Seminars in Arthritis and Rheumatism 79 (2026) 152998

Contents lists available at ScienceDirect

Seminars in Arthritis and Rheumatism

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

Suitability of the Oxford Shoulder Score for measuring pain in clinical trials evaluating interventions for people with shoulder disorders according to the OMERACT filter 2.2

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

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

ARTICLE INFO

Keywords:
Oxford Shoulder Score
OMERACT
Shoulder
Outcome measures
Systematic literature review

ABSTRACT

Background: Pain is a mandatory domain in Outcome Measures in Rheumatology (OMERACT) core outcome sets for shoulder disorder trials. We evaluated the Oxford Shoulder Score (OSS), a composite measure comprising four pain and eight function items but no separate subscales for these domains, for measuring pain using the OMERACT Filter 2.2.

Methods: Following the OMERACT Handbook, we assessed domain match and feasibility, then systematically reviewed OSS measurement properties in shoulder disorders (rotator cuff disease, adhesive capsulitis, instability, osteoarthritis, dislocation, humeral head fractures and unspecified pain). MEDLINE, EMBASE and CINAHL were searched to June 2023. Reviewers independently screened, appraised methodological quality and extracted data. Measurement properties were synthesised and rated (green, amber, red or white). Results were summarised in a Summary of Measurement Properties (SOMP) table and discussed at the OMERACT 2025 workshop.

Results: The OSS was rated amber for feasibility and domain match, reflecting concerns about its multidimensional nature while acknowledging interrelatedness of pain and function. Twenty-three studies were included in the systematic review: eleven examined construct validity, three test-retest reliability, ten responsiveness, five clinical trial discrimination and five thresholds of meaning. Thirty of 34 components were judged suitable to proceed (eight green; 22 amber). All studies assessed the OSS total score (pain and function); one assessed a 4-item pain subscale. For the total OSS score, construct validity and responsiveness were green and reliability, trial discrimination and thresholds of meaning were amber. Two studies reported no floor/ceiling effects, one was equivocal. At OMERACT 2025, 95% of respondents agreed multidimensional instruments should not be used to measure single domains without validated subscales.

Conclusion: The OSS total score shows adequate construct validity and responsiveness, but its combined assessment of pain and function makes it unsuitable for measuring pain alone.

* Corresponding author at: Musculoskeletal Health and Wiser Health Care Units, School of Public Health and Preventive Medicine, Monash University, 553St Kilda Road, Melbourne, Victoria, 3004, Australia.
E-mail address: romi.haas@monash.edu (R. Haas).

https://doi.org/10.1016/j.semarthrit.2026.152998

Available online 9 May 2026
0049-0172/© 2026 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Introduction

There is a lack of uniformity in the outcomes measured in clinical trials of people with shoulder disorders. A systematic review of 409 such trials identified 319 different instruments used to assess 32 distinct domains[1]. This variability limits our ability to pool data and compare findings across different trials assessing the same interventions. It also increases the risk of selective outcome reporting and contributes to research waste.

In response to this issue, the Outcome Measures in Rheumatology (OMERACT) Shoulder Working Group developed a core domain set for use in all clinical trials evaluating interventions for shoulder disorders (defined as including tendinopathy, impingement, subacromial bursitis, rotator cuff tears, adhesive capsulitis, instability, glenohumeral osteoarthritis, dislocation, proximal humeral or humeral head fractures, and unspecified shoulder pain)[2]. A preliminary core domain set[3] was developed based upon a systematic review of domains measured in shoulder trials[1], an international Delphi study involving patients, clinicians and researchers[4] and a qualitative synthesis of the lived experience of people with shoulder disorders[5] and the final core domain set was endorsed by OMERACT in 2018[2]. Mandatory core domains include pain, physical function, patient global-shoulder and adverse events.

Based upon earlier work[1], we identified 38 candidate instruments that measure pain in people with shoulder disorders. The next step in the OMERACT process is to determine which of these instruments meet the OMERACT Filter 2.2 criteria for truth, discrimination and feasibility.

The aim of this study was to use the OMERACT Filter 2.2 to assess the suitability of the Oxford Shoulder Score (OSS) as part of a core measurement set for assessing pain in clinical trials involving people with shoulder disorders. The OSS is a patient-reported composite (multidimensional) measure that includes items assessing both pain and physical function with no separate subscales for these domains. Originally developed to assess surgical outcomes for shoulder disorders[6], the OSS is widely used in both surgical and non-surgical trials[7] and is the most used shoulder-specific instrument in arthroplasty registries[8].

Methods

We followed the methods outlined in the OMERACT Handbook on Instrument Selection for Core Outcome Measurement Sets, which are based on the three pillars of truth, feasibility and discrimination[9]. This involved obtaining agreement from the working group that the instrument matched the target domain and was feasible, followed by a systematic review synthesising the measurement properties of the candidate instrument. The handbook's traffic light rating was applied at key decision points: Green indicated good to proceed, Amber signalled proceed with caution due to some concerns or limitations and Red indicated the need to stop (Fig. 1).

Oxford shoulder score

The OSS assesses shoulder problems over the last four weeks and includes four items related to pain (worst pain intensity, usual pain intensity, pain interference with usual work, and pain at night) and eight items related to daily function (dressing, use of knife and fork, brushing/combing hair, car transfer, shopping, carrying a tray across the room, bathing and dressing) (Supplementary Figure 1)[6]. Each item is rated on a 5-point Likert scale. In the original version of the instrument, items were scored from 1 (best) to 5 (worst), with total scores ranging from 12 to 60 (lower scores = better outcomes). In 2009, the direction of the scores were reversed (each item scored from 0 (worst) to 4 (best)), with total scores ranging from 0 to 48 (higher scores = better outcomes)[10]. Originally developed as a composite measure, it does not include validated subscales for pain and function.

Assessment of domain match and feasibility

We assessed the domain match and feasibility of the OSS during an online meeting of the Working Group in 2018. We assessed whether the OSS matched the domain definitions of pain, considering relevance, completeness, redundancy and clarity of the items, suitability of the response options, and appropriateness of the scoring method. Pain, reflecting pain intensity, was defined as the extent of shoulder pain experienced, including pain at rest, during and after activity, and at night[2].

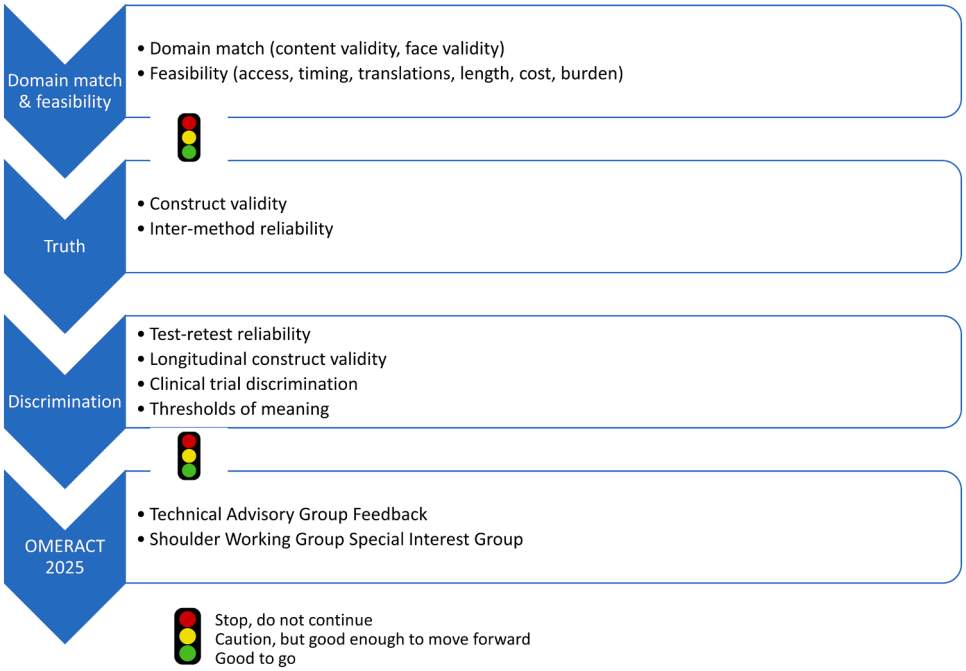

Fig. 1. Instrument selection process for assessing the suitability of the Oxford Shoulder Score to measure pain intensity in clinical trials of shoulder disorders.

For feasibility, we considered 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 the OSS would only proceed to the next step if it received a green or amber rating.

Systematic review of measurement properties

We conducted a systematic review of the available evidence on the measurement properties of the OSS for measuring pain in people with shoulder disorders. The study protocol was registered on PROSPERO CRD42023440148 and reported according to the PRISMA-COSMIN guidance for reporting systematic reviews of outcome measurement instruments[11].

Selection criteria and search methods

Table 1 provides more details about our inclusion and exclusion of studies based upon study population. We also excluded studies that evaluated the Oxford Shoulder Instability Score. Measurement properties assessed included floor and ceiling effects, construct validity, test-retest reliability, inter-method reliability, inter-rater reliability, responsiveness (longitudinal construct validity), clinical trial discrimination and thresholds of meaning.

We searched Medline via Ovid, Embase via Ovid, and Cumulative Index of Nursing and Allied Health (CINAHL) via Elton B. Stephens Company (EBSCO) from inception to 5 June 2023 with no restrictions on language or publication date. We used the Boolean operator "AND" to combine search terms relating to shoulder disorders, instruments and studies of measurement properties (Supplementary Table 1). A valid and sensitive search filter to identify studies of measurement properties was used in Medline and adapted for the other databases[12]. We identified randomised controlled trials (RCTs) that had included the OSS as an outcome from a systematic review of trials for shoulder disorders[1] and a more recent review of adult shoulder function measures[7]. Reference lists of included studies and relevant review articles were also screened.

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	• Oxford Shoulder Score (English version)	• Oxford Shoulder Instability Score
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 data were not available to assess these during the domain match stage.

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). Discrepancies were resolved through discussion with additional members of the investigator team (SR, SW, 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 SOMP table[13].

Methodological quality appraisal

Pairs of reviewers independently assessed the methodological quality of each study using the COSMIN—OMERACT Good Methods Checklist[9]. For each study evaluating one or more components of measurement property evidence reviewers completed the relevant checklist items and provided comments (Supplementary Table 2). 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).

Data extraction on measurement properties

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 initially performed independently by pairs of reviewers for two studies per measurement property. 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 3. For construct validity and responsiveness (longitudinal construct validity), observed correlations between the OSS and other instruments were compared with a priori hypotheses. For responsiveness, anticipated standardised response means (SRMs) or effect sizes over time were compared with a priori hypotheses. In situations where both SRMs and effect sizes were provided, SRMs were prioritised because they account for the variability in change scores within the sample, providing a more precise estimate of responsiveness and a better reflection of the change likely to occur in clinical trials. For clinical trial discrimination, between-group effect sizes were evaluated against expected differences. Where there were no explicit a priori hypotheses, we formulated hypotheses, based on prior literature where possible. Our a priori hypotheses for expected correlations between the OSS and other instruments, SRMs or effect sizes over time and between-group effect sizes in clinical trials are shown in Supplementary Tables 4, 5 and 6 respectively.

Adequacy of performance

For each included study, pairs of reviewers 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.

Evidence synthesis for each measurement property

The evidence supporting each measurement property was presented in the SOMP table[13]. The rating for each measurement property was synthesised in accordance with the OMERACT guide for synthesis ratings for each measurement property (Supplementary Table 7) considering the quality (rated as Green/Amber or Red), quantity, adequacy of performance (rated as +, ± or -) and the consistency 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 no evidence was available. Amber was assigned to all other situations.

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 session, online polling was conducted to address two specific questions: (a) whether participants agreed with our ratings for the synthesis of each measurement property? and (b) whether it is appropriate for multidimensional instruments to be assessed through the OMERACT Filter for a single domain (e.g., pain) when the instrument does not have a validated subscale (e.g., the OSS, which measures pain and function in one total score)?

Results

Domain match and feasibility

The Shoulder Working Group consensus on both the domain match and feasibility of the OSS for pain was Amber (Supplementary Table 8). For domain match, some concerns were raised about potential item redundancy, the appropriateness of the response options and the scoring approach. Although the multidimensional nature of the OSS was noted as a limitation, Patient Research Partners (PRPs) considered the items to be a good match to the pain domain, highlighting the interrelatedness of pain and function in everyday experience. They also reported that the questions were clear and easy to complete. For feasibility, the amber rating reflected concerns regarding copyright restrictions, including whether permission is required for use, as well as concerns about the availability of the OSS in relevant languages and cultural contexts.

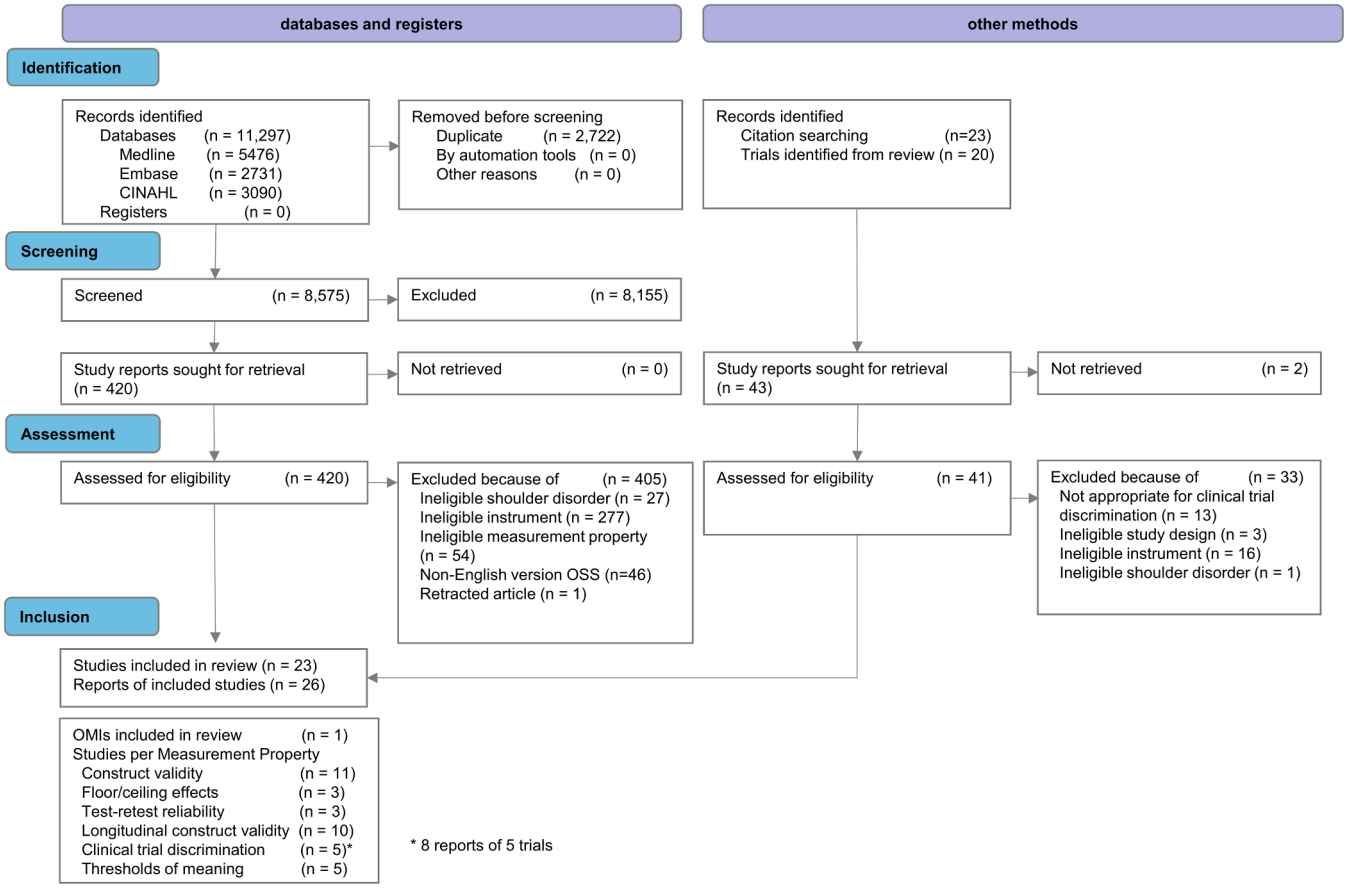

Fig. 2. PRISMA-COSMIN for Outcome Measurement Instruments (OMIs) flow diagram displaying the number of studies identified and included from databases and other sources. Abbreviations and footnote for Fig. 2:
PRISMA; Preferred Reporting Items for Systematic Reviews and Meta-Analyses:
COSMIN; COnsensus-based Standards for the selection of health Measurement INstruments:
OMI; Outcome Measurement Instrument
23 included studies comprise 15 identified through databases and registers and 8 identified through other methods.

Systematic review

Search results and description of included studies

After exclusion of duplicates, the search yielded 8575 studies; 463 full-text reports were assessed for eligibility and 23 studies (26 reports) were included (Fig. 2)[6,14–38]. Eleven studies examined construct validity[6,14–16,19,20,22,25,29,30,35], three floor and ceiling effects [23,33,34], three test-retest reliability[6,20,35], ten responsiveness[6, 15,16,20,23,25,29,31,34,35], five (eight reports) clinical trial discrimination[17,18,21,24,26–28,32], and five thresholds of meaning[34–38]. All 23 studies assessed the measurement properties of the OSS total score while one also assessed the measurement properties of a subscale comprising the four pain-specific items[19].

Setting and target populations

The included studies were published between 1996 and 2023, with nearly half (n = 10, 43 %) published between 2010 and 2020[16–18,23–26,29,31,32,35,36] and four (17 %) published after 2020[27,28,33,34,38] (Table 2). Most were observational in design (n = 18, 78 %), of which 17 were longitudinal [6,14–16,19,20,22,23,25,29,31, 33–38] and one was cross-sectional[30]. The remaining five studies (22 %) were RCTs[17,18,21,24,26–28,32]. Most studies were conducted in the United Kingdom (UK) (n = 16, 70 %)[6,14–24,26,29–33], followed by Singapore (n = 3, 13%)[36–38], with single studies performed in Australia[35], New Zealand[25], Denmark[34] and across both Norway and Sweden[27,28]. Most studies were hospital-based (n = 21, 91 %)[6, 14–20,22–25,27–38] and two (9 %) were conducted in primary care[21, 26].

Twelve studies included people with rotator cuff disorders of whom seven were undergoing surgery [14,22,25,34,36–38] and five were being treated nonoperatively [16–18,20,21,27,28]. Four studies included people with proximal humeral fractures, including one managed nonoperatively [15], one operatively [19] and two either operatively or nonoperatively [24,32,35]. One study included people with adhesive capsulitis being treated with manipulation under anaesthesia [30]. Six studies included mixed populations including five undergoing surgery [6,23,29,31,33] and one being managed nonoperatively [26].

The median (IQR) sample size of the studies was 111 (60–235) with most (n = 14, 61 %) including more than 100 participants[6,14,15,17, 18,20,21,23,24,27,28,31–33,36–38]. Among the longitudinal studies, the follow-up duration ranged from 12 weeks to 5 years. Mean age ranged from 48 to 79 years, and most studies (n = 14, 61 %) had more female than male participants[14,15,17,18,20,21,23–28,32,33,35–37].

Most studies (n = 12, 52 %) used the revised Oxford scoring system (0–48), where higher scores indicate better outcomes[16–18,23–28, 31–36]. Six studies (26 %) used the original scoring system (12–60), where higher scores indicate worse outcomes[6,15,19,21,22,30]. Two studies (9 %) reported transformed scores out of 100, one where higher scores reflect better outcomes[14] and one where higher scores reflect worse outcomes[20]. Three studies (13 %) did not specify the scoring system used[29,37,38].

Methodological quality of included studies

Of the 34 components of evidence appraised, only eight were rated as green, 22 as amber and four as red, including two components each for construct validity and responsiveness (Supplementary Table 9). The most common reasons for red quality ratings were high rates of attrition and inappropriate statistical measures.

Performance and synthesis ratings for each measurement property

The SOMP table (Table 3) summarises the evidence supporting each measurement property including the quality rating and adequacy of performance for each component of evidence and the final synthesis rating.

Construct validity

Of the nine studies with green or amber quality ratings for construct validity[6,15,16,19,20,22,25,29,35], seven were judged to have adequate performance[6,15,20,22,25,29,35], including two rated green [6,35] (Supplementary Table 10). These studies reported strong correlations with other composite measures of shoulder pain and function, including the Constant-Murley score, Shoulder Pain and Disability Index (SPADI), University of California - Los Angeles (UCLA) shoulder scale, American Shoulder and Elbow Surgeons (ASES) score, Disabilities of the Arm, Shoulder and Hand (DASH) Score and the Stanmore Percentage of Normal Shoulder Assessment (SPONSA).

A single study, rated amber, showed better OSS scores in healthy participants versus those with shoulder pain, supporting adequate construct validity using known-groups[16] (Supplementary Table 11). However, this study also reported inconsistent correlations with inertial metrics and was rated as having equivocal construct validity.

Only one study, rated amber, examined the association between the OSS pain subscale and other pain-specific instruments such as the pain Visual Analogue Scale (VAS)[19]. As strong correlations were anticipated, the observation of moderate correlations resulted in the OSS pain subscale being rating as having inadequate construct validity. In contrast, this same study demonstrated adequate construct validity of the OSS total score.

The synthesis rating for construct validity of the total OSS score was green, based on evidence from two good quality studies with adequate performance and sufficient consistency across findings [6,35].

Floor and ceiling effects

Three studies evaluated floor and ceiling effects of the total OSS score [23,33,34] (Supplementary Table 12). Two studies found no evidence of these effects and were rated as having adequate performance[23,34]. The third study, rated equivocal (±), reported no floor effects and no ceiling effects pre-operatively or at six months post-arthroplasty, but observed ceiling effects at three and five years post-operatively[33]. We did not report a synthesis rating for floor and ceiling effects because the COSMIN—OMERACT Good Methods Check does not provide guidance for appraising the methodological quality of studies evaluating these effects.

Test-retest reliability

Three studies evaluated test-retest reliability of the total OSS score[6, 20,35] (Supplementary Table 13). One study was rated green[35] and two as amber for methodological quality[6,20]. All studies were deemed to have adequate performance, with intraclass correlation coefficients (ICCs) of $\geq$0.75, although one study had a wide confidence interval[35]. Based on these findings, the synthesis rating was amber for test-retest reliability.

Responsiveness (Longitudinal construct validity)

Of the eight studies with green or amber quality ratings for responsiveness[6,16,23,25,29,31,34,35], six were judged to have adequate performance[6,16,25,29,34,35], including two rated green[34,35] (Supplementary Table 14). The remaining two studies, both rated amber for quality, were considered to have equivocal performance as they met only 25–75 % of the anticipated post-operative standardised response means (SRMs)[23,31].

The synthesis rating for responsiveness of the total OSS score was green, based on evidence from two good quality studies with adequate performance and sufficient consistency across findings. This included

Table 2
Characteristics of included studies.[a]

Observational studies												
Author	Year	Country	Design	Population	Setting	Measurement properties	Sample size	Follow-up (final)	Age (years) Mean (SD)	Male (%)	Baseline OSS Mean (SD)	Additional Instruments
Allom[14]	2009	UK	Longitudinal	Surgery for RCD	OPC	Construct	372	2 years	56 (range 19–87)	45	95 % CI 45.5 to 50.3 Out of 100, higher score=best outcome	Constant DASH
Baker[15]	2008	UK	Longitudinal	Proximal humerus fracture	OPC	Construct Responsiveness	103	12 mths	61.4 (range 15–85)	29	27.3 (12.7, range 12–57) 12–60 higher score=worse outcome	Constant
Bavan[16]	2020	UK	Longitudinal	Subacromial shoulder pain	Hospital PT Dept	Construct Responsiveness	25 Patients 50 Controls	92.0 (17.3) days	58.9 (10.5) 48.4 (21.2)	48 58	33.9 (7.15) 0–48 higher score=best outcome	Inertial metrics
Beastall[19]	2012	UK	Longitudinal	Proximal humerus fracture (internal fixation)	OPC	Construct	44	23 (IQR 16.25, 35) mths	58.4 (13.9)	66	19 (IQR 15, 26.75) 12–60 higher score=worse outcome	Constant QuickDASH UCLA
Cloke[20]	2005	UK	Longitudinal	Subacromial impingement <6 mths	Specialist clinic	Construct Test-retest Responsiveness	110	24 weeks	55 (range 24–89)	44	38.7 (20.4) Out of 100, higher score=worse outcome	SF-36 SPADI
Dawson[6]	1996	UK	Longitudinal	Shoulder surgery (not instability)	Pre-surgical clinic	Construct Test-retest Responsiveness	111	6 mths	median 57.4 (15.0, range 19.9 to 87.5)	54	36.3 (95 % CI 34.6 to 37.9) 12–60 higher score=worse outcome	Constant HAQ SF-36
Dawson[22]	2001	UK	Longitudinal	Subacromial impingement (surgical)	Surgical clinic	Construct	93	3.9 (2.2 to 5.6) years	57.8 (22.3 to 83.7)	66	33.0 (95 % CI 31.3 to 34.8) 12–60 higher score=worse outcome	Constant SF-36
Evans[23]	2018	UK	Longitudinal	Shoulder surgery (not instability)	Surgical clinic	Floor/ceiling Responsiveness	204	1 year	60.3	39	19.64 (6.52) 0–48 higher score=best outcome	EQ-5D-3L
Hapuarachchi [25]	2014	New Zealand	Longitudinal	Reverse shoulder arthroplasty for rotator cuff tear	Database	Construct Responsiveness	29	1 year	78.9 (6.0 range 63.3–87.6)	28	18.3 (6.8 range 4 to 32) 0–48 higher score=best outcome	ASES
Noorani[29]	2012	UK	Longitudinal	Shoulder surgery	Pre-surgical clinic	Construct Responsiveness	61	3–6 mths	NR	NR	NR	Constant SPONSA
Othman[30]	2004	UK	Cross-sectional	MUA for frozen shoulder	Surgical clinic	Construct	51 (60 shoulders)	33 (6–88) mths	53 (33–83)	52	15 (12–24) 12–60 higher score=worse outcome	Constant
Price[31]	2019	UK	Longitudinal	Shoulder surgery	Pre-surgical clinic	Responsiveness	594	6 mths	53.0 (16.7)	55	29.6 (10.3) 0–48 higher score=best outcome	EQ-5D-5L MSK-HQ

(*continued on next page*)

Table 2 (*continued*)

Observational studies												
Author	Year	Country	Design	Population	Setting	Measurement properties	Sample size	Follow-up (final)	Age (years) Mean (SD)	Male (%)	Baseline OSS Mean (SD)	Additional Instruments
Singh[33]	2021	UK	Longitudinal	Shoulder arthroplasty	Database	Floor/ceiling	46,824 separate observations	5 years	75.0 (7.0)	28	15.7 (8.6) 0–48 higher score=best outcome	-
Sorenson[34]	2021	Denmark	Longitudinal	Decompression for subacromial impingement	Pre-surgical clinic	Responsiveness Thresholds	58	6 mths	57.4 (10.1)	50	29.4 (6.6) 0–48 higher score=best outcome	EQ-5D-5L Pain VAS SSV
Van de Water [35]	2014	Australia	Longitudinal	Proximal humerus fracture	2 hospital clinics	Construct Test-retest Responsiveness Thresholds	22	13 weeks	68.1 (11.3)	20	NR 0–48 higher score=best outcome	Constant DASH FABQ-PA UCLA SSV
Xu[36]	2019	Singapore	Longitudinal	Rotator cuff repair for supraspinatus tear	Database	Thresholds	214	2 years	60.1 (10.0)	45	29 (11) 0–48 higher score=best outcome	Constant UCLA
Xu[37]	2020	Singapore	Longitudinal	Rotator cuff repair for supraspinatus tear	Database	Thresholds	306	2 years	60.2 (10.3)	45	28.1 (11.2) NR, assume as above given same authors	Constant UCLA
Zhou[38]	2023	Singapore	Longitudinal	Reverse shoulder arthroplasty for massive irreparable cuff tear	Database	Thresholds	131	5 years	75.0 (7.0)	66	39.1 (12.5) NR	Constant UCLA
Randomised controlled trials												
Author	Year	Country	Design	Population	Setting	Intervention arms	Sample size	Follow-up (final)	Age Mean (SD)	Male (%)	Baseline OSS	Instruments
Beard[17,18]	2018	UK	3-arm	Subacromial pain	32 hospitals	Decompression Arthroscopy No treatment	106 103 104	1 year	52.9 (10.3) 53.7 (10.5) 53.2 (10.2)	49 50 50	25.2 (9.5) 26.7 (8.8) 25.5 (8.3) 0–48 higher score=best outcome	Modified Constant HADS PainDETECT
Cloke[21]	2008	UK	Pilot 4-arm	Painful arc <6 months	Primary care	Control Physiotherapy Steroid injection PT+steroid	27 29 28 29	18 weeks	54.5 (range 23–88)	43	31.35 29.95 28.85 26.35 12–60 higher score=worse outcome	SF-36
Holt[26]	2013	UK	Pilot 2-arm	Rotator cuff tendinopathy or adhesive capsulitis	6 primary care practices	Steroid+lidnocaine Lidnocaine	19 21	12 weeks	61.5 (5.8) 56.0 (11.3)	42 29	26.4 (7.4) 25.4 (9.0) 0–48 higher score=best outcome	-
Moosmayer[27, 28]	2023	Norway Sweden	Pragmatic 3-arm	Calcific tendinopathy	6 hospitals	Lavage+steroid Sham lavage+steroid Sham	73 74 73	2 years	50.5 (8.5) 49.0 (8.8) 49.3 (9.0)	30 36 35	29.7 (7.4) 29.6 (7.6) 30.6 (6.3) 0–48 higher score=best outcome	EQ-5D-5L Pain VAS QuickDASH

(*continued on next page*)

Table 2 (*continued*)

Observational studies											
Author	Year	Country	Design	Population	Setting	Measurement properties	Sample size	Follow-up (final)	Age (years) Mean (SD)	Male (%)	Baseline OSS Mean (SD)	Additional Instruments
Rangan[24,32]	2015	UK	Pragmatic 2-arm	Proximal humerus fracture (displaced)	32 hospitals	Surgical Nonsurgical	125 125	2 years	66.6 (11.8) 65.4 (12.1)	22 24	NR NR 0–48 higher score=best outcome	SF-12

Abbreviations.
UK: United Kingdom; RCD: rotator cuff disorders; Constant: Constant-Murley Score MUA: manipulation under anaesthetic; OPC: outpatient clinic; PT: physical therapy; Mths: months; SD: standard deviation; IQR: interquartile range; NR: not reported; DASH: Disabilities of the Arm, Shoulder and Hand; UCLA: The University of California-Los Angeles Shoulder Scale; SF-12: Short Form 12 Health Survey; SF-36: Short Form 36 Health Survey; SPADI: Shoulder Pain and Disability Index; HAQ: Health Assessment Questionnaire; EQ-5D-3L: EuroQuol five-dimensional three-level; EQ-5D-5L: EuroQuol five-dimensional five-level; ASES: America Shoulder and Elbow Surgeons; SPONSA: Stanmore Percentage Of Normal Shoulder Assessment; MSK-HQ: Musculoskeletal Health Questionnaire; VAS: Visual Analogue Scale; SSV: Subjective Shoulder Value; FABQ-PA: Fear Avoidance Belief Scale - Physical Activity; HADS: Hospital Anxiety and Depression Scale.
[a] Some included studies were reported across multiple publications. Therefore, the number of references exceeds the number of unique studies.

adequate evidence of responsiveness using different measures including anticipated effect sizes and standardised response means over time (n = 6)[6,16,23,25,29,31], comparison of within-person change between improved versus unimproved groups (n = 3)[6,29,34], correlations in change scores with other instruments (n = 4)[23,31,34,35] and area under the curve calculations (n = 1)[34].

Clinical trial discrimination

Five studies evaluated clinical trial discrimination of the total OSS score[17,18,21,24,26–28,32] (Supplementary Table 15). Two were rated as green [17,18,21] and three as amber for methodological quality [24,26–28,32]. All but one study (amber quality) were judged to have equivocal performance[17,18,21,26–28]. Although these studies consistently showed no effect when none was expected, they did not detect a between-group difference of appropriate magnitude where this was expected (e.g., short-term effect of glucocorticoid injection). The study judged to have adequate performance also showed no effect of surgical versus nonsurgical management of a displaced proximal humerus fracture over two years when none was expected[24,32]. Based on these findings, the synthesis rating for clinical trial discrimination was amber.

Thresholds of meaning

All five studies that evaluated thresholds of meaning of the total OSS score were judged to have adequate performance due to the use of at least one meaningful anchor and triangulation of results[34–38] (Supplementary Table 16). However, only one study was rated green for quality, reporting a minimal important change (MIC) of 6 points (out of 48) six months after arthroscopic decompression for subacromial impingement[34]. Therefore, the synthesis rating was amber.

The remaining four studies were rated amber for quality due to attrition[36–38] or incomplete reporting[35]. Of these, three reported minimal clinically important differences (MCIDs) ranging from 2.6 to 11.4 following reverse shoulder arthroplasty[38], rotator cuff tear[37], and proximal humerus fracture[35].

Feedback from the OMERACT TAG and outcome of the OMERACT 2025 special interest group workshop

We agreed with TAG concerns about the suitability of the OSS for assessing pain. They considered the OSS to be more suited to measuring function or a combined construct of function and pain, rather than pain alone. The TAG reflected on evidence provided from the systematic review that although one study reported correlations between a four-item pain subscale of the OSS and other pain measures such as the VAS, stronger correlations were observed between the total OSS score and composite instruments such as the Constant-Murley score, which assesses both pain and function. This issue was further compounded because the pain subscale used in the single study had not been formally validated as a measure of pain.

There were 19 attendees at the OMERACT 2025 Shoulder Special Interest Group workshop. Sixteen of 18 (89 %) who responded, agreed with proposed ratings for the synthesis of each property. Two attendees (11 %) disagreed, citing concerns about the appropriateness of using the OSS total score to measure the pain domain. Eighteen (95 %) attendees agreed it was inappropriate to evaluate multidimensional instruments (e.g., the OSS) using the OMERACT Filter for a single domain (e.g., pain) in the absence of a validated subscale.

Discussion

We evaluated the measurement properties of the OSS using the OMERACT filter for assessing pain in clinical trials of shoulder disorders after it received an amber rating for domain match and feasibility based

Table 3
Summary of Findings of Measurement Properties (SOMP) for the Oxford Shoulder Score (OSS).

Instrument studied (& version): Oxford Shoulder Score (English)								
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	Feb 2025
Target Population:		**Intervention:**			**Control:**		**Desired use:**	
Shoulder complaints		Any			Any		Change within clinical trials	
Review findings*								
Source		Truth			**Discrimination**			
Author	**Year**	**Construct Validity**	**Floor & ceiling effects**	**Inter-method reliability**	**Test- retest reliability**	**Longitudinal construct validity†**	**Clinical trial discrimination**	**Thresholds of meaning**
Observational studies								
Allom[14]	2009							
Baker[15]	2008	+						
Bavan[16]	2020	± dissimilar constructs +known groups				+		
Beastall[19]	2012	- OSS pain +OSS total						
Cloke[20]	2005	+			+			
Dawson[6]	1996	+			+	+		
Dawson[22]	2001	+						
Evans[23]	2018		+			±		
Hapuarachchi[25]	2014	+				+		
Noorani[29]	2012	+				+		
Othman[30]	2004							
Price[31]	2019					±		
Singh[33]	2021		±					
Sorenson[34]	2021		+			+		+
Van de Water[35]	2014	+			+	+		+
Xu[36]	2019							+
Xu[37]	2020							+
Zhou[38]	2023							+
Randomised controlled trials								
Beard[17, 18]	2018						±	
Cloke[21]	2008						±	
Holt[26]	2013						±	
Moosmayer[27, 28]	2023						±	
Rangan[24, 32]	2015						+	
Synthesis								
Studies per property (n)		11	3	0	3	10	5	5
Used in synthesis		9	0	0	3	8	5	5
Synthesis statement per property‡			N/A*					
		We do not plan to seek OMERACT endorsement for this instrument for the pain domain as most of the evidence assessed the total OSS score which considers pain and function combined						

NB. All studies except for Beastall 2012 assessed the OSS total score, which incorporates both pain and function.
Blank: property not assessed.
Methods quality assessment: Green: sufficient; amber: equivocal, use with some caution; red: insufficient.
Performance assessment (not shown for studies with insufficient methods quality or for studies assessing floor and ceiling effects):.
+: findings above threshold of performance; -: findings below threshold of performance; ± equivocal findings.
‡ Synthesis statement per property reflects the quantity and quality of evidence to support the performance on each measurement property: Green: sufficient evidence of adequate performance; amber: equivocal evidence or equivocal performance; red: good or equivocal evidence of poor performance; N/A (grey): measurement property not needed. White: insufficient evidence to synthesize.
N/A Not applicable.
*Methodological quality for floor and ceiling effects was not appraised since this measurement property is not included in the COSMIN—OMERACT Good Methods Checklist.
† Also called responsiveness or sensitivity to change.

on patient partner feedback. Our systematic review identified 23 studies of which only one evaluated measurement properties of a pain subscale rather than the full OSS. Of the 34 components of evidence appraised, only eight were rated as having good methodological quality. Although the total OSS demonstrated adequate construct validity and responsiveness for measuring a multidimensional construct of pain and function, we concluded that it was unlikely to be suitable for measuring pain in isolation. Our review also resulted in amber ratings for test-retest reliability, clinical trial discrimination and thresholds of meaning of the total OSS, indicating limitations or uncertainties in the evidence.

Our finding that the total OSS score is not appropriate for measuring pain in clinical trials involving people with shoulder disorders is consistent with previous research using a refined International Classification of Functioning, Disability and Health (ICF) linking framework, which found that ten of the 14 main OSS concepts relate to constructs other than pain[39]. Although prior evidence has suggested potential for a four-item OSS pain subscale in rotator cuff repair populations[40], other studies do not support its use in populations with frozen shoulder or proximal humeral fracture[41]. Furthermore, only two of the four items in the subscale specifically assess pain intensity, which was reflected in our definition of the pain domain. While this review focused specifically on the suitability of the OSS for measuring pain, the same conceptual limitation applies to the domain of physical function. Because the OSS combines pain and function into a single score, it is not appropriate for evaluating either domain in isolation.

Given these limitations, alternative instruments for measuring pain require consideration. One instrument that shows promise is the Shoulder Pain and Disability Index (SPADI) pain subscale, which is a validated tool comprising five items that specifically measure pain intensity[42], and is commonly used in shoulder disorder trials[1]. The Patient-Reported Outcomes Measurement Information System (PROMIS) Pain Intensity scale[43] may also be a suitable candidate with evidence of favourable measurement properties in other musculoskeletal conditions[44,45]. It has demonstrated strong responsiveness following arthroscopic rotator cuff repair and total shoulder arthroplasty in longitudinal studies, but it has not, to our knowledge, yet been used in shoulder trials[46,47].

Although we concluded that the OSS total score is not suitable for measuring a single domain such as pain, this does not preclude its potential utility as a composite measure of pain and function. Nor does it rule out the possibility of further validation of a pain subscale. The OSS may also be of value in clinical practice, where a brief multidimensional tool is appropriate. It is quick to complete, taking only two and four minutes[7], user-friendly, and the total OSS has demonstrated adequate construct validity and responsiveness across a range of surgical and nonoperative shoulder conditions, including glenohumeral or acromioclavicular arthritis, rotator cuff disease, proximal humerus fractures, and adhesive capsulitis. The OSS has now been translated and culturally adapted into more than 21 languages across 24 countries[48], supporting its broader applicability in global clinical settings.

From a trial design perspective, the availability of quantitative thresholds of meaning for the total OSS remains limited with the evidence rated amber due to methodological limitations. Only one high-quality study reported a minimal important change of 6/48 points at six months following arthroscopic decompression, but this estimate reflects within-patient change and does not provide a between-group minimal important difference for trial design. Three additional studies-limited by attrition or incomplete reporting-estimated minimal clinically important differences ranging from 2.6 to 11.4 points following reverse shoulder arthroplasty, rotator cuff tear and proximal humerus fracture. Further high-quality studies are needed to establish robust thresholds across diverse clinical contexts to support trial design, sample size calculations and interpretation of treatment effects.

Given the interrelatedness of pain and function in shoulder disorders, composite endpoints may offer a pragmatic approach for reflecting mandatory domains of pain and function within a single measure. However, formal evaluation of the OSS as a composite outcome, including considerations such as domain weighting and validation within this framework, was beyond the scope of the present study. Future research is warranted to formally assess the suitability of the OSS as a composite measure of pain and function for use in clinical trials of shoulder disorders, particularly if suitable outcome measures for the separate domains of pain and physical function cannot be identified.

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 has not been updated since June 2023 meaning that some studies may not have been included. However, we do not believe additional evidence would have changed our conclusion that the OSS is unsuitable for measuring pain.

Conclusion

The OSS total score shows adequate construct validity and responsiveness, but its combined assessment of pain and function makes it unsuitable for measuring pain alone. While it does not pass the OMERACT Filter as a measure of pain in clinical trials involving people with shoulder disorders, it may still be suitable as a composite measure of pain and function and may have value in clinical practice. Further validation of separate pain and function subscales may enhance its applicability.

Data availability

Data extracted from included studies are provided in the supplementary materials. Requests for other materials can be sent to the corresponding author.

CRediT authorship contribution statement

Romi Haas: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Hana Marmura:** Writing – review & editing, Writing – original draft, Visualization, Investigation, Formal analysis, Data curation. **Rochelle Furtado:** Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation. **Sofia Ramiro:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Joel J. Gagnier:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Arianne Verhagen:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Samuel Whittle:** Writing – review & editing, Supervision, Conceptualization. **Dorcas Beaton:** Writing – review & editing, Methodology, Investigation, Formal analysis, Conceptualization. **Beverley Shea:** Writing – review & editing, Methodology. **Pamela Richards:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Conceptualization. **Danielle Berkovic:** Writing – review & editing, Investigation, Formal analysis, Conceptualization. **Andrew Firth:** Writing – review & editing, Validation, Investigation. **Rachelle Buchbinder:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

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.

Acknowledgements

We acknowledge the valuable contribution of Diane Horrigan, information specialist at Monash University, who designed and executed the search strategy and provided expert guidance on search terms.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.semarthrit.2026.152998.

References

[1] Page MJ, Huang H, Verhagen AP, et al. Outcome reporting in randomized trials for shoulder disorders: literature review to inform the development of a core outcome set. Arthritis Care Res (Hoboken) 2018;70:252–9.
[2] Ramiro S, Page MJ, Whittle SL, et al. The OMERACT core domain set for clinical trials of shoulder disorders. J Rheumatol 2019;46:969–75.
[3] Buchbinder R, Page MJ, Huang H, et al. A preliminary core domain set for clinical trials of shoulder disorders: a report from the OMERACT 2016 shoulder core outcome set special interest group. J Rheumatol 2017;44:1880–3.
[4] Page MJ, Huang H, Verhagen AP, et al. Identifying a core set of outcome domains to measure in clinical trials for shoulder disorders: a modified Delphi study. RMD Open 2016;2:e000380.
[5] Page MJ, O'Connor DA, Malek M, et al. Patients' experience of shoulder disorders: a systematic review of qualitative studies for the OMERACT Shoulder Core Domain Set. Rheumatology 2019;58:1410–21.
[6] Dawson J, Fitzpatrick R, Carr A. Questionnaire on the perceptions of patients about shoulder surgery. J Bone Joint Surg Br 1996;78:593–600.
[7] Buchbinder R, Ramiro S, Huang H, Gagnier JJ, Jia Y, Whittle SL, et al. Measures of adult shoulder function. Arthritis Care Res 2020;72(10):250–93. https://doi.org/10.1002/acr.24230.
[8] Bohm ER, Kirby S, Trepman E, et al. Collection and reporting of patient-reported outcome measures in arthroplasty registries: multinational survey and recommendations. Clin Orthop Relat Res® 2021;479:2151–66.
[9] Beaton D., Maxwell L., Grosskleg S., et al. The OMERACT handbook: for establishing and implementing core outcomes in clinical trials across the spectrum of rheumatoloic conditions. Version 2.1. April 2021.
[10] Dawson J, Rogers K, Fitzpatrick R, et al. The Oxford shoulder score revisited. Arch Orthop Trauma Surg 2009;129:119–23.
[11] Elsman EB, Mokkink LB, Terwee CB, et al. Guideline for reporting systematic reviews of outcome measurement instruments (OMIs): PRISMA-COSMIN for OMIs 2024. J Clin Epidemiol 2024;173:111422.
[12] Terwee CB, Jansma EP, Riphagen II, et al. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res 2009;18:1115–23.
[13] Beaton D, Boers M, Bingham CO, et al. Summary of findings tables for measurement property reviews: the evolution and application of OMERACT's summary of measurement properties (SOMP) Table. Seminars in arthritis and rheumatism, 72. Elsevier; 2025. p. 152664.
[14] Allom R, Colegate-Stone T, Gee M, et al. Outcome analysis of surgery for disorders of the rotator cuff: a comparison of subjective and objective scoring tools. J Bone Joint Surg Br 2009;91:367–73.
[15] Baker P, Nanda R, Goodchild L, et al. A comparison of the Constant and Oxford shoulder scores in patients with conservatively treated proximal humeral fractures. J Shoulder Elbow Surg 2008;17:37–41.
[16] Bavan L, Wood J, Surmacz K, et al. Instrumented assessment of shoulder function: a study of inertial sensor based methods. Clin Biomech 2020;72:164–71.
[17] Beard D, Rees J, Rombach I, Cooper C, Cook J, Merritt N, Gray A, Gwilym S, Judge A, Savulescu J, Moser J. The CSAW Study (Can Shoulder Arthroscopy Work?)–a placebo-controlled surgical intervention trial assessing the clinical and cost effectiveness of arthroscopic subacromial decompression for shoulder pain: study protocol for a randomised controlled trial. Trials 2015;16:210. https://doi.org/10.1186/s13063-015-0725-y.
[18] Beard DJ, Rees JL, Cook JA, et al. Arthroscopic subacromial decompression for subacromial shoulder pain (CSAW): a multicentre, pragmatic, parallel group, placebo-controlled, three-group, randomised surgical trial. Lancet 2018;391:329–38.
[19] Beastall JE, Fielding S, Christie E, et al. Shoulder outcome measures: is there a right answer? Eur J Trauma Emerg Surg 2012;38:659–64.
[20] Cloke D, Lynn S, Watson H, et al. A comparison of functional, patient-based scores in subacromial impingement. J Shoulder Elbow Surg 2005;14:380–4.
[21] Cloke DJ, Watson H, Purdy S, et al. A pilot randomized, controlled trial of treatment for painful arc of the shoulder. J Shoulder Elbow Surg 2008;17:S17–21.
[22] Dawson J, Hill G, Fitzpatrick R, et al. The benefits of using patient-based methods of assessment: medium-term results of an observational study of shoulder surgery. J Bone Joint Surg Br 2001;83:877–82.
[23] Evans J, Dattani R, Ramasamy V, et al. Responsiveness of the EQ-5D-3L in elective shoulder surgery: does it adequately represent patient experience? J Orthop Surg 2018;26(2). 2309499018774922.
[24] Handoll H, Brealey S, Rangan A, et al. Protocol for the ProFHER (PROximal Fracture of the Humerus: evaluation by Randomisation) trial: a pragmatic multicentre randomised controlled trial of surgical versus non-surgical treatment for proximal fracture of the humerus in adults. BMC Musculoskelet Disord 2009;10:1–11.
[25] Hapuarachchi KS, Poon PC. A correlation study of the American Shoulder and Elbow Society Score and the Oxford Shoulder Score with the use of regression analysis to predict one score from the other in patients undergoing reverse shoulder joint arthroplasty for cuff tear arthropathy. Shoulder Elbow 2014;6:81–9.
[26] Holt TA, Mant D, Carr A, Gwilym S, Beard D, Toms C, Yu LM, Rees J. Corticosteroid injection for shoulder pain: single-blind randomized pilot trial in primary care. Trials 2013;14(1):425. https://doi.org/10.1186/1745-6215-14-425.
[27] Moosmayer S, Ekeberg OM, Hallgren HB, et al. KALK study: ultrasound guided needling and lavage (barbotage) with steroid injection versus sham barbotage with and without steroid injection-protocol for a randomized, double-blinded, controlled, multicenter study. BMC Musculoskelet Disord 2017;18(1):138.
[28] Moosmayer S, Ekeberg OM, Hallgren HB, et al. Ultrasound guided lavage with corticosteroid injection versus sham lavage with and without corticosteroid injection for calcific tendinopathy of shoulder: randomised double blinded multi-arm study. BMJ 2023:383.
[29] Noorani AM, Roberts DJ, Malone AA, et al. Validation of the Stanmore percentage of normal shoulder assessment. Int J Shoulder Surg 2012;6(1):9.
[30] Othman A, Taylor G. Is the constant score reliable in assessing patients with frozen shoulder? 60 shoulders scored 3 years after manipulation under anaesthesia. Acta Orthop Scand 2004;75(1):114–6.
[31] Price AJ, Ogollah R, Kang S, et al. Determining responsiveness and meaningful changes for the Musculoskeletal Health Questionnaire (MSK-HQ) for use across musculoskeletal care pathways. BMJ Open 2019;9:e025357.
[32] Rangan A, Handoll H, Brealey S, et al. Surgical vs Nonsurgical Treatment of Adults With Displaced Fractures of the Proximal Humerus: the PROFHER Randomized Clinical Trial. JAMA, J Am Med Assoc 2015;313(10):1037–47.
[33] Singh HP, Haque A, Taub N, et al. Floor and ceiling effects in the Oxford Shoulder Score: an analysis from the National Joint Registry. Bone Joint J 2021;103:1717–24.
[34] Sørensen L, van Tulder M, Johannsen HV, et al. Responsiveness and minimal important change of the Oxford Shoulder Score, EQ-5D, and the Fear-Avoidance Belief Questionnaire Physical Activity subscale in patients undergoing arthroscopic subacromial decompression. JSES Int 2021;5:869–74.
[35] van de Water AT, Shields N, Davidson M, et al. Reliability and validity of shoulder function outcome measures in people with a proximal humeral fracture. Disabil Rehabil 2014;36:1072–9.
[36] Xu S, Chen JY, Lie HME, et al. Determination of threshold scores for treatment success after arthroscopic rotator cuff repair using Oxford, constant, and University of California, Los Angeles shoulder scores. Arthroscopy 2019;35:304–11.
[37] Xu S, Chen JY, Lie HME, et al. Minimal clinically important difference of Oxford, Constant, and UCLA shoulder score for arthroscopic rotator cuff repair. J Orthop 2020;19:21–7.
[38] Zhou A, Xu S, Yew KSA, et al. Minimal clinically important differences for Oxford, Constant, and University of California Los Angeles shoulder scores after reverse shoulder arthroplasty to allow interpretation of patient-reported outcome measures and future statistical power analyses. Arthroscopy 2023;39:1405–14.
[39] Røe Y, Buchbinder R, Grotle M, et al. What do the OMERACT shoulder core set candidate instruments measure? An analysis using the refined international classification of functioning, disability, and health linking rules. J Rheumatol 2020;47:1557–64.

[40] Dawson J, Harris KK, Doll H, et al. A comparison of the Oxford shoulder score and shoulder pain and disability index: factor structure in the context of a large randomized controlled trial. Patient Relat Outcome Meas 2016:195–203.
[41] Simpson J, Keding A, Spencer S, et al. Factor structure of the Oxford Shoulder Score: secondary analyses of the UK FROST and PROFHER trial populations. J Orthop Surg Res 2023;18:846.
[42] Roach KE, Budiman-Mak E, Songsiridej N, et al. Development of a shoulder pain and disability index. Arthritis Rheumat 1991;4:143–9.
[43] Revicki D, Cook K. PROMIS pain-related measures: an overview. Pract Pain Manag 2015;15.
[44] Braaksma C, Wolterbeek N, Veen M, et al. Assessing the measurement properties of PROMIS Computer Adaptive Tests, short forms and legacy patient reported outcome measures in patients undergoing total hip arthroplasty. J Patient Rep Outcomes 2024;8:121.
[45] Stephan A, Stadelmann VA, Preiss S, et al. Measurement properties of PROMIS short forms for pain and function in patients receiving knee arthroplasty. J Patient Rep Outcomes 2023;7:18.
[46] Alben MG, Romeo PV, Papalia AG, et al. Does the addition of Patient-Reported Outcome Measure Information System (PROMIS) pain instruments improve the sensitivity of PROMIS upper extremity scores after arthroscopic rotator cuff repair? J Shoulder Elbow Surg 2025;34:595–605.
[47] Romeo PV, Alben MG, Papalia AG, et al. Addition of PROMIS pain instruments to PROMIS upper extremity physical function improves the responsiveness of PROMIS scores compared to legacy scores in patients undergoing total shoulder arthroplasty: a prospective study. J Shoulder Elbow Surg 2025;34(8):e702–712.
[48] Oxford University Innovation. OSS Available languages, https://innovation.ox.ac.uk/outcome-measures/oxford-shoulder-score-oss/(accessed July 29 2025).