaire Short Form (QuickDASH), Shoulder Disability Questionnaire (Dutch Version), and a Global Rating of Change (GRC) measure [16,17,23]

Clinical trial discrimination

Four RCTs were used to evaluate clinical trial discrimination [12,14, 18,19] All were rated amber for methodological quality. Two were deemed to have adequate performance (no difference between groups was hypothesized for NPRS scores, and no significant difference between groups was reported), [14,19] and the two other studies reported equivocal and negative findings, respectively due to some level of between-group difference being reported when one was not expected (Supplementary Tables 11,12) [12,18]

There were no studies in which a significant difference in NPRS scores between the tested interventions was expected or reported. We therefore considered two additional longitudinal studies as silver evidence for clinical trial discrimination because these studies reported differences between stable/not improved and improved groups of participants [15,16] In one of the studies, negative evidence was reported, as the effect size between stable and improved groups was smaller than hypothesized [16] In the other study adequate clinical trial discrimination (+) was reported, as the hypothesized large effect size between improved and not improved groups was reported [15]

Thresholds of meaning

In three studies evaluating thresholds of meaning, two green-rated studies reported positive findings,[15,16] and one amber-rated study reported equivocal findings [22] The thresholds of meaning that were calculated differed across studies and included the minimal clinically important difference (MCID, n=2),[15,16] threshold based on area under the curve calculations (n=1),[15] sensitivity (n=1),[15] specificity (n=1),[15] minimal detectable change (MDC, n=1),[16] patient acceptable symptom state (PASS, n=1),[22] and the minimal clinically important improvement (MCII, n=1) [22] (Supplementary Table 13)

Synthesis and consensus

We found insufficient evidence of the measurement properties for any single NPRS. With guidance from the TAG, we therefore asked attendees at the 2025 OMERACT SIG the following: "We synthesized all NPRSs regardless of context (type of pain, anchor statements, timeframe, etc.). Do you agree this is appropriate?". Sixteen of the 19 session attendees (including methodologists, clinicians, and patient partners) voted. Twelve (75%) responded "No" and four (25%) responded "Yes". Based on the majority response, a synthesis statement of the measurement properties of the NPRS across all studies was not completed (Table 4).

We also considered whether it would be possible to synthesize the evidence for subsets of NPRSs studies that had used the same pain description (i.e., background, evoked, averaged, not specified). However, most attendees also considered this inappropriate due to the heterogeneity even within these subsets in the prompt wording and anchors, as well the different definitions and perceptions of evoked, averaged and background pain studies. Therefore, no synthesis statement of measurement properties for any NPRS version was completed.

Discussion

The purpose of this study was to use the OMERACT methods to assess the NPRS through Filter 2.2 to understand whether it could be endorsed as part of a core outcome set for assessing pain intensity in individuals with shoulder disorders. Our systematic review identified 12 studies comprising 18 components of evidence across five measurement properties. Current findings from individual studies suggest that the NPRS has adequate performance for longitudinal construct validity and thresholds of meaning (studies rated green and (+)). However, caution should be used when interpreting these findings due to the heterogeneity across NPRS instruments. Additional high-quality research is required to confirm the construct validity, test-retest reliability, longitudinal construct validity and clinical trial discrimination of the NPRS.

Arguably the most important finding of this study was the significant variation in format of the NPRS scales assessed across studies and whether these should be considered different instruments. We found insufficient evidence to appropriately assess any single version's suitability for inclusion as a measure of pain in the core outcome set. No two studies assessing the same measurement property used the same NPRS instrument. Even when studies used NPRS assessing the same category of pain (i.e., evoked pain), the wording, scale anchors and description differed. This heterogeneity limited our ability to pool data across studies or to draw any firm conclusions about the measurement properties of the NPRS for patients with shoulder disorders.

The complexity of assessing pain with a single item scale was further illustrated by patient research partner comments during the OMERACT Shoulder Special Interest Group wherein patient partners described conflicting definitions of pain, resting pain and average pain and conflicting interpretations of NPRS question formats. A systematic review of unidimensional pain scales in varying patient populations reported similar findings to our study [24] A total of 41 different NPRS instruments were used across 37 studies. Fifteen different anchor descriptors were reported including options like no pain/worst pain, mild/excruciating pain, or not specified. Like our study, variability in the wording of scale anchors was also observed and there were eight different sized scales ranging from 6- to 101-point scales with the most common being the 11-point (0 to 10) and 101 (0 to 100) point scales.

In the absence of a recommended instrument to measure pain in shoulder trials, trialists should consider what specific components of pain and the pain experience are most meaningful to the type of patients that would be candidates for their trial. Ideally there should also be good evidence in support of the measurement properties of any chosen

instrument. If trialists decide on an NPRS, our review has highlighted that there are many different NPRSs that could be used. Consideration should be given to the best description of the pain for the research question and trial population, time frame and anchors, as well as the available evidence of their measurement properties. They could also consider how their trial could add to this evidence base.

This study has important implications for next steps in instrument selection for the pain domain in our core outcome set. From our work, we hypothesize that similar single-item instruments measuring pain, such as the pain verbal rating scale (VRS) and pain visual analog scale (VAS) may have similar limitations as the NPRS. The format and reporting of these scales are inconsistent, [24] and it is likely that similar issues would arise. We will therefore focus on assessing the suitability of multi-item pain scales such as the subscale of the Shoulder Pain and Disability Index and the Shoulder Pain Score as they may be reported more consistently across the literature.

A key strength of our study is the use of the OMERACT Filter 2.2, which provided a systematic and rigorous framework for evaluating outcome measures. We also employed robust methods, including a comprehensive literature search, independent screening and data assessment, and a collaborative review process involving clinical experts, methodologists, and patient partners. Our search was conducted in June 2023 and has not been updated since. However additional evidence would most likely not alter our conclusions of the NPRS at this time and may in fact add to our results of varying reported formats of the NPRS.

Conclusion

The NPRS cannot progress to the endorsement stage of instrument selection for a core outcome set for clinical trials for shoulder conditions due to heterogeneity amongst NPRS versions used within primary studies and insufficient evidence to properly synthesize results for a single NPRS version. Each version should be considered a distinct instrument, and current evidence is insufficient to support its use in clinical trials.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Romi Haas reports travel was provided by OMERACT. Pam Richards reports travel was provided by OMERACT. Dorcas Beaton reports travel was provided by OMERACT. Dorcas Beaton reports a relationship with OMERACT that includes: board membership. Sofia Ramiro reports a relationship with AbbVie Inc that includes: consulting or advisory, funding grants, speaking and lecture fees, and travel reimbursement. Sofia Ramiro reports a relationship with Alfasigma SpA that includes: consulting or advisory and funding grants. Sofia Ramiro reports a relationship with Eli Lilly and Company that includes: consulting or advisory, funding grants, and travel reimbursement. Sofia Ramiro reports a relationship with Johnson & Johnson MedTech that includes: consulting or advisory. Sofia Ramiro reports a relationship with Merck & Co Inc that includes: consulting or advisory. Sofia Ramiro reports a relationship with Novartis Pharmaceuticals Corporation that includes: consulting or advisory and funding grants. Sofia Ramiro reports a relationship with Pfizer Inc that includes: consulting or advisory, funding grants, and speaking and lecture fees. Sofia Ramiro reports a relationship with UCB Pharma SA that includes: consulting or advisory, funding grants, and travel reimbursement. Co-author Sofia Ramiro is a member of the Executive Committee of the Assessment of Spondyloarthritis International Society (ASAS) and a member of the EULAR Quality of Care Committee. No other authors declare that they have known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.semarthrit.2026.152954.

References

[1] Ramiro S, Page MJ, Whittle SL, Huang H, Verhagen AP, Beaton DE, Buchbinder R. The OMERACT core domain set for clinical trials of shoulder disorders. The Journal of rheumatology 2019;46(8):969–75.

[2] Buchbinder R, Page MJ, Huang H, Verhagen AP, Beaton D, Kopkow C, Gagnier JJ. A preliminary core domain set for clinical trials of shoulder disorders: a report from the OMERACT 2016 shoulder core outcome set special interest group. The Journal of rheumatology 2017;44(12):1880–3.

[3] Page MJ, Huang H, Verhagen AP, Buchbinder R, Gagnier JJ. Identifying a core set of outcome domains to measure in clinical trials for shoulder disorders: a modified Delphi study. RMD open 2016;2(2):e000380.

[4] Page MJ, O'Connor DA, Malek M, Haas R, Beaton D, Huang H, Ramiro S, Richards P, Voshaar MJH, Shea B, Verhagen AP, Whittle SL, van der Windt DA, Gagnier JJ, Buchbinder R, Shoulder Core Set Working Group OMERACT. Patients' experience of shoulder disorders: a systematic review of qualitative studies for the OMERACT Shoulder Core Domain Set. Rheumatology (Oxford, England) 2019.

[5] Beaton D, Maxwell L, Grosskleg S, et al. Striving to improve endpoint outcome measurement through a data driven, iterative consensus process involving relevant stakeholder groups.

[6] Downie WW, Leatham PA, Rhind VM, Wright V, Branco JA, Anderson JA. Studies with pain rating scales. Annals of the rheumatic diseases 1978;37(4):378–81.

[7] Young I, Dunning J, Escaloni J, Maselli F, Prall J, Mourad F, Fernández-de-Las-Peñas C. Reliability, construct validity, responsiveness and minimum clinically important difference of the numeric pain rating scale and shoulder pain and disability index in patients with subacromial pain syndrome. Musculoskeletal Science and Practice 2025:103372.

[8] Furtado R, Bobos P, Ziebart C, Vincent J, MacDermid J. Patient-reported outcome measures used for shoulder disorders: an overview of systematic reviews. Journal of Hand Therapy 2022;35(2):174–85.

[9] Elsman EB, Mokkink LB, Terwee CB, et al. Guideline for reporting systematic reviews of outcome measurement instruments (OMIs): PRISMA-COSMIN for OMIs 2024. Journal of Clinical Epidemiology 2024;173:111422.

[10] Buchbinder R, Ramiro S, Huang H, Gagnier JJ, Jia Y, Whittle SL. Measures of adult shoulder function 2020.

[11] Boers M, Kirwan JR, Wells G, Beaton D, Gossec L, d'Agostino MA, Tugwell P. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. Journal of clinical epidemiology 2014;67(7):745–53.

[12] Atkinson M, Mathews R, Brantingham JW, Globe G, Cassa T, Bonnefin D, Korporaal C. A randomized controlled trial to assess the efficacy of shoulder manipulation vs. placebo in the treatment of shoulder pain due to rotator cuff tendinopathy. Journal of the American Chiropractic Association 2008;45(9).

[13] Hummel-Berry KH. Assessment of shoulder function and functional impact of clinic physical therapy versus home exercise for patients with shoulder stiffness: A randomized controlled trial. University of Washington; 2001.

[14] Kardouni JR, Shaffer SW, Pidcoe PE, Finucane SD, Cheatham SA, Michener LA. Immediate changes in pressure pain sensitivity after thoracic spinal manipulative therapy in patients with subacromial impingement syndrome: a randomized controlled study. Manual therapy 2015;20(4):540–6.

[15] Michener LA, Snyder AR, Leggin BG. Responsiveness of the numeric pain rating scale in patients with shoulder pain and the effect of surgical status. Journal of sport rehabilitation 2011;20(1):115–28.

[16] Mintken PE, Glynn P, Cleland JA. Psychometric properties of the shortened disabilities of the Arm, Shoulder, and Hand Questionnaire (QuickDASH) and Numeric Pain Rating Scale in patients with shoulder pain. Journal of Shoulder and Elbow Surgery 2009;18(6):920–6.

[17] O'Halloran B, Wright A, Cook CE. Criterion validation of the rate of recovery, a single alphanumeric measure, in patients with shoulder pain. International Journal of Sports Physical Therapy 2013;8(6):784.

[18] Rangan A, Brealey SD, Keding A, Corbacho B, Northgraves M, Kottam L, Venateswaran B. Management of adults with primary frozen shoulder in secondary care (UK FROST): a multicentre, pragmatic, three-arm, superiority randomised clinical trial. The Lancet 2020;396(10256):977–89.

[19] Rhon DI, Boyles RB, Cleland JA. One-year outcome of subacromial corticosteroid injection compared with manual physical therapy for the management of the unilateral shoulder impingement syndrome: a pragmatic randomized trial. Annals of internal medicine 2014;161(3):161–9.

[20] Riley SP, Bialosky J, Cote MP, Swanson BT, Tafuto V, Sizer PS, Brismée JM. Thoracic spinal manipulation for musculoskeletal shoulder pain: can an instructional set change patient expectation and outcome? Manual therapy 2015;20(3):469–74.

[21] Riley SP, Tafuto V, Cote M, Brismée JM, Wright A, Cook C. Reliability and relationship of the fear-avoidance beliefs questionnaire with the shoulder pain and disability index and numeric pain rating scale in patients with shoulder pain. Physiotherapy Theory and Practice 2019;35(5):464–70.

[22] Tubach F, Dougados M, Falissard B, Baron G, Logeart I, Ravaud P. Feeling good rather than feeling better matters more to patients. Arthritis Care & Research: Official Journal of the American College of Rheumatology 2006;55(4):526–30.
[23] van Meeteren J, Roebroeck ME, Selles RW, Stam HJ. Responsiveness of isokinetic dynamometry parameters, pain and activity level scores to evaluate changes in patients with capsulitis of the shoulder. Clinical Rehabilitation 2006;20(6):496–501. https://doi.org/10.1191/0269215506cr983oa.
[24] Hjermstad MJ, Fayers PM, Haugen DF, Caraceni A, Hanks GW, Loge JH, Fainsinger R, Aass N, Kaasa S. European Palliative Care Research Collaborative (EPCRC) (2011). Studies comparing Numerical Rating Scales, Verbal Rating Scales, and Visual Analogue Scales for assessment of pain intensity in adults: a systematic literature review. Journal of pain and symptom management 2010;41(6):1073–93. https://doi.org/10.1016/j.jpainsymman.2010.08.016.