








BMJ Open Clinical outcome measures and their evidence base in degenerative cervical myelopathy: a systematic review to inform a core measurement set (AO Spine RECODE-DCM)

Alvaro Yanez Touzet ¹, Aniqah Bhatti,² Esmee Dohle,² Faheem Bhatti ², Keng Siang Lee ³, Julio C Furlan,^{4,5,6} Michael G Fehlings,⁷ James S Harrop,⁸ Carl Moritz Zipser ⁹, Ricardo Rodrigues-Pinto ^{10,11}, James Milligan,¹² Ellen Sarewitz,¹³ Armin Curt,⁹ Vafa Rahimi-Movaghar,¹⁴ Bizhan Aarabi,¹⁵ Timothy F Boerger ¹⁶, Lindsay Tetreault,¹⁷ Robert Chen,^{17,18} James D Guest,¹⁹ Sukhvinder Kalsi-Ryan,⁶ Angus GK McNair ^{20,21}, Mark Kotter,^{22,23} Benjamin Davies,²⁴ On behalf of the AO Spine RECODE-DCM Steering Committee

To cite: Yanez Touzet A, Bhatti A, Dohle E, *et al*. Clinical outcome measures and their evidence base in degenerative cervical myelopathy: a systematic review to inform a core measurement set (AO Spine RECODE-DCM). *BMJ Open* 2022;**12**:e057650. doi:10.1136/bmjopen-2021-057650

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2021-057650>).

Received 24 September 2021
Accepted 22 December 2021



© Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to
Dr Benjamin Davies;
bd375@cam.ac.uk

ABSTRACT

Objectives To evaluate the measurement properties of outcome measures currently used in the assessment of degenerative cervical myelopathy (DCM) for clinical research.

Design Systematic review

Data sources MEDLINE and EMBASE were searched through 4 August 2020.

Eligibility criteria Primary clinical research published in English and whose primary purpose was to evaluate the measurement properties or clinically important differences of instruments used in DCM.

Data extraction and synthesis Psychometric properties and clinically important differences were both extracted from each study, assessed for risk of bias and presented in accordance with the Consensus-based Standards for the selection of health Measurement Instruments criteria.

Results Twenty-nine outcome instruments were identified from 52 studies published between 1999 and 2020. They measured neuromuscular function (16 instruments), life impact (five instruments), pain (five instruments) and radiological scoring (five instruments). No instrument had evaluations for all 10 measurement properties and <50% had assessments for all three domains (ie, reliability, validity and responsiveness). There was a paucity of high-quality evidence. Notably, there were no studies that reported on structural validity and no high-quality evidence that discussed content validity. In this context, we identified nine instruments that are interpretable by clinicians: the arm and neck pain scores; the 12-item and 36-item short form health surveys; the Japanese Orthopaedic Association (JOA) score, modified JOA and JOA Cervical Myelopathy Evaluation Questionnaire; the neck disability index; and the visual analogue scale for pain. These include six scores with barriers to application and one score with insufficient criterion and construct validity.

Strengths and limitations of this study

- Consensus-based reporting guidelines were used to evaluate the properties and clinically important differences of degenerative cervical myelopathy measurement instruments.
- Only instruments that are currently in use were evaluated in this study.
- Interpretability was used as an important characteristic to make recommendations, a posteriori, due to the absence of category A Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) recommendations.
- Interpretability and feasibility were evaluated using bespoke criteria adapted, a priori, from the COSMIN methodology.

Conclusions This review aggregates studies evaluating outcome measures used to assess patients with DCM. Overall, there is a need for a set of agreed tools to measure outcomes in DCM. These findings will be used to inform the development of a core measurement set as part of AO Spine RECODE-DCM.

INTRODUCTION

The most common adult spinal cord disease, degenerative cervical myelopathy (DCM), is both measured and reported inconsistently across clinical research.^{1–4} DCM is a progressive spinal cord disease caused by degenerative changes in the cervical spine that lead to stress and injury to the cervical spinal cord. It usually initially presents as a loss of digital dexterity, subtle gait disturbances and mild

pain which, if left untreated, can potentially lead to tetraplegia and wheelchair dependence.⁵

In 2019, AO Spine launched the Research Objectives and Common Data Elements for Degenerative Cervical Myelopathy (AO Spine RECODE-DCM; www.aospine.org/recode) initiative with the aim of creating a 'research toolkit' to help accelerate knowledge discovery and improve outcomes in DCM.^{3 6} The initiative identified the need to improve consistency in measurement and reporting across DCM research to enable studies to be compared and/or aggregated, and to ensure the most meaningful aspects of the disease are captured.^{7 8} This process started by creating a list of essential outcomes (ie, core outcome set) and baseline characteristics (ie, core data elements). To truly enable consistent reporting, however, these datasets should be partnered with a core measurement set (CMS): a set of agreed tools that are used to measure the outcomes and data elements of DCM.^{9–17}

Several approaches have been employed to form a CMS, ranging from the development of novel measurement instruments to adopting the use of existing ones.^{18–20} For AO Spine RECODE-DCM, it was decided to recommend existing instruments and, preferably, those already used in DCM. This was to allow a more rapid introduction of the CMS, cognisant that many new tools are in development and the CMS can be updated in the future.

Consequently, we sought to examine the tools used in DCM research and assess their quality²¹ using objective criteria. In recognition of variable quality among reported outcome measures, the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) initiative has developed clinimetric tools to assess instrument quality.²² We searched the literature for studies evaluating one or more psychometric properties defined by the COSMIN guidelines, as well as studies that defined clinically important differences such as the minimally clinical important difference (MCID) and substantial clinical benefits (SCBs). Data were rated, aggregated and assessed for methodology bias using the COSMIN manual for systematic reviews of patient-reported outcome measures (PROMs).^{23–25} This work builds on the protocol for the AO Spine RECODE-DCM initiative^{3 6 26} and complements two earlier reviews of outcome measures in DCM.^{2 21}

METHODS

Search

A search string was developed to identify original research assessing the psychometric properties of instruments currently used in the clinical research of DCM.²⁷ This comprised synonyms of 'psychometric' and 'DCM' (online supplemental table 1). The search was developed with oversight of a medical librarian (IK) and informed by previously developed search filters for DCM.^{27–29} The search was applied to MEDLINE and EMBASE, from inception until 4 August 2020, using OVID (Wolters

Table 1 Inclusion and exclusion criteria

Inclusion	Exclusion
Publication type	
<ul style="list-style-type: none"> ▶ Article written in English ▶ Primary clinical research articles 	<ul style="list-style-type: none"> ▶ Article not written in English ▶ Conference abstracts or posters ▶ Editorials, commentaries, opinion papers or letters ▶ Book chapters or theses
Study type	
<ul style="list-style-type: none"> ▶ Study includes primary clinical data 	<ul style="list-style-type: none"> ▶ Study uses only secondary data ▶ Case reports ▶ Narrative reviews ▶ Systematic reviews ▶ Meta-analyses
Populations	
<ul style="list-style-type: none"> ▶ Human studies 	<ul style="list-style-type: none"> ▶ Non-human studies
Indications	
<ul style="list-style-type: none"> ▶ Exclusively DCM (CSM, OPLL, cervical stenosis, spondylosis, spinal cord compression, cervical myelopathy) 	<ul style="list-style-type: none"> ▶ Populations with DCM and at least one other condition (eg, radiculopathy)
Comparator	
<ul style="list-style-type: none"> ▶ At least one assessment tool^{2 21 30} 	
Outcomes	
<ul style="list-style-type: none"> ▶ At least one psychometric property ▶ At least one MCID or SCB 	

CSM, Cervical spondylotic myelopathy; DCM, degenerative cervical myelopathy; MCID, minimally clinical important difference; OPLL, Ossification of the posterior longitudinal ligament; SCB, substantial clinical benefits.

Kluwer, Netherlands). The search also focused on DCM tools identified in previous scoping reviews.^{2 21 30}

Study selection

All titles and abstracts were screened independently against a set of predefined eligibility criteria by four reviewers (AYT, AB, ED and FB). A full list of inclusion and exclusion criteria of studies are stated in [table 1](#).

Potentially eligible studies were selected for full-text analysis. In the event of multiple publications analysing the same cohort for the same purpose, the most recent paper was used for evaluation. At each stage, two reviewers independently (AYT, AB, ED, FB) reviewed all the screened studies for inclusion to ensure reliability of study selection (online supplemental table 2). Disagreements were resolved by consensus or appeal to a third senior reviewer (BMD).

Table 2 Definitions of domains, measurement properties and aspects of measurement properties, adapted from the COSMIN guidelines^{23–25 48} and studies of clinically important differences^{49 50}

Domain	Measurement property	Aspect	Definition
Reliability			The degree to which the measurement is free from measurement error. The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions; over time; by different persons on the same occasion; or by the same persons on different occasions.
	Internal consistency		The degree of inter-relatedness among the items included in a measurement instrument.
	Reliability		The proportion of the total variance in the measurements which is due to ‘true’* differences between patients.
	Measurement error		The systematic and random error of a patient’s score that is not attributed to true changes in the construct to be measured.
Validity			The degree to which a measurement instrument measures the construct(s) it purports to measure.
	Content validity		The degree to which the content of a measurement tool is an adequate reflection of all facets of a given construct.
	Construct validity		The degree to which the scores of a measurement instrument are consistent with hypotheses (for instance, with regard to internal relationships, relationships to scores of other instruments or differences between relevant groups) based on the assumption that the instrument validly measures the construct to be measured.
		Structural validity	The degree to which the scores of a measurement instrument are an adequate reflection of the dimensionality of the construct to be measured.
		Hypotheses testing	Idem construct validity.
		Cross-cultural validity	The degree to which the performance of the items on a translated or culturally adapted measurement instrument are an adequate reflection of the performance of the items of the original version of the instrument.
	Criterion validity		The degree to which the scores of a measurement instrument are an adequate reflection of a ‘gold standard’.
Responsiveness			The ability of a measurement instrument to detect change over time in the construct to be measured.
	Responsiveness		Idem responsiveness.
Interpretability†			Interpretability is the degree to which one can assign qualitative meaning—that is, clinical or commonly understood connotations—to a PROM’s quantitative scores or change in scores.
	Clinically important differences		
		Minimal clinically important difference	The smallest measured change score that patients perceive to be important, also known as the MCID or MID
		Substantial clinical benefit	The change in outcome associated with patient perception of a large meaningful improvement.

*The word ‘true’ must be seen in the context of the classical test theory, which states that any observation is composed of two components—a true score and error associated with the observation. ‘True’ is the average score that would be obtained if the scale was applied infinite number of times. It refers only to the consistency of the score, and not to its accuracy.⁵¹

†Interpretability is not considered a measurement property, but an important characteristic of a measurement instrument.

COSMIN, Consensus-based Standards for the selection of health Measurement Instruments; MCID, minimally clinical important difference; MID, minimally important difference; PROMs, patient-reported outcome measures.

Quality assessment

The quality of included studies was assessed using the COSMIN risk of bias checklist.^{23–25} Briefly, the COSMIN risk of bias tool assesses 10 measurement properties, including nine psychometric properties (ie, content validity, structural validity, internal consistency, cross-cultural validity/measurement invariance, reliability, measurement error, criterion validity, hypotheses testing for construct validity and responsiveness) and clinically important differences. A list of definitions is presented in table 2. Interpretability and feasibility were also evaluated using criteria adapted a priori from the COSMIN

methodology (online supplemental tables 3 and 4), respectively). Namely, interpretability was evaluated for each measurement instrument through the availability of anchor-based MCIDs,^{23–25} while feasibility was assessed with respect to the ease of application of the instrument.

The methodological quality of each study was scored as ‘very good’, ‘adequate’, ‘doubtful’, ‘inadequate’ or ‘not applicable’. Overall ratings were then made for each property using the modified Grading of Recommendations Assessment, Development, and Evaluation approach from the COSMIN risk of bias checklist.^{23–25} For each study, one review author (AYT) assessed the quality,

feasibility and interpretability from included studies and a second (BD) checked the assessments. Disagreements were resolved by consensus.

Data extraction

A proforma adapted from COSMIN was employed by one reviewer (AYT) to extract the following: study details, sample size, patient demographics, measurement properties and qualitative and/or quantitative results for each property. This was checked by a second reviewer (BD) and any disagreements were resolved by consensus. Examples of qualitative and quantitative results included observations (eg, narrative syntheses) and statistics (eg, correlation coefficients). These result types are specific for each measurement property and are listed in the COSMIN guidelines.^{23–25}

Data analysis

Each result was rated as ‘sufficient’, ‘indeterminate’ or ‘insufficient’. All results were qualitatively summarised and given an overall rating as ‘sufficient’, ‘indeterminate’, ‘inconsistent’ or ‘insufficient’. The definitions of these ratings are available in the COSMIN guidelines.^{23–25} Measurement instruments were categorised into three recommendation groups:

1. Instruments with evidence of sufficient content validity and at least low-quality evidence of sufficient internal consistency.
2. Instruments categorised not in 1 or 3.
3. Instruments with high-quality evidence of an insufficient measurement property.^{23–25}

Recommendations for each instrument were presented in tandem with interpretability and feasibility assessments and reported as a narrative synthesis.³¹ We used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses checklist when writing our report.³²

Patient and public involvement

This project forms part of a larger, international multi-stakeholder co-production initiative called AO Spine RECODE-DCM, which aims to develop a framework to accelerate knowledge discovery that can improve outcomes in DCM. Patients and the public were therefore involved in its overall design, conduct, management, and dissemination, and are recognised among the authors of this article. For further information, please refer to www.aospine.org/recode.

RESULTS

Literature search

The primary literature search identified a total of 3239 unduplicated studies (MEDLINE: 2389, EMBASE: 1550). Abstract and full-text screening excluded 3187 studies. Therefore, this review included a total of 52 studies (figure 1 and online supplemental table 2).

Study properties

The 52 included studies assessed a total of 7395 patients worldwide (female: 3217, male: 4178) with 29 instruments

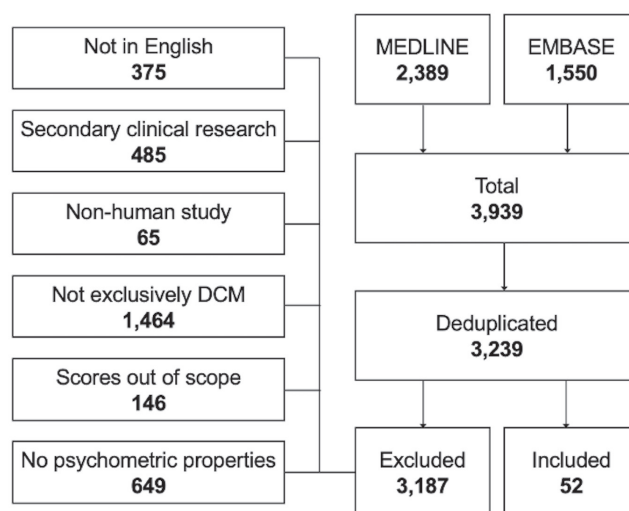


Figure 1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow chart. A systematic review of Medline and EMBASE was conducted through 4 August 2020 to identify original research on the measurement properties of instruments currently used in degenerative cervical myelopathy research.

(table 3). These were classified into four domains based on the DCM core outcome set³³: neuromuscular function, life impact, pain, and radiological scoring.

Measurement properties

The measurement properties of the 29 instruments were evaluated using the COSMIN methodology for systematic reviews.^{23–25} A summary of findings is presented in table 4¹: the overall feasibility rating,² the overall interpretability rating and³ the overall recommendation category based on existing evidence. Included studies reported on at least one of the 10 COSMIN properties for all instruments. No instrument had evidence for all 10 properties and <50% (13/29) of instruments had evidence for at least one property per measurement domain (figure 2).

Content validity

Only three measurement instruments were evaluated for content validity: the JOA Cervical Myelopathy Evaluation Questionnaire (JOACMEQ), the modified JOA (mJOA) score and the Berg Balance Scale (BBS) (online supplemental table 5). The overall ratings for content validity, however, were indeterminate due to the uncertainty of the methods used to assess comprehensibility, and the very low quality of the evidence.

Structural validity

No instruments were assessed for structural validity.

Internal consistency

Ten measurement instruments were evaluated for internal consistency, including the JOACMEQ, JOA, mJOA, 12-Item Short Form Health Survey (SF-12) and SF-36 (online supplemental table 6). Since structural validity is required for the interpretation of internal consistency,

Table 3 Study properties

Property	Number	%
Total studies, included	52	100
Prospective	31	60
Retrospective	21	40
Total patient sample	7395	100
Male	4178	56
Female	3217	44
Measurement instruments by domain*	29	100
Neuromuscular function	16	55
Life impact	5	17
Pain	5	17
Radiological scoring	5	17
Publication year		
Maximum year of publication	2020	–
Median year of publication	2014	–
Mean year of publication	2012	–
Minimum year of publication	1999	–
Countries, by number of patients	7395	100
Japan	2014	27
USA	1802	24
Canada	1361	18
South Korea	726	10
Global/multicentre	601	8
China	255	3
India	121	2
Iran	87	1
Brazil	85	1
Italy	75	1
Hong Kong	72	1
Thailand	70	1
Taiwan	45	1
UK	41	1
France	40	1

*Instrument counts per domain do not add up to the total due to the one-to-many relationship between certain instruments and domains (eg, JOACMEQ is used both for life impact and neuromuscular function; see table 4).

JOACMEQ, Japanese Orthopaedic Association Cervical Myelopathy Evaluation Questionnaire.

the overall ratings for internal consistency were indeterminate, given the aforementioned absence of studies on structural validity.

Cross-Cultural validity

Only three measurement instruments were evaluated for cross-cultural validity: JOACMEQ, JOA and mJOA (online supplemental table 7). The overall ratings were indeterminate due to the absence of multiple group factors

analyses and differential item functioning analyses. The quality of evidence was also very low due to the uncertainty of the approaches used to analyse the data.

Reliability

Seventeen measurement instruments were evaluated for reliability, including JOACMEQ, JOA and mJOA (online supplemental table 8). The reported measures of reliability were test–retest reliability, intraobserver reliability and interobserver reliability. No instrument attained high-quality evidence for sufficient or insufficient reliability due to (1) imprecision (sample sizes <100), (2) serious inconsistency and/or (c) serious risk of bias.

Measurement error

Nine instruments were evaluated for measurement error, including JOACMEQ, JOA, mJOA, NDI, SF-36 and Visual Analogue Scale (VAS) for pain (online supplemental table 9)). The measures of error reported were minimal detectable change and distribution-based MCID.^{23–25 34} The mJOA was the only score to attain high-quality evidence for sufficiency (distribution-based MCID range: 1.2–1.4, total sample size: 868). Due to the inconsistency of results, the quality of the evidence of most other instruments could not to be rated.

Criterion validity

Twelve measurement instruments were evaluated for criterion validity, including the JOACMEQ, JOA, mJOA, NDI and SF-36 (online supplemental table 10). Both the mJOA and the patient-derived version of the mJOA (P-mJOA) attained high-quality evidence for sufficient criterion validity as whole scales. However, three of four items of the mJOA, along with the 10 s step test and foot tapping test, attained high-quality evidence for insufficient criterion validity (ie, these subdomains lack criterion validity for their use as separate measures). The quality of the evidence of most of the remaining instruments was not high due to (1) imprecision (ie, sample sizes <100) or (b) important methodological flaws in the design or statistical methods.

Construct validity

Sixteen measurement instruments were evaluated for construct validity, including JOACMEQ, JOA, mJOA, NDI, arm and neck pain scores and SF-12 (online supplemental table 11). From these, 8 of 16 attained high-quality evidence for sufficient construct validity; these included the NDI, arm and neck pain scores and SF-12. Two instruments achieved high-quality evidence for insufficient construct validity. Notably, the mJOA had both high-quality sufficiency and insufficiency depending on the comparator tool (eg, sufficiency with respect to the NDI and SF-36 and insufficiency with respect to the 30 m walking test (30MWT) and EuroQol-5 Dimension (EQ-5D)). While the designs and statistical methods applied were adequate for the research questions posed, the quality of the evidence of most of the remaining tools ranged from ‘low’ to ‘moderate’ due to imprecision (ie,

**Table 4** Summary of findings

Domain	Instrument	Feasibility	Interpretability	Recommendation category	Recommendation justification
Life impact					
	EQ-5D	+	+	C	High-quality evidence for insufficient construct validity
	SF-12	-	+	B	Indeterminate result rating for internal consistency
	SF-36	-	+	B	Indeterminate result rating for internal consistency
	WHOQOL-Bref	+	-	B	Indeterminate result rating for internal consistency
Life impact and neuromuscular function					
	JOACMEQ	+	+	B	
Neuromuscular function					
	10 s step test	+	-	C	High-quality evidence for insufficient criterion validity
	30MWT	+	-	C	High-quality evidence for insufficient responsiveness
	9-Hole peg test	++	-	B	
	BBS	++	-	B	
	European Myelopathy Scale	+	-	B	
	Foot tapping test	+	-	C	High-quality evidence for insufficient criterion validity
	Grip-and-release test	+	-	B	
	JOA	-	+	B	
	MDI	+	-	B	
	mJOA	-	+	C	High-quality evidence for insufficient criterion and construct validity
	Nurick scale	+	-	B	
	P-mJOA	+	-	B	
	Ranawat classification of disease severity	-	-	B	
	Triangle step test	+	-	B	
Pain and neuromuscular function					
	QuickDASH	-	-	B	
Pain					
	NDI	+	+	B	
	Arm pain score	-	+	B	
	Neck pain score	+	+	B	
	VAS for pain	+	+	B	
Radiology					
	Cobb's method	+	-	B	
	CT (Tsuayama's classification, 2D and 3D)	+	-	B	

Continued

Table 4 Continued

Domain	Instrument	Feasibility	Interpretability	Recommendation category	Recommendation justification
	CT (Tsuyama's classification, lateral + axial)	+	–	B	
	Isihara's cervical curvature index	+	–	B	
	MRI (depiction of intramedullary hyperintensity at eight cervical disc levels, T2W, 1.5-T or 3-T)	+	–	B	
	MRI (Kang's classification, 1.5-T or 3-T)	+	–	B	
	MRI (Muhle's classification, 1.5-T)	+	–	B	
	MRI (Vaccaro's classification, 1.5-T)	+	–	B	
	X-rays (computer-assisted measurement of length and thickness)	+	–	B	

Feasibility: ++=No barriers; +=Minimal barriers; –=Barriers

Interpretability: +=Interpretable; –= Uninterpretable, due to absence of anchor-based MCIDs ^{23–25}

Recommendation category: A=measurement instruments with evidence for sufficient content validity (any level) AND at least low-quality evidence for sufficient internal consistency; B=measurement instruments categorised not in A or C; C=measurement instruments with high-quality evidence for an insufficient measurement property.

BBS, Berg Balance Scale; EQ-5D, EuroQol-5 Dimension; JOA, Japanese Orthopaedic Association; JOACMEQ, Japanese Orthopaedic Association Cervical Myelopathy Evaluation Questionnaire; MDI, Myelopathy Disability Index; mJOA, modified Japanese Orthopaedic Association; 30MWT, 30-m Walking Test; NDI, Neck Disability Index; P-mJOA, patient-derived version of the mJOA; SF-12, 12-Item Short Form Health Survey; SF-36, 36-Item Short Form Health Survey; VAS, Visual Analogue Scale; WHOQOL-Bref, World Health Organisation Quality of Life.

sample sizes <100). Importantly, only one study formulated a hypothesis a priori.³⁵

Responsiveness

Sixteen measurement instruments were evaluated for responsiveness, including the JOACMEQ, JOA, mJOA, NDI, SF-12 and SF-36 (online supplemental table 12). The mJOA was the only score to attain high-quality evidence for sufficient responsiveness (effect size range: 0.87–1.0, total sample size: 352). The 30MWT, on the other hand, was the only score to attain high-quality evidence for insufficient responsiveness (standardised response mean: 0.3, total sample size: 484). The quality of the evidence of most of the remaining tools ranged from 'very low' to 'moderate' due to (1) imprecision (ie, sample sizes <100) and (b) uncertainty of the statistical methods.

Clinically important differences

Ten measurement instruments were evaluated for clinically important differences, including the JOACMEQ, JOA, mJOA, NDI, arm and neck pain scores, SF-12, SF-36 and VAS for pain (online supplemental table 3). From these, 7 of 10 attained a sufficient rating, including the JOACMEQ, JOA, mJOA, NDI and SF-36. Only anchor-based measures were accepted for the assessment of the MCID.^{23–25 36–39}

Interpretability and feasibility

Interpretability and feasibility were described using criteria adapted from the COSMIN methodology (online supplemental tables 3 and 4, respectively). Interpretability was summarised in terms of the degree to which

clinicians may assign qualitative meaning to the scores or change in scores (ie, the clinically important differences), while feasibility was described in terms of the ease of application of the measurement instrument. No or minimal application barriers were identified for most outcome measures (table 4). Nine instruments were, however, deemed uninterpretable due to the absence of anchor-based MCIDs.^{23–25}

Recommendations

No category A recommendations were made as no measurement instrument had sufficient evidence for content validity (table 4 and figure 2). Furthermore, five instruments were recommended for category C due to the availability of high-quality evidence for insufficient criterion validity, construct validity and/or responsiveness. Most instruments were classed into category B due to the notable absence of high-quality evidence for most measurement properties.

In light of these results, and given both (1) the very strict quality standards of the COSMIN framework and (2) that the absence of category A evidence is not the same as presence of poor-quality evidence, we propose that instruments most suitable for use should be interpretable by clinicians and offer qualitative meaning to either clinicians or people with lived experience of DCM (ie, they should have an available assessment of clinically important differences). To this end, the measurement properties of the nine interpretable instruments are presented in table 5: the arm and neck pain scores; SF-12 and SF-36; JOA, mJOA and JOACMEQ; NDI; and VAS

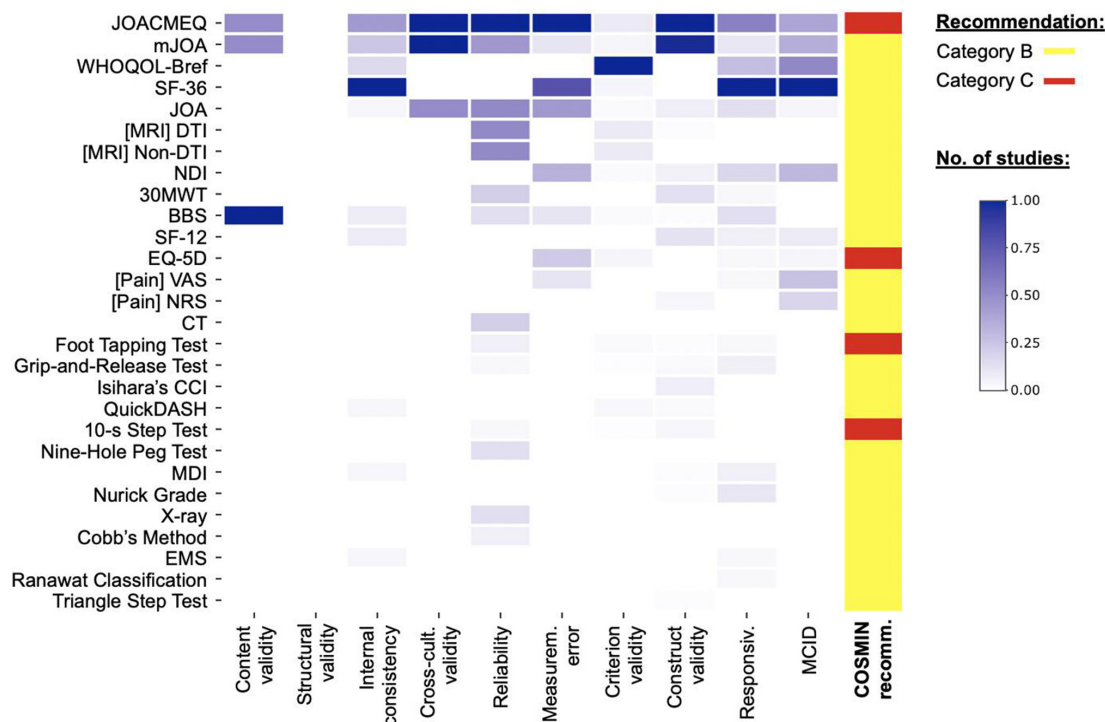


Figure 2 Number of studies for each outcome measure and property (normalised). Included studies reported on at least one of the 10 COSMIN properties for all instruments. No instrument had evidence for all 10 properties and <50% (13/29) of instruments had evidence for at least one property per measurement domain (see table 2 for definitions). Notably, no instruments were evaluated for structural validity, attained sufficient evidence for content validity or obtained a category A recommendation based on COSMIN criteria. 30MWT, 30-m Walking Test; BBS, Berg Balance Scale; COSMIN, Consensus-based Standards for the selection of health Measurement Instruments; EQ-5D, EuroQol-5 Dimension; JOA, Japanese Orthopaedic Association; JOACMEQ, Japanese Orthopaedic Association Cervical Myelopathy Evaluation Questionnaire; MDI, Myelopathy Disability Index; mJOA, modified Japanese Orthopaedic Association; NDI, Neck Disability Index; P-mJOA, patient-derived version of the mJOA; SF-12, 12-Item Short Form Health Survey; SF-36, 36-Item Short Form Health Survey; VAS, Visual Analogue Scale; WHOQOL-Bref, World Health Organisation Quality of Life

for pain. These include one score with insufficient criterion and construct validity (ie, mJOA) and six scores with barriers to application.

DISCUSSION

DCM is measured and reported inconsistently across clinical trials.¹⁻⁴ In light of these inconsistencies, AO Spine launched RECODE-DCM (www.aospine.org/recode) with the aim of creating a 'research toolkit' that helps to accelerate knowledge discovery and improve outcomes in DCM. One of the objectives of the RECODE-DCM initiative was to develop a CMS.^{3 6 26} This systematic review consists of an initial step towards building this CMS by identifying tools that have been used in DCM research and examining their quality, in accordance with the COSMIN standards.²³⁻²⁵

Overall, we identified 29 instruments with at least 1 in 10 measurement properties evaluated (figure 2); none, however, had evaluations for all 10 properties and <50% had more than one property evaluated per measurement domain (ie, reliability, validity and responsiveness) (table 2). We also noted a paucity in the quantity and quality of studies evaluating DCM instruments; this is visible by the absence of category A recommendations

and the classification of most tools in category B (table 4). Acknowledging both the stringency of the COSMIN standards and that absence of category A evidence is not equivalent to presence of poor-quality evidence, we proposed nine instruments that seem interpretable to clinicians and appear to offer qualitative meaning to clinicians and people with lived experience of DCM. These instruments are the SF-12 and SF-36; JOA, mJOA, and JOACMEQ; NDI; and VAS for pain (table 5).

The fact that most outcomes received B-category recommendations due to absence of high-quality evidence is not unexpected. In this review, the most common reasons for low-quality evidence, as per the COSMIN guidelines, were (1) important methodological flaws in study design or statistical methods, (2) uncertainty of approaches used to analyse the data and (3) imprecision due to sample size below the recommended power and significance levels. The rigour (or stringency) of the COSMIN standards may have accentuated these limitations due to the highly specific nature of some standards and the expectation of psychometric expertise within the DCM context. For example, results for internal consistency must be rated 'indeterminate' if there is not at least low-quality evidence for structural validity. No such studies were available in

Table 5 Interpretable measurement instruments

Domain	Instrument	Psychometric properties*	Feasibility	Recommendation category
Life impact				
	SF-12	Cronbach's α coefficient (0.77)	-	B
	MCS	SCB (51.5)		
	PCS	SCB (30.1) Responsiveness: SF-12 PCS (mean change score: 8.17)		
	SF-36	Cronbach's α coefficient (0.79–0.93) Responsiveness: SF-36 (normalised change: 0.32)	-	B
	MCS	MDC or SDC (distribution: 3.3–5.7) MCID (distribution: 3.4–6.8, anchor: 3.0–7.4) Construct validity: Arm pain score (Pearson's correlation: -0.23) mJOA scale (Pearson's correlation: 0.19) NDI (Spearman's rank correlation: -0.17) Neck pain score (Pearson's correlation: -0.28) SF-12 PCS (Pearson's correlation: 0.01) Responsiveness: SF-36 MCS (effect size range: 0.81, sensitivity: 0.67)		
	PCS	MDC or SDC (distribution: 5.2–5.7, anchor: 4.9) MCID (distribution: 2.9–5.5, distribution: 10, anchor: 3.9–9.6) SCB ¹⁶ Criterion validity (Likert scale): AUC: 0.67–0.69 Construct validity: Arm pain score (Pearson's correlation: -0.44) mJOA scale (Pearson's correlation: 0.43) NDI (Spearman's rank correlation: -0.49) Neck pain score (Pearson's correlation: -0.41) SF-12 PCS (Pearson's correlation: -0.29) Responsiveness: SF-36 PCS (effect size range: 0.84, sensitivity: 0.85)		
Life impact and neuromuscular function				
	JOACMEQ	Patient comprehensibility: 'No questions elicited no answer or 'I am not sure' in more than 5% of patients' Test-retest stability: Cronbach's α coefficient (0.91) Forward-backward translation (Persian and Thai): n/a	+	B
	Bladder function	Cronbach's α coefficient (0.32–0.74) Test-retest stability: ICC (0.62) MDC or SDC (distribution: 7.7) MCID (anchor: 6.0) Responsiveness: JOACMEQ bladder function (AUC: 0.82, effect size: 0.33, mean change score: 18.0)		
	Cervical spine function	Cronbach's α coefficient (0.77–0.78) Test-retest stability: ICC (0.63) MDC or SDC (distribution: 12.9, anchor: 12.5) MCID (anchor: 2.5) Criterion validity (Likert scale): AUC: 0.58 Responsiveness: JOACMEQ cervical spine function (AUC: 0.72, Effect size: 0.28, Mean change score: 25.8)		

Continued

Table 5 Continued

Domain	Instrument	Psychometric properties*	Feasibility	Recommendation category
	Lower extremity function	Cronbach's α coefficient (0.80–0.86) Test–retest stability: ICC (0.83) MDC or SDC (distribution: 6.6, anchor: 8.5) MCID (anchor: 8.5–9.5) Criterion validity (Likert scale): AUC: 0.66–0.70 Construct validity: NDI (Pearson's correlation: –0.66) SF-12 MCS (Spearman's rank correlation: 0.40) SF-12 PCS (Spearman's rank correlation: 0.29) Responsiveness: JOACMEQ quality of life (AUC: 0.83, effect size: 0.46, mean change score: 23.7)		
	QOL	Cronbach's α coefficient (0.80–0.86) Test–retest stability: ICC (0.83) MDC or SDC (distribution: 6.6, anchor: 8.5) MCID (anchor: 8.5–9.5) Criterion validity (Likert scale): AUC: 0.66–0.70 Construct validity: NDI (Pearson's correlation: –0.66) SF-12 MCS (Spearman's rank correlation: 0.40) SF-12 PCS (Spearman's rank correlation: 0.29) Responsiveness: JOACMEQ quality of life (AUC: 0.83, effect size: 0.46, mean change score: 23.7)		
	Upper extremity function	Cronbach's α coefficient (0.72–0.74) Test–retest stability: ICC (0.93) MDC or SDC (distribution: 9.5, anchor: 6.1) MCID (anchor: 2.5–13.0) Responsiveness: JOACMEQ upper extremity function (AUC: 0.74, effect size: 0.17, mean change score: 10.7)		
Neuromuscular function				
	JOA	Cronbach's α coefficient (0.72) Forward–backward translation (Brazilian Portuguese): Comprehension rate (>81.2%) Interobserver reliability: ICC (0.81) MDC or SDC (distribution: 1.0, anchor: 2.5) LOA (1.2(–1.2 to 3.6)) MCID (anchor: 2.5) Criterion validity (Likert scale): AUC: 0.59–0.62 Construct validity: JOACMEQ QOL (Spearman's rank correlation: 0.41) mJOA (Spearman's rank correlation: 0.87) NDI (Spearman's rank correlation: –0.50 to –0.76) SF-12 MCS (Spearman's rank correlation: –0.05) SF-12 PCS (Spearman's rank correlation: 0.50) Responsiveness: JOA (mean change score: 4.6, normalised change: 0.21) JOA motor function of lower extremity (mean change score: 0.60) mJOA (Spearman's rank correlation: 0.75)	–	B
	Bladder function	Intraobserver reliability ($\kappa=0.64$) Interobserver reliability ($\kappa=0.47$)		
	Motor function of fingers	Intraobserver reliability ($\kappa=0.68$) Interobserver reliability ($\kappa=0.53$)		

Continued

Table 5 Continued

Domain	Instrument	Psychometric properties*	Feasibility	Recommendation category
	Motor function of shoulder and elbow	Intraobserver reliability ($\kappa=0.50$) Interobserver reliability ($\kappa=0.31$)		
	Motor function of lower extremity	Intraobserver reliability ($\kappa=0.55$) Interobserver reliability ($\kappa=0.49$)		
	Sensory function of lower extremity	Intraobserver reliability ($\kappa=0.54$) Interobserver reliability ($\kappa=0.58$)		
	Sensory function of upper extremity	Intraobserver reliability ($\kappa=0.51$) Interobserver reliability ($\kappa=0.42$)		
	mJOA	Cronbach's α coefficient (0.60–0.63) Forward-backward translation (Brazilian Portuguese and Italian): n/a Test-retest stability (Spearman's rank correlation: 0.91) Intraobserver reliability (ICC: 0.87) Interobserver reliability (ICC: 0.97, $\kappa=0.80$) MDC or SDC (distribution: 2.1) MCID (distribution: 1.2–1.4, anchor: 1.3–3.1) SCB ¹⁴ Criterion validity (Nurick scale): Spearman's rank correlation: –0.41 Pearson's correlation: –0.62 to –0.63 Construct validity: 30MWT (Pearson's correlation: –0.38) EQ-5D (Spearman's rank correlation: 0.42) JOACMEQ QOL (Spearman's rank correlation: 0.41) NDI (Spearman's rank correlation: –0.51, Pearson's correlation: –0.33 to –0.34) SF-12 MCS (Pearson's correlation: 0.03) SF-12 PCS (Pearson's correlation: 0.42) SF-36 MCS (Pearson's correlation: 0.25) SF-36 PCS (Pearson's correlation: 0.30) Responsiveness: mJOA (effect size: 0.87–1.0, normalised change: 1.47)	–	C
	Motor dysfunction of lower extremities	Interobserver reliability (ICC: 0.73) Criterion validity (Nurick scale): Pearson's correlation: –0.65 to –0.68 Construct validity: 30MWT (Pearson's correlation: –0.43) NDI (Pearson's correlation: –0.31) SF-36 MCS (Pearson's correlation: 0.21) SF-36 PCS (Pearson's correlation: 0.31–0.50)		
	Motor dysfunction of upper extremities	Interobserver reliability (ICC: 0.77) Criterion validity (Nurick scale): Pearson's correlation: –0.42 Construct validity: 30MWT (Pearson's correlation: –0.21) NDI (Pearson's correlation: –0.24) SF-36 MCS (Pearson's correlation: 0.20) SF-36 PCS (Pearson's correlation: 0.22)		
	Sensory dysfunction of sphincter dysfunction	Interobserver reliability (ICC: 0.78) Criterion validity (Nurick scale): Pearson's correlation: –0.25 Construct validity: 30MWT (Pearson's correlation: –0.23) NDI (Pearson's correlation: –0.16) SF-36 MCS (Pearson's correlation: 0.08) SF-36 PCS (Pearson's correlation: 0.06)		

Continued



Table 5 Continued

Domain	Instrument	Psychometric properties*	Feasibility	Recommendation category
	Sensory dysfunction of upper extremities	Interobserver reliability (ICC: 0.93) Criterion validity (Nurick scale): Pearson's correlation: -0.23 Construct validity: 30MWT (Pearson's correlation: -0.05) NDI (Pearson's correlation: -0.23) SF-36 MCS (Pearson's correlation: 0.19) SF-36 PCS (Pearson's correlation: 0.19)		
Pain				
	NDI	MDC or SDC (distribution: 6.2%, anchor: 5.2%) MCID (anchor: 5–13) SCB (anchor: 9.5–36) Criterion validity (Likert scale): AUC: 0.66–0.75 Construct validity: Arm pain score (Pearson's correlation: 0.68) mJOA (Pearson's correlation: -0.36) Neck pain score (Pearson's correlation: 0.64) SF-12 MCS (Pearson's correlation: -0.40) SF-12 PCS (Pearson's correlation: -0.54) Responsiveness: Anchor (AUC: 0.66) NDI (mean change score: -15.8)	+	B
	Pain, 'Numeric rating scale' (arm pain score)	MCID (anchor: 2.5) SCB (3.5) Construct validity: mJOA (Pearson's correlation: -0.19) Neck pain score (Pearson's correlation: 0.72)	-	B
	Pain, 'Numeric rating scale' (neck pain scores)	MCID (anchor: 2.5) SCB (3.5) Construct validity: mJOA (Pearson's correlation: -0.07)	-	B
	VAS for pain	MDC or SDC (distribution: 3.1) MCID (distribution: 24.0–30.0, anchor: 0.4–2.7) SCB (1.1)	+	B

n/a=No info available

Feasibility: +=No barriers;+=Minimal barriers; -=Barriers

Interpretability: +=Interpretable; -=Uninterpretable, due to absence of anchor-based MCIDs^{23–25}

Recommendation category: A=measurement instruments with evidence for sufficient content validity (any level) AND at least low-quality evidence for sufficient internal consistency; B=Measurement instruments categorised not in A or C; C=measurement instruments with high-quality evidence for an insufficient measurement property.

*Comparators shown as indented tools

AUC, area under curve; EQ-5D, EuroQol-5 Dimension; JOA, Japanese Orthopaedic Association; JOACMEQ, Japanese Orthopaedic Association Cervical Myelopathy Evaluation Questionnaire; LOA, limits of agreement; MCID, minimal clinically important difference; MCS, mental component summary; MDC, minimal detectable change; mJOA, modified Japanese Orthopaedic Association; 30MWT, 30-m Walking Test; NDI, Neck Disability Index; PCS, physical component summary; SCB, substantial clinical benefit; SDC, smallest detectable change; SF-12, 12-Item Short Form Health Survey; SF-36, 36-Item Short Form Health Survey.

this review, possibly because this is a more recent and complex criterion, or because of the search or selection criteria. Similarly, studies on content validity cannot score higher than 'inadequate' if there are no recordings/verbatim transcriptions of patient focus groups or interviews. Likewise, analyses of reliability cannot score higher than 'doubtful' if statistics other than the Pearson or Spearman correlation coefficients are used. These thresholds of acceptability may account for some of the lacking information and are an important entry challenge for instruments into DCM research—a field where the routine involvement of stakeholders with lived experience is at an

early stage,^{3 8} inconsistent study reporting is prevalent,^{2 4} few studies have involved >100 patients, and where there is a bias in the availability of measurement literature (ie, some tools, such as the SF-12, are used because they are the only tools available and, therefore, have available literature due to their routine use). From the application of these COSMIN criteria in other research fields, however, it appears that these methodological deficiencies are not exclusive to DCM instruments, including those in current use.^{40–42} The lack of high-quality assessments, thus, should not necessarily imply that (1) the identified

outcome measures are generally inadequate, or (b) that the COSMIN standards are not fit for the DCM context.

Measurement rigour is universally important and, in DCM, particularly relevant as the development of new instruments is a top 10 research priority. This rank reinforces the decision of the steering committee to make the initial CMS recommendations based existing on tools, rather than on tools under development.²⁶ This decision was taken recognising that the success of a CMS requires widespread adoption, and that the adoption of clinical recommendations can be challenging without stakeholder awareness, familiarity and/or confidence.^{43–46} We hypothesised that asking the global field to align with new innovations would be more challenging, and premature, at this stage. Thus, for this first iteration of the DCM CMS, there is a focus on current instruments in academic usage. While, currently, few have met the bar set by the COSMIN methodology, there are nine reasonable candidates using our post-hoc thresholds (table 5). Ultimately, the CMS process will need to lean significantly on the expertise of those involved in the consensus phase in order to make final recommendations that are methodologically rigorous and representative of those with lived experience.

Despite its conscientious design, this systematic review has limitations. In searching for existing instruments, we have neither identified nor assessed tools under development, or those currently being translated into clinical or research settings or published in languages other than English. To the extent that DCM instruments are currently in use, however, this review only identified tools in four of the six core domains from RECODE-DCM's minimum dataset,³³ and did not consider the construct of the disease as a factor in evaluating the outcomes. For those missing outcomes, focused scoping reviews (informed by a gap analysis that will be published separately) will be conducted in the future. Next, clinician-reported outcome measures and performance-based outcome measures were analysed with the exact same methods as PROMs. While COSMIN explicitly allows this,^{23–25} methods may be differentially adapted to tailor to these distinct instrument types; we chose not to do so out of prudence and consistency, and results across these instrument groups should be interpreted accordingly. Feasibility and interpretability were also evaluated using bespoke criteria which, despite being adapted from the COSMIN methodology, may not weigh all criteria accurately. Importantly, our decision to shortlist the clinically interpretable instruments was made a posteriori due to the unexpected absence of category A recommendations. This decision was informed by our judgement that instruments in a CMS should be interpretable by clinicians and offer qualitative meaning to clinicians and people with lived experience. While the COSMIN taxonomy does indeed class interpretability as an important and standalone characteristic,^{23–25} the aforementioned shortlist may inevitably represent a placement bias. Notably, some nuances of different versions of measurement instruments

(eg, mJOA) were not extensively evaluated.⁴⁷ Lastly, and as is frequently the case in this body of reviews,^{40–42} none of the authors is specifically trained in measurement theory and, therefore, this work represents our best attempt to implement the guidelines and standards set forward by the COSMIN methodology in the context of DCM.

CONCLUSIONS

Currently, none of the measurement instruments used in DCM holds sufficient evidence to meet the COSMIN criteria for a strong recommendation for use. However, there are leading contenders that appear to offer qualitative meaning to clinicians and people with lived experience of DCM; namely, the SF-12 and SF-36; JOA, mJOA, and JOACMEQ; NDI; and VAS for pain. The findings of this review will inform a consensus process to form a CMS for DCM. As the development of new assessments for DCM is an active research priority, greater awareness of the COSMIN framework is pertinent to DCM researchers.

Author affiliations

¹School of Medical Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK

²School of Clinical Medicine, University of Cambridge, Cambridge, UK

³Bristol Medical School, Faculty of Health Sciences, University of Bristol, Bristol, UK

⁴Department of Medicine, Division of Physical Medicine and Rehabilitation, University of Toronto, Toronto, Ontario, Canada

⁵Division of Physical Medicine and Rehabilitation, Toronto Rehabilitation Institute, University Health Network, Toronto, Ontario, Canada

⁶KITE Research Institute, University Health Network, Toronto, Ontario, Canada

⁷Division of Neurosurgery and Spinal Program, Toronto Western Hospital, University of Toronto, Toronto, Ontario, Canada

⁸Thomas Jefferson University, Jefferson Health System, St Louis, Philadelphia, USA

⁹Spinal Cord Injury Center, Balgrist University Hospital, Zurich, Switzerland

¹⁰Spinal Unit (UVM), Department of Orthopaedics, Centro Hospitalar Universitário do Porto, Porto, Portugal

¹¹Instituto de Ciências Biomédicas Abel Salazar, Porto, Portugal

¹²Department of Family Medicine, McMaster University, Hamilton, Ontario, Canada

¹³Myelopathy.org, Cambridge, UK

¹⁴Academic Department of Neurological Surgery, Sina Trauma and Surgery Research Center, Tehran University of Medical Sciences, Tehran, Iran

¹⁵Division of Neurosurgery, University of Maryland School of Medicine, Baltimore, Maryland, USA

¹⁶Department of Neurosurgery, Medical College of Wisconsin, Milwaukee, Wisconsin, USA

¹⁷Toronto Western Hospital, University of Toronto, Toronto, Ontario, Canada

¹⁸Krembil Research Institute, Toronto, Ontario, Canada

¹⁹Department of Neurological Surgery and The Miami Project to Cure Paralysis, University of Miami Miller School of Medicine, Miami, Florida, USA

²⁰Centre for Surgical Research, Bristol Medical School: Population Health Sciences, University of Bristol, Bristol, UK

²¹GI Surgery, North Bristol NHS Trust, Bristol, UK

²²Department of Clinical Neurosurgery, University of Cambridge, Cambridge, UK

²³Department of Clinical Neurosciences, Ann McLaren Laboratory of Regenerative Medicine, Cambridge, UK

²⁴Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK

Twitter Alvaro Yanez Touzet @AYanezTouzet and Angus GK McNair @angusgkmcnair

Acknowledgements We thank Isla Khun for her assistance with the systematic search and all patients who advised, and continue to advise, the AO Spine RECODE-DCM initiative.

Contributors BD was responsible for conceiving the article and is the guarantor. KSL conducted the search and AYT, AB, ED and FB conducted the screening. AYT

and BD extracted and analysed the data and wrote the manuscript. AYT, JCF, MGF, JSH, CMZ, RR-P, JM, ES, AC, VR-M, BA, TFB, LT, RC, JDG, SK-R, AGKM, MK and BD provided critical appraisal of the manuscript. All authors critically revised and approved the manuscript.

Funding This work was supported by AO Spine through the AO Spine Knowledge Forum Spinal Cord Injury, a focused group of international Spinal Cord Injury experts. AO Spine is a clinical division of the AO Foundation, which is an independent medically guided not-for-profit organisation. Study support was provided directly through the AO Spine Research Department. An award/grant number is not applicable.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study does not involve human participants.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Alvaro Yanez Touzet <http://orcid.org/0000-0001-9309-1885>

Faheem Bhatti <http://orcid.org/0000-0003-3897-4196>

Keng Siang Lee <http://orcid.org/0000-0003-2308-0579>

Carl Moritz Zipser <http://orcid.org/0000-0002-4396-4796>

Ricardo Rodrigues-Pinto <http://orcid.org/0000-0002-6903-348X>

Timothy F Boerger <http://orcid.org/0000-0003-1587-3704>

Angus GK McNair <http://orcid.org/0000-0002-2601-9258>

REFERENCES

- Fehlings MG, Tetreault LA, Riew KD, *et al*. A clinical practice guideline for the management of patients with degenerative cervical myelopathy: recommendations for patients with mild, moderate, and severe disease and Nonmyelopathic patients with evidence of cord compression. *Global Spine J* 2017;7:70S–83.
- Davies BM, McHugh M, Elgheriani A, *et al*. Reported outcome measures in degenerative cervical myelopathy: a systematic review. *PLoS One* 2016;11:e0157263.
- Davies BM, Khan DZ, Mowforth OD, *et al*. RE-CODE DCM (REsearch Objectives and Common Data Elements for Degenerative Cervical Myelopathy): A Consensus Process to Improve Research Efficiency in DCM, Through Establishment of a Standardized Dataset for Clinical Research and the Definition of the Research Priorities. *Global Spine J* 2019;9:65S–76.
- Davies BM, McHugh M, Elgheriani A, *et al*. The reporting of study and population characteristics in degenerative cervical myelopathy: a systematic review. *PLoS One* 2017;12:e0172564-e.
- Davies BM, Mowforth OD, Smith EK, *et al*. Degenerative cervical myelopathy. *BMJ* 2018;360:k186.
- Spine AO. Ao spine RECODE-DCM: research objectives and common data elements for degenerative cervical myelopathy, 2021. Available: www.aospine.org/recode
- Davies B, Mowforth O, Sadler I, *et al*. Recovery priorities in degenerative cervical myelopathy: a cross-sectional survey of an international, online community of patients. *BMJ Open* 2019;9:e031486.
- Boerger TF, Davies BM, Sadler I, *et al*. Patient, sufferer, victim, casualty or person with cervical myelopathy: let us decide our identifier. *Integrated Healthcare Journal* 2020;2:e000023.
- Clarke M. Standardising outcomes for clinical trials and systematic reviews. *Trials* 2007;8:39.
- Kirkham JJ, Dwan KM, Altman DG, *et al*. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ* 2010;340:c365.
- Kirkham JJ, Gargon E, Clarke M, *et al*. Can a core outcome set improve the quality of systematic reviews?—a survey of the Coordinating Editors of Cochrane Review Groups. *Trials* 2013;14:21.
- Prinsen CAC, Vohra S, Rose MR, *et al*. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" - a practical guideline. *Trials* 2016;17:449.
- Williamson PR, Altman DG, Bagley H, *et al*. The comet Handbook: version 1.0. *Trials* 2017;18:280.
- Boers M, Kirwan JR, Wells G, *et al*. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. *J Clin Epidemiol* 2014;67:745–53.
- Schmitt J, Apfelbacher C, Spuls PI, *et al*. The Harmonizing outcome measures for eczema (home) roadmap: a methodological framework to develop core sets of outcome measurements in dermatology. *J Invest Dermatol* 2015;135:24–30.
- Kirkham JJ, Davis K, Altman DG, *et al*. Core outcome Set-STAndards for development: the COS-STAD recommendations. *PLoS Med* 2017;14:e1002447.
- Kirkham JJ, Gorst S, Altman DG, *et al*. Core outcome Set-STAndards for reporting: the COS-STAR statement. *PLoS Med* 2016;13:e1002148.
- Potter S, Davies C, Holcombe C, *et al*. International development and implementation of a core measurement set for research and audit studies in implant-based breast reconstruction: a study protocol. *BMJ Open* 2020;10:e035505.
- Grieve S, Perez RSGM, Birklein F, *et al*. Recommendations for a first core outcome measurement set for complex regional pain syndrome clinical sTudies (compact). *Pain* 2017;158:1083–90.
- Davies CF, Macefield R, Avery K, *et al*. Patient-Reported outcome measures for post-mastectomy breast reconstruction: a systematic review of development and measurement properties. *Ann Surg Oncol* 2021;28:386–404.
- Kalsi-Ryan S, Singh A, Massicotte EM, *et al*. Ancillary outcome measures for assessment of individuals with cervical spondylotic myelopathy. *Spine* 2013;38:S111–22.
- Mokkink LB, Terwee CB, Knol DL, *et al*. Protocol of the COSMIN study: consensus-based standards for the selection of health measurement instruments. *BMC Med Res Methodol* 2006;6:2.
- Prinsen CAC, Mokkink LB, Bouter LM, *et al*. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res* 2018;27:1147–57.
- Mokkink LB, de Vet HCW, Prinsen CAC, *et al*. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Qual Life Res* 2018;27:1171–9.
- Terwee CB, Prinsen CAC, Chiarotto A, *et al*. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Qual Life Res* 2018;27:1159–70.
- Davies BM, Yanez Touzet A, Mowforth OD. Development of a core measurement set for research in degenerative cervical myelopathy: a study protocol (AO Spine RECODE-DCM CMS). *medRxiv* 2021. doi:10.1101/2021.11.11.21266170
- Davies BM, Goh S, Yi K, *et al*. Development and validation of a Medline search filter/hedge for degenerative cervical myelopathy. *BMC Med Res Methodol* 2018;18:73.
- Khan DZ, Khan MS, Kotter MR, *et al*. Tackling research inefficiency in degenerative cervical myelopathy: illustrative review. *JMIR Res Protoc* 2020;9:e15922p <http://europepmc.org/abstract/MED/32525490https://doi.org/10.2196/15922https://europepmc.org/articles/PMC7317636>
- Khan MA, Mowforth OD, Kuhn I, *et al*. Development of a validated search filter for Ovid Embase for degenerative cervical myelopathy. *Health Info Libr J*;12.
- Singh A, Tetreault L, Casey A, *et al*. A summary of assessment tools for patients suffering from cervical spondylotic myelopathy: a systematic review on validity, reliability and responsiveness. *Eur Spine J* 2015;24 Suppl 2:209–28.
- Campbell M, McKenzie JE, Sowden A, *et al*. Synthesis without meta-analysis (swim) in systematic reviews: reporting guideline. *BMJ* 2020;368:l6890.
- Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.
- AO Spine. Minimum dataset, 2021. Available: <https://aospine.aofoundation.org/research/recode-dcm/minimum--dataset>
- de Vet HCW, Terwee CB, Mokkink LB. *Measurement in medicine: a practical guide*. Cambridge: Cambridge University Press, 2011.

- 35 Longo UG, Berton A, Denaro L, *et al.* Development of the Italian version of the modified Japanese orthopaedic association score (mJOA-IT): cross-cultural adaptation, reliability, validity and responsiveness. *Eur Spine J* 2016;25:2952–7.
- 36 de Vet HCW, Terwee CB. The minimal detectable change should not replace the minimal important difference. *J Clin Epidemiol* 2010;63:804–5.
- 37 de Vet HCW, Ostelo RWJG, Terwee CB, *et al.* Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. *Qual Life Res* 2007;16:131–42.
- 38 de Vet HC, Terwee CB, Ostelo RW, *et al.* Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health Qual Life Outcomes* 2006;4:54.
- 39 Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol* 2003;56:395–407.
- 40 Francis DO, McPheeters ML, Noud M, *et al.* Checklist to operationalize measurement characteristics of patient-reported outcome measures. *Syst Rev* 2016;5:129.
- 41 Francis DO, Daniero JJ, Hovis KL, *et al.* Voice-Related patient-reported outcome measures: a systematic review of instrument development and validation. *J Speech Lang Hear Res* 2017;60:62–88.
- 42 Terwee CB, de Vet HC, Prinsen CAC. Comment on “Checklist to operationalize measurement characteristics of patient-reported outcome measures. Available: <https://www.cosmin.nl/wp-content/uploads/Letter-comment-on-Francis.pdf>
- 43 Kirkham JJ, Clarke M, Williamson PR. A methodological approach for assessing the uptake of core outcome sets using ClinicalTrials.gov: findings from a review of randomised controlled trials of rheumatoid arthritis. *BMJ* 2017;357:j2262.
- 44 Bauer MS, Kirchner J. Implementation science: what is it and why should I care? *Psychiatry Res* 2020;283:112376.
- 45 Gupta DM, Boland RJ, Aron DC. The physician's experience of changing clinical practice: a struggle to unlearn. *Implement Sci* 2017;12:28.
- 46 Braithwaite J, Churrua K, Long JC, *et al.* When complexity science meets implementation science: a theoretical and empirical analysis of systems change. *BMC Med* 2018;16:63.
- 47 Furlan JC, Catharine Craven B. Psychometric analysis and critical appraisal of the original, revised, and modified versions of the Japanese orthopaedic association score in the assessment of patients with cervical spondylotic myelopathy. *Neurosurg Focus* 2016;40:E6.
- 48 Mokkink LB, Terwee CB, Patrick DL, *et al.* The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010;63:737–45.
- 49 van Kampen DA, Willems WJ, van Beers LWAH, *et al.* Determination and comparison of the smallest detectable change (SDC) and the minimal important change (MIC) of four-shoulder patient-reported outcome measures (PROMs). *J Orthop Surg Res* 2013;8:40.
- 50 Nwachukwu BU, Beck EC, Kunze KN, *et al.* Defining the clinically meaningful outcomes for arthroscopic treatment of femoroacetabular impingement syndrome at minimum 5-year follow-up. *Am J Sports Med* 2020;48:901–7.
- 51 Streiner DL, Norman GR, Cairney J. Health measurement Scales A practical guide to their development and use: a practical guide to their development and use: Oxford university press; 2015 2015:01.