BMJ Open

Systematic Review of the Measurement Properties of Performance-based Functional Tests in Patients with Neck Disorders

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-031242
Article Type:	Research
Date Submitted by the Author:	24-Apr-2019
Complete List of Authors:	McGee, Steven; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Sipos, Taylor; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Alin, Thomas; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Chen, Celia; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Greco, Alexandra; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Bobos, Pavlos; Western University, Health and Rehabilitation Sciences; University of Toronto, Dalla Lana School of Public Health, Institute of Health Policy Management and Evaluation MacDermid, Joy ; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Group, CATWAD; Michele Sterling m.sterling@uq.edu.au, Anne Söderlund anne.soderlund@mdh.se, Michele Curatolo, curatolo@uw.edu, Jim Elliott j-elliott@northwestern.edu, David Walton dwalton5@uwo.ca, Helge Kasch helgkasc@rm.dk, Carroll, Linda linda.carroll@ualberta.ca Hans.Westergren@skane.se, McLean, Samuel A Samuel_McLean@med.unc.edu, Gwendolen Jull g.jull@uq.edu.au, Genevieve Grant genevieve.grant@monash.edu Luke Connelly l.connelly@uq.edu.au, MacDermid, Joy jmacderm@uwo.ca, Mandy Nielsen mandy.nielsen@gri
Keywords:	functional, psychometric properties, neck pain, cervical, outcome measures

Title: Systematic Review of the Measurement Properties of Performance-based Functional Tests in Patients with Neck Disorders

[1]Steven McGee, PT

[2]Taylor Sipos, PT

[3]Thomas Allin, PT

[4]Celia Chen, PT

[5]Alexandra Greco, PT

[6]Pavlos Bobos, PT, PhD(c) (corresponding author)

[7]Joy MacDermid, PT, PhD

[8]CATWAD

Authors' information

[1]Steven McGee PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (smcgee7@uwo.ca)

[2]Taylor Sipos PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (jsipos@uwo.ca)

[3]Thomas Allin PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (tallin@uwo.ca)

[4]Celia Chen PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (qchen224@uwo.ca)

[5]Alexandra Greco PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (agreco33@uwo.ca)

[6]Pavlos Bobos PT, PhD(c), (corresponding author) Doctoral Candidate, Western's Bone and Joint Institute, Department of Health and Rehabilitation Sciences, Western University, Elborn College, 1201 Western Road, N6G 1H1, London, Ontario, Dalla Lana School of Public Health, Institute of

Health Policy Management and Evaluation, Department of Clinical Epidemiology, University of Toronto, Canada, (pbobos@uwo.ca), tel: +1 519 661 2111 x88912

[7]Joy C MacDermid BScPT, PhD, Professor, Physical Therapy and Surgery, Western University, London, ON and Co-director Clinical Research Lab, Hand and Upper Limb Centre, St. Joseph's Health Centre, London, Ontario; Professor Rehabilitation Science McMaster University, Hamilton, ON, Canada (jmacderm@uwo.ca)

[5]CATWAD Coauthors: Michele Sterling m.sterling@uq.edu.au, Anne Söderlund anne.soderlund@mdh.se, Michele Curatolo, curatolo@uw.edu, Jim Elliott j-elliott@northwestern.edu, David Walton dwalton5@uwo.ca, Helge Kasch helgkasc@rm.dk, Carroll, Linda linda.carroll@ualberta.ca Hans.Westergren@skane.se, McLean, Samuel A Samuel_McLean@med.unc.edu, Gwendolen Jull g.jull@uq.edu.au, Genevieve Grant genevieve.grant@monash.edu, Luke Connelly l.connelly@uq.edu.au, MacDermid, Joy jmacderm@uwo.ca, Mandy Nielsen mandy.nielsen@griffith.edu.au, Pierre Cote pierre.cote@uoit.ca, Tonny Elmose Andersen tandersen@health.sdu.dk, Trudy Rebbeck trudy.rebbeck@sydney.edu.au, Annick Maujean a.maujean@uq.edu.au, Sarah Robins s.robins1@uq.edu.au, Kenneth Chen k.chen8@uq.edu.au, Julia Treleaven j.treleaven@uq.edu.au

Key Words: functional, psychometric properties, neck, cervical, outcome measures

Word Count: 4239

Abstract

Objective: The purpose of this systematic review is to identify and synthesize studies evaluating performance-based outcome measures designed to evaluate the functional abilities of patients with mechanical neck pain.

Setting: Not applicable

Participants: Participants with neck disorders

Methods: A literature search using PubMed, Scopus, CINAHL, Embase, COCHRANE, Google Scholar, and a citation mapping strategy was conducted through June 2018. Selected articles were appraised using the COSMIN risk of bias checklist tool and the Quality Appraisal for Clinical Measurement Research Reports Evaluation Form (QACMRR). Relevant data were then extracted from selected articles using an extraction guide.

Results: The search obtained 12 articles which reported on 4 outcome measures reporting to assess the functional abilities in patients with mechanical neck pathology. Of the selected papers: 1 reports content validity, 5 construct validity, 4 reliability, 1 sensitivity to change, and 1 both reliability and construct validity. COSMIN sub-scores ranged from "inadequate" to "very good" and QACMRR scores ranged from 68% to 95%.

Conclusions: A limited number of performance-based tests have been developed or validated for assessing neck function. The pool of research in this area is sparse and insufficient to make conclusive recommendations.

Prospero registration: CRD42018112358

Strengths and limitations of this study

- The psychometric properties of performance outcome measures for neck pain were synthesized and critically appraised
- This study assessed the risk of bias and the quality of measurements properties
- The feasibility or usability of these tools was not assessed

Introduction

Neck pain has been associated with high disability and is regarded as a substantial societal burden. (1,2) Approximately 70% of people experience neck pain within their lifetime and about 33% of adults experience neck pain every year. (3,4) Further concern is warranted as it has been suggested that the incidence of neck pain is increasing. (5,6,7) The economic burden due to neck disorders is high, including lost wages, costs of treatment, and compensation expenditures to injured people. (8,9) Neck pain is second only to low back pain in annual workers' compensation costs in the United States.(7)

Outcome measures are a crucial component in monitoring patients with neck pain to determine the effects of treatment, evaluation of interventions, guiding return to work, and justifying treatment. Several self-reported outcome measures currently exist to assess disability and function in those with neck pain (e.g. the Neck Disability Index (NDI) or the numeric pain rating scale (NPRS). (10) Evidence-based clinical practice guidelines suggest that measures assessing physical performance should also be used for people with neck pain. (11) Performance-based testing is where the assessment is based on actual performance of a task or activity. Physical performance can be assessed by testing a person's ability to execute a standardized activity in a standardized environment (i.e. clinical setting). (12) Time to complete the activity, number of repetitions performed, and weight lifted are frequently used to quantify

the physical performance. (13) Conversely, self-report measures examine patients' perception and experience of their ability to perform functional tasks. (12) Previous research has demonstrated poor to fair relationships between physical performance and self-report measures of ability in patients with various musculoskeletal disorders suggesting that these measures assess different constructs of function. (13,14) Consequently, physical performance tests and self-report measures complement each other and may each contribute unique information about a patient's function. (15)

A fundamental component of monitoring outcomes is having reliable and valid tools with known measurement properties. (16,17) While recent research has investigated the psychometric properties of patient-reported outcomes in people with neck pain (1,10, 18,19,20) there is a gap in knowledge with respect to performance-based functional outcomes. The purpose of this systematic review was to identify and synthesize clinical measurement studies that evaluate psychometric properties of performance-based functional tests in patients with neck disorders.

METHODS

Patient and Public Involvement

No patient involved

Study Design and Protocol Registration

We conducted a systematic review to evaluate the psychometric properties of performance-based functional tests for people with mechanical neck disorders. The protocol was registered in PROSPERO register with registration number CRD42018112358.

Search Strategy

A database search using CINAHL, PubMed, Scopus and Google Scholar was performed to identify articles published before July 2018. The following search strategy was used to search all databases for eligible studies: (Reliability OR validity OR responsiveness OR calibration OR validation OR (minimal detectable change) OR (clinically important difference) OR (psychometric properties) AND cervical OR neck OR c-spine AND (performance measure) OR (functional test) OR (functional outcome) OR (performance outcome)). A citation map of articles and systematic reviews selected for the full-text review was performed. This strategy was included to minimize the risk of publication bias. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) process (21) was followed to ensure all appropriate steps were taken in the selection process (**FIGURE 1**).

Inclusion and Exclusion Criteria

Articles were included in the final review if all of the following criteria were met: 1) >50% of the study's patient population had neck pain or a musculoskeletal neck disorder 2) Patients in the study completed a functional-based test 3) Clinometric properties of at least one performance-based test were reported. Definitions for the properties can be found in **APPENDIX A.**

Article Selection

Titles and abstracts generated by the search strategy were screened by two authors independently. Articles that met the inclusion criteria and selected for a full text review were also

reviewed in pairs of authors. Disagreements were resolved by the most experienced author (JCM)

Data Extraction

Data extraction and critical appraisal was performed in pairs of two raters among the authors, after the completion of a calibration session. When reviewers disagreed during data extraction and/or critical appraisal, and consensus could not be met, a third author arbitrated. A data extraction form (17) (**APPENDIX A and APPENDIX B**), developed by one of the authors (JCM.), was used to ensure systematicity. Authors extracted sample size, patient population characteristics, functional tests performed and reported psychometric properties.

Risk of Bias and Quality Assessment

Two authors used the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) (22) checklist to assess risk of bias in the articles selected for publication. The COSMIN checklist was recently adapted to evaluate risk of bias in studies on measurement properties of patient reported outcome measures (PROMs). (22) After completing a calibration session, each article was scored on the 4-point scale as “very good”, “adequate”, “doubtful” or “inadequate” for each of the checklist criteria for relevant measurement properties (e.g. reliability, responsiveness, etc.). To determine the overall score for each measurement property, the worst score counts method was used wherein the lowest score for the checklist criteria of the relevant property was taken as the overall score. (23) Pairs of authors critically appraised the quality of each study using a standardized 12-item evaluation tool (QACMRR) designed to assess the quality of studies determining measurement properties in

outcome measures (**APPENDIX C**). (24) Total scores on the tool can range from 0 to 24, with a higher score indicating higher quality. Scores can be normalized to range between 0-100%. This tool has been found to have good to excellent pre-consensus inter-rater reliability (ICC: 0.69-0.91) across a number of systematic reviews. (17,24-28) Raw scores were converted to standardized percentage scores and ranked based on percentage values. There were no formal mechanisms developed to weight the studies based on quality scores.

RESULTS

The search strategy resulted in 840 published articles. After duplications were removed, 31 articles were deemed relevant and were screened at full text. Overall, 12 articles met our inclusion criteria (**FIGURE 1**). The characteristics of the included studies and the summary of psychometric properties are presented in **TABLE 1.** The risk of bias and the quality assessment is summarized and presented in **TABLE 2-3**. The 12 articles that were included for review provided properties on the following performance based tests: Functional Capacity Evaluations (FCE) (29,30,31,32,33,34), The Baltimore Therapeutic Equipment Work Simulator II (BTEWS II) (35), Functional Impairment Test- Hand and Neck/Shoulder/Arm (FIT-HaNSA) (36), as well as items off of a physiotherapy test package including a cervical and lumbar Progressive Isoinertial Lifting Evaluation (PILE-C, PILE-L) test (37,38,39,40) and 2 x 20 m with burden walking test (2x20M-WWB) (37,38,39,40). Descriptions of all performance-based tests and their relevant subtasks are provided in **APPENDIX D.**

FCE

Six articles reported measurement properties for a FCE battery. We identified multiple versions of the FCE in the literature with one article reporting properties on the Workwell FCE (30), two reporting on the Whiplash Associated Disorder (WAD) FCE (29,31) and three reporting on the neck-FCE. (32,33,34) These test batteries include various combinations of muscular strength, endurance and functional based tests. The measurement properties of the functional based tests used by the FCE are outlined in **TABLE 4.**

An article evaluating the Workwell FCE (30) reported convergent validity and predictive criterion validity of future work capacity in workers diagnosed with WAD I or II. Correlations between FCE sub scores and baseline work capacity ranged between r=0.06 and r=0.39. FCE subscores did not predict future work capacity at 1, 3, 6 and 12 months.

An article evaluating the WAD FCE (29) evaluated test-retest reliability and measurement error in sick listed workers diagnosed with WAD grade 1 or 2. Interclass Correlation Coefficients (ICC) ranged from 0.66 to 0.96 (moderate to excellent). Limits of agreement relative to mean performance ranged from 21 to 57% for functional based sub-tests. Another WAD FCE article (31) evaluated convergent validity and known-groups validity. FCE subscales showed small to moderate correlations with each of: pain, self-reported functional ability, self-reported disability, anxiety and depression. It was found that the FCE had known-group sex validity (males vs females) for 1 of 3 functional subtests (lifting waist-overhead) and reported significant performance differences between culture groups (german vs non-german language groups).

Reesink et al. developed an independent FCE for patients with musculoskeletal neck disorders (neck FCE). (34) They performed a review of epidemiological literature and identified four physical risk factors for work-related neck disorders and used that information to develop an

FCE consisting of eight performance-based tests. Content validity was established by following operational definitions of the risk factors when searching the literature and using current literature to provide a rationale to guide their development of the tasks comprising the FCE. Because of the unconventional methods used by this study to establish content validity, the authors of this review determined that the tools used to critically appraise other articles would be inappropriate and were given scores of N/A for the COSMIN and QACMRR. An additional article measured test-retest reliability of the subscales of the neck FCE in patients with multifactorial neck pain. (32) Test retest ICC's ranged from poor to excellent. Limits of agreement relative to mean performance range from 32.0% to 56.5% for functional based sub tests. Convergent validity was performed against the Neck Disability Index (NDI) items and total score. (33) The authors found weak to moderate Pearson correlations for the FCE sub scores to both NDI individual items and the NDI total score.

BTEWS II

Lomond and Cote reported on the reliability, measurement error, minimum detectable change (MDC) and validity of the power output (PO) task during the BTEWS II test in patients with chronic neck and shoulder pain (**TABLE 5**). (35) Test-retest reliability, measured with Spearman Rank correlations and ICC's was measured at ρ=0.37 and $ICC_{2,1}$ = 0.54, respectively. The standard error of measurement (SEM) and the minimal detectable change at 90% confidence (MDC_{90}) for the PO task were measured as 30.25 and 70.59, respectively.

Weak Spearman Rank correlations between the PO task and the NDI, Shoulder Pain and Disability Index (SPADI) and Numeric Rating Scale (NRS) for pain tests were recorded. There were no significant performance differences between control and pain groups for the PO task.

Fit-HaNSA

Pierrynowski and colleagues reported on the reliability, measurement error, MDC and validity of the Fit-HaNSA test in a sample of people with WAD II following motor vehicle collision (MVC) (**TABLE 6**). (36) Intra-rater reliability ICC's for patient subtask and total scores ranged between 0.70-0.78. (36) Inter-rater reliability ICC's for patient subtask and total scores ranged between 0.54-0.84. (36) The Bland and Altman plot for the patient group showed a 26 s bias in terms of improved performance on the second test (possible learning effect). The standard deviation of difference was 124 s and 95% Limits of Agreement (LoA_{95}) was 248 s. (36) The SEM for people with WAD II was reported to be 76 s. (36) The MDC_{90} was measured as 176 s. (36)

Spearman rank correlations were also calculated between the Fit-HANSA, Numeric Pain Rating Scale (NPRS), NDI, the disabilities of arm, hand and shoulder (DASH) and 6 cervical range of motion measures. Most (59 of 78) of the correlations between performance and comparator measures were poor (r=<0.4). (36) All correlations between total Fit-HaNSA scores and subtask scores had good correlations (r=<0.75), except for Task 1-Task 3. (36) Significant performance differences between WAD II and control groups (known group validity) were recorded for the total Fit-HaNSA score and all 3 subtask scores. (36)

Physiotherapy Test Package Subtests

Ljungquist et al published a series of articles which evaluated the clinometric properties of a physiotherapy test package for patients with spinal pain (**TABLE 7**). (37,38,39,40) This package included muscular strength & endurance tests, submaximal endurance tests, and three

functional tests. These functional tests included the PILE-C, PILE-L, and 2x20M-WWB test. Ljungquist's series of articles reported on convergent validity, known-groups validity, reliability, measurement error and sensitivity to change for these tests. (37,38,39,40)

In a 1999 article (38), correlations between the tests of the package and pain (CR-10) and perceived exertion (Borg RPE) were determined. All correlations were weak, except for a moderate correlation between the PILE-C test and pain intensity and a moderate correlation between 2x20M-WWB test and pain intensity.

In a paper from 1999, the PILE-C, PILE-L and 2x20M-WWB tests were found to have significant discriminative abilities in distinguishing healthy subjects from patients with spinal pain. (37) The sensitivity and specificity for this known group discrimination for the PILE-C test, were reported to be 0.93 and 0.69, respectively. (37) The sensitivity and specificity for the PILE-L test were reported to be 0.85 and 0.65, respectively. In a 2003 article, the PILE-C, PILE-L and 2x20M-WWB tests were tested to determine their ability to discriminate between known-groups (neck pain vs back pain). (40) Subjects with spinal pain completed the CR-10, the University of Alabama Pain Behavior scale (UAB) and the Borg RPE test. Specific cut points were used to distinguish patients with high vs. low pain intensity, high vs. low pain behavior, and high vs. low perceived exertion in patients, respectively. Participants then completed the test package and it was determined if each subtest could discriminate between participants with high vs. low pain intensity. The functional tests were able to discriminate between all 3 subgroups with the exception of the PILE-C being unable to discriminate between participants with high vs. low perceived exertion.

The inter and intra rater reliability were tested on participants with spinal pain. (38) Limits of agreement were used to measure inter rater reliability and repeatability, defined as 2x

the within-subject standard deviation of each variable. Interrater agreement for 2 tests was deemed "acceptable", while all 3 functional tests had "clinically acceptable" intrarater reliability. (38) Sensitivity-to-change was evaluated in the test package following 6 months of a physiotherapy intervention. Using ROC curves, Wilcoxon sign ranked tests and spearman correlation coefficients, only the 2x20m-WWB test and the PILE-C (women only) were deemed to be sensitive to change. (39) Additionally, moderate to high effect sizes were found for all test components.

DISCUSSION

This study synthesized 12 studies assessing clinometric properties of 4 different performance-based functional assessments. Given the limited number of studies, the substantial variation in the types of tests examined, the methods used to assess the clinical measurement properties, and the study populations, the current state of knowledge does not allow firm conclusions regarding recommendations for an optimal performance-based test at this time. Overall, there is weak to strong evidence for a range of properties of the 4 different assessments in patients with acute or chronic neck pain that is musculoskeletal in origin.

FCE

The breadth of a performance-based test is variable and defined by the developers. An advantage of the functional assessment designed by Reesink et al. (34) is that they mapped the eight subtests to risk factors identified in the literature for work-related neck disorders. The eight subtests consist of: material handling tasks, lifting floor to waist, overhead lift test, one-handed and two-handed carrying, overhead working, repetitive reaching, overhead lifting, and repetitive bending and overhead reaching. Given the systematic approach and rationale these authors used

in developing the FCE and this approach being used in previous research (41), we suggest that this test has strong content validity. However, the nature of the reporting of content validity made it difficult to formally assess this paper using the COSMIN tool.

Six articles address the clinical measurement properties of this FCE. There is adequate evidence that the FCE is stable over test-retest time of 7-14 days. (29,32) These measures demonstrate longer stability over time compared to self-report measures such as the Neck Disability Index (NDI) which has demonstrated test-retest reliability within only a short period of 0-3 days. (17) Whether this longer-term stability is a characteristic of performance-based tests or reflects differences in study populations in context requires further testing. Although test-retest reliability has been assessed, inter-rater and intra-rater reliability has yet to be researched. Unlike self-report measures, we expect measurement error due to the evaluator and performance-based tests. Thus, future research should explore these aspects of reliability.

Convergent validity is often examined in clinical measurement studies. We suggest that this may be because these comparisons are easily performed by correlating different tests rather than providing strong confidence in the validity of the measurement. Often convenient comparisons are performed rather than those most relevant. Across many domains and measures it has become clear that the relationship between self-reported function and performance-based function or physical impairment is often low to moderate. Therefore the value of assessment of these relationships as a form of validation has limited value. Several studies of varying quality have reported on the convergent validity of the FCE. (30,31,33) One article of adequate quality found the relationship between the FCE and work capacity to be poorly associated with one another. (30) The same study found that the ability of the FCE to predict future work capacity was poor. This may be considered a more important comparison since ideally performance-based

tests would relate to important outcomes like return to work. No studies to our knowledge report the responsiveness or sensitivity to change of the FCE. This is an important gap since the focus of rehabilitation is often to remediate limitations in goal impairments or work capacity, and assessment of these changes is critical to clinical decision-making and reporting outcomes. Thus, future research should evaluate the responsiveness of the FCE to provide insight in the measure's ability to detect change after an intervention.

FIT-HaNSA

One very good quality study assessed the FIT-HaNSA, a test consisting of two reaching tasks (waist and eye-level) and sustained overhead task performance. (36) Overall, the FIT-HaNSA demonstrates excellent inter-rater reliability and strong intra-rater reliability. The specific subtests included within the FIT-HaNSA similarly demonstrate moderate to strong inter-rater and intra-rater reliability. The FIT-HaNSA also demonstrated a clear ability to distinguish between people with WAD 2 and healthy controls. Correlations between the FIT-HaNSA and other patient self-report disability and functional outcome measures (NPRS, NDI, DASH, CROM and FIT-HaNSA) were generally poor ($\rho < 0.4$), consistent with other studies comparing performance and self-report. (13,14) The largest limitation in critically synthesizing information for this test is that only a single study was found that reported the measurement properties for people with neck disorders. It should be noted however that it has been validated in other MSK disorders. (1–6) Although others have noted the lag in development of performance-based measures in comparison to self-report measures, FIT-HaNSA was recommended as a performance-based measure for people with shoulder disorders. (2)

BTEWS II

One study of doubtful to adequate quality according to the COSMIN risk of bias tool assessed the efficacy of the BTEWS II where the participants performed a dynamic pushing and pulling task in which power output was recorded over a 10 second sample. (35) While the convergent validity aspect of this paper was assessed as adequate through the critical appraisal process, the relationship between the power output on the BTEWS and measures of pain and disability (NDI, SPADI, NRS) were poorly associated with each other. In addition, the power output component was not found to be significantly different between people with neck pain and healthy controls which suggests it might not be discriminative. Discrimination between patients and those without any symptoms is a low benchmark, and tests that cannot fulfil this benchmark should be viewed with caution. Because of the weak measurement properties demonstrated by the power output component of the BTEWS II, it does not appear to be a desirable performance-based measure to assess function in people with neck pain. However, we acknowledge for all of the performance-based tests the evidence pool is so shallow that there is high potential that future studies might lead to different conclusions.

Physiotherapy Test Package Subtests

Four studies assessing relevant items from a physiotherapy test package, including a lift from floor-to-waist and a waist-to-shoulder task and a two-handed carrying task, ranged in quality from “inadequate” to “very good”. The properties of these assessment items include weak to moderate correlations to pain, perceived exertion, and had “adequate” reliability. The 2x20m-WWB and PILE-C tests were found to be sensitive-to-change which is valuable information as no other study has assessed this property in performance-based measures in patients with neck disorders. Thus, this measure may be of value in clinical settings when assessing functional capacity before and after a treatment intervention. All tests had discriminative ability for

detecting participants with spinal pain vs healthy controls. Most of the three tests demonstrated poor construct validity in that they were poorly related to pain and perceived exertion, although this was observed in a study of "doubtful" quality. Thus, further research of better quality is necessary to investigate these constructs.

Limitations

A challenge in synthesizing clinical measurement evidence is the wide range of properties and indicators that need to be considered. Unlike effectiveness studies where one can focus on the effect size of treatment there are many considerations that would affect the recommendations made about outcome measures. This is further complicated when the pool of evidence is shallow. Although the COSMIN and the quality assessment tool (QACMRR) developed by one of the authors of this review which assess risk of bias and the quality of design of individual studies respectively, were useful for interpreting the evidentiary pool, there is no clear method to synthesize the extracted clinical measurement evidence. While some systematic reviews on treatment might only report findings from high-quality studies, it is important to see how outcome measures perform in different contexts. Further, the assessment of risk of bias and quality are complicated given that clinical measurement studies have so many dimensions. Therefore, exclusion of lower quality studies has questionable value. Thus, a more practical approach is to consider quality when interpreting the findings, rather than excluding studies.

The COSMIN and the QACMRR provide different perspectives since one focuses on the risk of bias and the other the quality of the research design. For example, the article by Van de Meer et al. was determined to be doubtful according to the COSMIN which is the lowest score attainable on the tool whereas the QACMRR yielded a score of 86%. Additionally, the COSMIN score for the Reneman 2017 paper in this review was found to be adequate, a much better result

than many other articles in this review but yielded the lowest score on the QACMRR of 67%. This difference is likely attributed to the QACMRRs focus on different design issues. For example, it provides lower scores where there are problems with small sample size or poor subject retention, whereas the COSMIN did not ask any specific questions that captured these qualities. The QACMRR focuses on whether the authors made appropriate decisions in selecting the scope and methods of their clinical measurement evaluations within a given study and provides descriptors of poor fair or good design options. Quality focuses on issues that might affect risk of bias or imprecision in estimates; whereas risk of bias assessments focusses on items that might result in a biased estimate. For example, insufficient power is a precision (quality) issue, not a risk of bias. Although it is difficult to interpret the meaning of the percentage of the QACMRR as there are no established cut-offs for distinguishing good and poor-quality studies, it provides one way of ranking the articles in order of quality. Since the COSMIN rates bias according to specific measurement properties whereas the the QACMRR evaluates the overall study design, we found that these tools provide complementary perspectives on the studies. Therefore, agreement on the scores was not expected.

Another limitation in this review was that the feasibility or usability of these tools was not assessed. While feasibility was not the focus of this review, information on the practical application of these performance-based measures provides valuable information to clinicians for determining whether these tests are appropriate to use in their given setting. Thus, future research should not only investigate further the psychometric properties of these tools, but also report the feasibility of using these tests so that they may be used in clinical settings and to identify limitations that restrict their application in practice.

CONCLUSION

This study confirms that performance-based tests have had far less development and evaluation than self-report measures. Limitations include the number of tests and insufficient body of evidence to make confident recommendations with respect to performance-based testing. It is clear that self-report and performance-based measures provide different perspectives. Theoretically, performance-based tests are important to inform our understanding about the mechanisms of intervention and how interventions increase capacity. Overall more work is required to further establish the psychometric properties of performance-based tests in persons with neck disorders, including sensitivity-to-change, responsiveness, and predictive validity. The data presented suggest that the FIT-HaNSA has the strongest clinometric properties though this is based on a single high-quality paper specific to neck disorder. (36, 5) Importantly, normative data have been published (6), it has been validated in multiple studies in patients with shoulder conditions (1,3,4) and has been recommended when compared to other measures (2). The FCE has a limited evidence base from which to draw, though it was developed with strong content validity and further evaluation may demonstrate its usefulness. Performance-based evaluation in people with neck disorders is an area needing much research attention both to establish the measurement properties of existing measures, potentially to develop innovative new measures and to perform head-to-head comparisons of measures before an optimal performance-based tests can be identified.

Authors' contributions

SM contributed significantly to conception and design of the study, data extraction, critical appraisal, interpretation of data and drafting of the manuscript. TS, TA, PB, and CC were involved in literature search, critical appraisal and interpretation of data and drafting. AG was involved in

critical appraisal and drafting. JM was also involved in the conception and design of the study, drafting, and revised the manuscript for important intellectual content. PB and CATWAD were involved in the drafting and review of the manuscript. All authors have given their final approval on the manuscript to be published

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study

Funding Statement

This work was supported by the Canadian Institutes of Health Research (CIHR) with funding reference number (FRN: SCA-145102).

Competing Interest Statement

None to report.

REFERENCES

1. Bobos P, MacDermid JC, Walton DM, Gross A, Santaguida PL. Patient-Reported Outcome Measures Used on Neck Disorders. An Overview of Systematic Reviews. *J Orthop Sport Phys Ther*. June 2018:1-76. doi:10.2519/jospt.2018.8131.

2. Carroll LJ, Hogg-Johnson S, Côté P, et al. Course and Prognostic Factors for Neck Pain in Workers. *Spine (Phila Pa 1976)*. 2008;33(Supplement):S93-S100. doi:10.1097/BRS.0b013e31816445d4.
3. Croft PR, Lewis M, Papageorgiou AC, et al. Risk factors for neck pain: a longitudinal study in the general population. *Pain*. 2001;93(3):317-325. http://www.ncbi.nlm.nih.gov/pubmed/11514090. Accessed July 11, 2018.
4. Vos T, Allen C, Arora M, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet*. 2016;388(10053):1545-1602. doi:10.1016/S0140-6736(16)31678-6.
5. Blanpied PR, Gross AR, Elliott JM, et al. Neck Pain: Revision 2017. *J Orthop Sport Phys Ther*. 2017;47(7):A1-A83. doi:10.2519/jospt.2017.0302.
6. Nygren A, Berglund A, von Koch M. Neck-and-shoulder pain, an increasing problem. Strategies for using insurance material to follow trends. *Scand J Rehabil Med Suppl*. 1995;32:107-112.
7. Wright A, Mayer TG, Gatchel RJ. Outcomes of disabling cervical spine disorders in compensation injuries. A prospective comparison to tertiary rehabilitation response for chronic lumbar spinal disorders. *Spine (Phila Pa 1976)*. 1999;24:178-183.
8. Rempel DM, Harrison RJ, Barnhart S. Work-related cumulative trauma disorders of the upper extremity. *JAMA*. 1992;267:838-842. https://doi.org/10.1001/jama.1992.03480060084035

9. Borghouts JA, Koes BW, Vondeling H, Bouter LM. Cost-of-illness of neck pain in The Netherlands in 1996. *Pain*. 1999;80:629-636. https://doi.org/10.1016/S0304-3959(98)00268-1

10. Alreni ASE, Harrop D, Lowe A, Tanzila Potia, Kilner K, McLean SM. Measures of upper limb function for people with neck pain. A systematic review of measurement and practical properties. *Musculoskelet Sci Pract*. 2017;29:155-163. doi:10.1016/j.msksp.2017.02.004.

11. Childs JD, Cleland JA, Elliott JM, et al. Neck pain: Clinical practice guidelines linked to the International Classification of Functioning, Disability, and Health from the Orthopedic Section of the American Physical Therapy Association. *J Orthop Sports Phys Ther*. 2008;38(9):A1-A34. doi:10.2519/jospt.2008.0303.

12. Finch E, Canadian Physiotherapy Association. *Physical Rehabilitation Outcome Measures : A Guide to Enhanced Clinical Decision Making*. BC Decker; 2002. https://www.ncbi.nlm.nih.gov/nlmcatalog/101175542. Accessed July 19, 2018.

13. Simmonds MJ, Olson SL, Jones S, et al. Psychometric characteristics and clinical usefulness of physical performance tests in patients with low back pain. *Spine (Phila Pa 1976)*. 1998;23(22):2412-2421. http://www.ncbi.nlm.nih.gov/pubmed/9836355. Accessed July 19, 2018.

14. Stratford PW, Kennedy D, Pagura SMC, Gollish JD. The relationship between self-report and performance-related measures: questioning the content validity of timed tests. *Arthritis Rheum*. 2003;49(4):535-540. doi:10.1002/art.11196.

15. Novy DM, Simmonds MJ, Lee CE. Physical performance tasks: what are the underlying constructs? *Arch Phys Med Rehabil*. 2002;83(1):44-47. http://www.ncbi.nlm.nih.gov/pubmed/11782832. Accessed July 19, 2018.

16. MacDermid JC, Stratford P. Applying evidence on outcome measures to hand therapy practice. *Journal of Hand Therapy*. 2004;*17*(2):165-173. doi: 10.1197/j.jht.2004.02.005.

17. Macdermid, J. C., Walton, D. M., Avery, S., Blanchard, A., Etruw, E., Mcalpine, C., & Goldsmith, C. H. (2009). Measurement properties of the neck disability index: a systematic review. *Journal of orthopaedic & sports physical therapy*, *39*(5), 400-C12.

18. Misailidou V, Malliou P, Beneka A, Karagiannidis A, Godolias G. Assessment of patients with neck pain: a review of definitions, selection criteria, and measurement tools. *J Chiropr Med*. 2010;9(2):49-59. doi:10.1016/j.jcm.2010.03.002.

19. Holly LT, Matz PG, Anderson PA, et al. Functional outcomes assessment for cervical degenerative disease. *J Neurosurg Spine*. 2009;11(2):238-244. doi:10.3171/2009.2.SPINE08715.

20. Pietrobon R, Coeytaux RR, Carey TS, Richardson WJ, DeVellis RF. Standard scales for measurement of functional outcome for cervical pain or dysfunction: a systematic review. *Spine (Phila Pa 1976)*. 2002;27(5):515-522. http://www.ncbi.nlm.nih.gov/pubmed/11880837. Accessed July 19, 2018.

21. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *J Clin Epidemiol*. 2009;62(10):1006-1012. doi:10.1016/j.jclinepi.2009.06.005.

22. Mokkink, L. B., De Vet, H. C., Prinsen, C. A., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN risk of Bias checklist for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, *27*(5), 1171-1179.

23. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012; 21: 651–7

24. MacDermid J. Appendix G: critical appraisal of study quality for psychometric articles, evaluation form. In: Law M, MacDermid J, eds. *Evidence-Based Rehabilitation: A Guide to Prac- tice*. 2nd ed. Thorofare, NJ: SLACK Incorporated; 2008:387-388.

25. Roy J-S, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for musculoskeletal disorders: a systematic review. *J Rehabil Med*. 2011;43(1):23-31. doi:10.2340/16501977-0643.

26. Vessey J, Strout TD, DiFazio RL, Walker A. Measuring the youth bullying experience: a systematic review of the psychometric properties of available instruments. *J Sch Health*. 2014;84(12):819-843. doi:10.1111/josh.12210.

27. Roy J-S, MacDermid JC, Woodhouse LJ. A systematic review of the psychometric properties of the Constant-Murley score. *J shoulder Elb Surg*. 2010;19(1):157-164. doi:10.1016/j.jse.2009.04.008.

28. Roy J-S, MacDermid JC, Woodhouse LJ. Measuring shoulder function: A systematic review of four questionnaires. *Arthritis Rheum*. 2009;61(5):623-632. doi:10.1002/art.24396.

29. Trippolini MA, Reneman MF, Jansen B, Dijkstra PU, Geertzen JHB. Reliability and safety of functional capacity evaluation in patients with whiplash associated disorders. *J Occup Rehabil*. 2013;23(3):381-390. doi:10.1007/s10926-012-9403-z.

30. Trippolini MA, Dijkstra PU, Côté P, Scholz-Odermatt SM, Geertzen JH, Reneman MF. Can functional capacity tests predict future work capacity in patients with whiplash-associated disorders? *Arch Phys Med Rehabil*. 2014;95(12):2357-2366. doi:10.1016/j.apmr.2014.07.406.

31. Trippolini MA, Dijkstra PU, Geertzen JHB, Reneman MF. Construct Validity of Functional Capacity Evaluation in Patients with Whiplash-Associated Disorders. *J Occup Rehabil*. 2015;25(3):481-492. doi:10.1007/s10926-014-9555-0.

32. Reneman MF, Roelofs M, Schiphorst Preuper HR. Reliability and Agreement of Neck Functional Capacity Evaluation Tests in Patients With Chronic Multifactorial Neck Pain. *Arch Phys Med Rehabil*. 2017;98(7):1476-1479. doi:10.1016/j.apmr.2016.12.005.

33. Van der Meer S, Reneman MF, Verhoeven J, van der Palen J. Relationship between self-reported disability and functional capacity in patients with whiplash associated disorder. *J Occup Rehabil*. 2014;24(3):419-424. doi:10.1007/s10926-013-9473-6

34. Reesink DD, Jorritsma W, Reneman MF. Basis for a functional capacity evaluation methodology for patients with work-related neck disorders. *J Occup Rehabil*.

35. Lomond K V, Côté JN. Shoulder functional assessments in persons with chronic neck/shoulder pain and healthy subjects: Reliability and effects of movement repetition. *Work*. 2011;38(2):169-180. doi:10.3233/WOR-2011-1119.

36. Pierrynowski M, McPhee C, P Mehta S, C MacDermid J, Gross A. Intra and Inter-Rater Reliability and Convergent Validity of FIT-HaNSA in Individuals with Grade Π

Whiplash Associated Disorder. *Open Orthop J*. 2016;10(1):179-189. doi:10.2174/1874325001610010179.

37. Ljungquist T, Fransson B, Harms-Ringdahl K, Björnham A, Nygren A. A physiotherapy test package for assessing back and neck dysfunction--discriminative ability for patients versus healthy control subjects. *Physiother Res Int*. 1999;4(2):123-140. http://www.ncbi.nlm.nih.gov/pubmed/10444762. Accessed July 11, 2018.
38. Ljungquist T, Harms-Ringdahl K, Nygren A, Jensen I. Intra- and inter-rater reliability of an 11-test package for assessing dysfunction due to back or neck pain. *Physiother Res Int*. 1999;4(3):214-232. http://www.ncbi.nlm.nih.gov/pubmed/10581627. Accessed July 11, 2018.
39. Ljungquist T, Nygren A, Jensen I, Harms-Ringdahl K. Physical performance tests for people with spinal pain--sensitivity to change. *Disabil Rehabil*. 2003;25(15):856-866. doi:10.1080/0963828031000090579.
40. Ljungquist T, Jensen IB, Nygren A, Harms-Ringdahl K. Physical performance tests for people with long-term spinal pain: aspects of construct validity. *J Rehabil Med*. 2003;35(2):69-75. http://www.ncbi.nlm.nih.gov/pubmed/12691336. Accessed July 11, 2018.
41. Reneman MF, Dijkstra PU, Westmaas M, Göeken LNH. Test-retest reliability of lifting and carrying in a 2-day functional capacity evaluation. *J Occup Rehabil*. 2002;12(4):269-275. http://www.ncbi.nlm.nih.gov/pubmed/12389478. Accessed July 19, 2018.

613
614
615
616

TABLE 1. Summary of Studies Reporting Psychometric Properties of Functional-based Tests in Neck Disorder Patients

Study	Population	Sample Size (n)	Functional Tests	Intervention/Test Interval
Ljungquist et al. 1999	Neck pain, back pain, multiple pain sites, chronic pain	53	PILE-C, PILE-L	N/A
Ljungquist et al. 1999	Neck pain, lumbar pain, thoracic pain, shoulder pain, multiple pain sites, chronic pain	68	PILE-C, PILE-L, 2 x 20m WWB	8 days
Ljungquist et al. 2003	Neck pain, lumbar pain, thoracic pain, shoulder pain, lower extremity pain, multiple pain sites, chronic pain	235	PILE-C, PILE-L, 2 x 20m WWB	N/A
Ljungquist et al. 2003	cervical pain, lumbar pain, cervical and lumbar pain, multiple pain sites, chronic pain	186	PILE-C, PILE-L, 2 x 20m WWB	6 months
Lomond and Cote. 2011	Chronic neck and shoulder pain	32	BTEWS II	9.5 days
Pierrynowski et al. 2016	Sub-acute and chronic WAD II	66	FIT-HaNSA	2-7 days
Reesink et al. 2007	N/A	N/A	Neck-FCE	N/A
Reneman et al. 2017	Chronic multifactorial neck pain	18	Neck-FCE	2 weeks
Trippolini et al. 2013	Sub acute and chronic WAD I and II	32	WAD FCE	7 days
Trippolini et al. 2014	Sub acute and chronic WAD I and II	267	Workwell FCE	N/A
Trippolini et al. 2015	Sub acute and chronic WAD I and II	314	WAD FCE	N/A
Van der Meer et al. 2013	Chronic WAD I and II	40	Neck FCE	N/A

PILE-C, Progressive Isoinertial Lifting Evaluation-Cervical; PILE-L, Progressive Isoinertial Lifting Evaluation; CBT, Cognitive-Behavioural Therapy; PT, Physical Therapy; NRPS, Numeric Pain Rating Scale; BTEWS II, Baltimore Therapeutic Equipment Work Simulator II; WAD, Whiplash Associated Disorder; MVA, Motor Vehicle Accident; FIT-HaNSA, Functional Impairment Test-Hand and Neck/Shoulder/Arm; FCE, Functional Capacity Evaluation; EXP, Experimental; M, Male; F, Female

TABLE 2. Summary of Psychometric Properties Reported in Studies and COSMIN risk of bias checklist scores

Study	Psychometric Properties Reported	COSMIN Score
Ljungquist et al. 1999	Known-groups Validity	Adequate
	Convergent Validity	Very Good
Ljungquist et al. 1999	Reliability	Inadequate
	Measurement Error	Adequate
Ljungquist et al. 2003	Known-groups Validity	Very Good
Ljungquist et al. 2003	Sensitivity to Change	Doubtful
Lomond and Cote. 2011	Reliability	Doubtful
	Measurement Error	Doubtful
	Known-groups Validity	Doubtful
	Convergent Validity	Adequate
Pierrynowski et al. 2016	Reliability	Very Good
	Measurement Error	Adequate
	Known-groups Validity	Very Good
	Convergent Validity	Very Good
Reesink et al. 2007	Content Validity	N/A*
Reneman et al. 2017	Reliability	Adequate
	Measurement Error	Adequate
Trippolini et al. 2013	Reliability	Adequate
	Measurement Error	Adequate
Trippolini et al. 2014	Convergent Validity	Very Good
	Predictive Criterion Validity	Very Good
Trippolini et al. 2015	Known-groups Validity	Very Good
	Convergent Validity	Inadequate
Van der Meer et al. 2013	Convergent Validity	Doubtful

COSMIN, Consensus-based Standards for the Selection of health Measurement Instruments
*Paper is not applicable for completion of COSMIN checklist

TABLE 3. Quality of Studies on Psychometric Properties of Functional-based Tests Evaluated in Neck Disorder Patients

Study	Item Evaluation Criteria												
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Total (%)
Trippolini et al, 2014	2	2	2	2	1	2	2	2	2	2	1	2	92%
Lomond and Cote, 2011	2	2	1	2	0	2	2	2	2	2	2	2	88%
Pierrynowski et al, 2016	2	2	1	2	0	2	2	2	2	2	2	2	88%
Trippolini et al, 2015	2	2	2	0	1	N/A	2	2	2	2	2	2	86%
Van der Meer et al, 2013	2	1	2	1	2	N/A	2	1	2	2	1	2	86%
Ljungquist et al 2003 KGV	2	2	2	0	0	N/A	2	2	2	2	2	2	82%
Ljungquist et al 1999 Rel	2	1	1	2	0	2	2	2	2	2	1	2	79%
Ljungquist et al 2003 STC	1	1	1	2	1	1	2	2	2	2	2	2	79%
Trippolini et al, 2013	2	2	1	1	0	0	2	2	2	2	2	2	75%
Ljungquist et al 1999 KGV	2	1	1	2	0	N/A	2	1	2	2	1	2	68%
Reneman et al, 2017	1	2	1	1	1	0	1	2	2	2	2	1	67%
Reesink, 2007*	-	-	-	-	-	-	-	-	-	-	-	-	N/A

*Paper is not applicable for completion of study quality tool

TABLE 4. Psychometric Properties of the Functional Capacity Evaluation

FCE Battery	Type of Properties	Statistical Test	Value	Quality
Neck FCE	Test-retest	ICC	0.39-0.96	Poor-excellent
	Measurement Error	Ratio of LoA	32.0-56.5%	
	Convergent Validity	Pearson or Spearman correlation	NDI total: 0.39-0.62 NDI items: 0.03-0.63	Weak to moderate Negligible to moderate
WAD FCE	Test-retest Reliability	ICC	0.66-0.96	Moderate-excellent
	Convergent Validity	Pearson Correlation	Pain* 0.31-0.39 SFS: 0.42-0.61 NDI: 0.34-0.45 HADS-A: 0.27-0.36 HADS-D: 0.30-0.41	Weak Moderate Weak Negligible-weak Weak
	Known-groups Validity (German vs Non-German)	Linear Regression Analysis	$p<0.001$	Significant for All Tasks
	Known-groups Validity (sex)	t-test	$p<0.001$	Significant for Two Tasks
Workwell FCE	Convergent Validity	Pearson or Spearman Correlation	Work Capacity: 0.1-0.3	Weak
	Predictive Validity	Pearson or Spearman Correlation	0.06-0.39	Weak
		Linear Mixed Model Regression of All Predictors	β=-0.04, 95% CI: -0.15 – 0.06 p=0.428 (task 6)	Not Significant

FCE, Functional Capacity Evaluation; ICC, Intraclass correlation coefficient; LoA, Limits of Agreement; NDI, Neck Disability Index; Mod., Moderate; Neg., Negligible; SFS, Spinal Function Sort; HADS-A, Hospital Anxiety and Depression Scale – Anxiety; HADS-D, Hospital Anxiety and Depression Scale – Depression; CI, Confidence Interval Sig., Significant
*Pain measured via Numeric Rating Scale

TABLE 5. Summary of Fit-HaNSA's psychometric properties in neck disorder patients

Test	Type of Property	Statistical Test	Value	Quality
Fit-HaNSA	Intra-rater Reliability	ICC	0.78	Strong
Fit-HaNSA	Inter-rater Reliability	ICC	0.84	Strong
Fit-HaNSA	Measurement Error	SEM LOA_{95} MDC_{90}	76 s 248 s 176 s	
Fit-HaNSA	Convergent Validity	Spearman Rank Correlation	<0.4 - >0.75	Moderate - Strong
Fit-HaNSA	Known-groups Validity WAD II vs Control	F-test	62.6, <p,0.001	Significant
Fit-HaNSA Functional Sub-tasks	Intra-rater reliability	ICC	0.70-0.72	Strong
	Inter-reliability	ICC	0.54-0.80	Moderate
	Convergent Validity	Spearman Rank Correlation	<0.4 - >0.75	Moderate - Strong
	Known-groups Validity WAD II vs Control	F-test	42.0-53.3, p<0.001	Significant

Fit-HaNSA, Functional Impairment Test, Hand and Neck/Shoulder/Arm; ICC, Intraclass correlation coefficient; SEM, Standard Error of Measurement; LOA_{95}, 95% Limits of Agreement; MDC_{90}, 90% Minimal Detectable Change; WAD, Whiplash Associated Disorder; Mod, Moderate

*Correlations completed with Numeric Pain Rating Scale, Neck Disability Index, Disabilities of Arm, Shoulder, Hand and 6 cervical range of motion tests

TABLE 6. Psychometric Properties of Baltimore Therapeutic Equipment Work Simulator II – Power Output Task

Test	Type of Property	Statistical Test	Value	Quality
BTEWS II	Test-retest reliability	ICC	0.53	Moderate
		Spearman	0.37	Poor
BTEWS II	Measurement Error	SEM	30.25	
		MDC_{90}	70.59	
BTEWS II	Convergent Validity*	Spearman	Not Reported	Weak
BTEWS II	Known-groups Validity (Pain vs Control)	Two-way Repeated Measures ANOVA	Not Reported	Non-significant

ICC, Intraclass correlation coefficient; SEM, Standard Error of Measurement; MDC_{90}, 90% Minimal Detectable Change; ANOVA, Analysis of Variance

*Spearman correlations completed with Numeric Rating Scale, Neck Disability Index and Shoulder Pain and Disability Index

TABLE 7. Psychometric Properties of performance-based tests included in physiotherapy test package

Test	Type of Property	Statistical Test	Value	Quality
PILE-C	Inter-rater Reliability	Mean Difference LoA	-0.24 -2.46 and 1.82	
PILE-C	Inter-rater Reliability	Repeatability (2X SD) % of Range	M=3.93; F=1.19 M=10.5%; F=6.1%	
PILE-C	Convergent Validity	Spearman Correlation	CR-10: Unreported* Borg RPE: Unreported	Moderate Low
PILE-C	KGV: spinal pain vs. control	Sensitivity and Specificity	0.93, 0.69	
PILE-C	KGV: spinal pain vs. control	Wilcoxon Sign Ranked Test	p=0.008	Significant
PILE-C	KGV: High vs. low pain intensity	Mann-Whitney U	p=0.003	Significant
PILE-C	KGV: High vs. low Pain behavior	Mann-Whitney U	p=0.005	Significant
PILE-C	KGV: High vs. low perceived exertion	Mann-Whitney U	p=0.154	Non-significant
PILE-C	Sensitivity to Change	Effect Size	Subjects improving: 0.39 - 0.73 Subjects deteriorating: 0 – 0.4	Low – Moderate Negligible – Low
PILE-L	Inter-rater Reliability	Mean Difference LoA	-0.11 -2.33 and 2.11	
PILE-L	Intra-rater Reliability	Repeatability % of Range	M=4.0; F=3.59 M=10.7%; F=18.5%	
PILE-L	Convergent Validity	Spearman Correlation	CR-10: Unreported Borg RPE: Unreported	Low Low
PILE-L	KGV: spinal pain vs no spinal pain	Sensitivity and Specificity	0.85, 0.65	
PILE-L	KGV: spinal pain vs control	Wilcoxon Sign Ranked Test	p=0.002	Significant

PILE-L	KGV: High vs. low pain intensity	Mann-Whitney U	p=0.001	Significant
PILE-L	KGV: High vs. low pain behaviour	Mann-Whitney U	p<0.001	Significant
PILE-L	KGV: High vs. low perceived exertion	Mann-Whitney U	p<0.001	Significant
PILE-L	Sensitivity to change	Effect Size	Subjects improving: 0.02 – 1.08 Subjects deteriorating 0.42-0.81	Negligible – Strong Weak – Strong
2 x 20m WWB	Inter-rater Reliability	Mean Difference LoA	0.05 -1.33 and 1.43	
2 x 20m WWB	Intra-rater Reliability	Repeatability % of Range	3.2 10.7%	
2 x 20m WWB	Convergent Validity	Spearman Correlation	CR-10: Unreported Borg RPE: Unreported	Moderate Low
2 x 20m WWB	KGV: spinal pain vs control	Wilcoxon Sign Ranked Test	p=0.014	Significant
2 x 20m WWB	KGV: High vs. low pain intensity	Mann Whitney U	p<0.001	Significant
2 x 20m WWB	KGV: High vs. low pain behaviour	Mann Whitney U	p<0.001	Significant
2 x 20m WWB	KGV: High vs. low perceived exertion	Mann Whitney U	p<0.001	Significant
2 x 20m WWB	Sensitivity to change	Effect Size	Subjects improving: 0.38-0.78 Subjects deteriorating: 0.13-0.62	Weak – Moderate Negligible – Moderate

PILE-C, Progressive Iso-intertial Lifting Evaluation – Cervical; PILE-L, Progressive Iso-intertial Lifting Evaluation – Lumbar; LoA, Limits of Agreement; SD, Standard Deviation; M, Male; F, Female; RPE, Rating of perceived exertion; KGV, Known-groups Validity; Neg., Negligible; Mod., Moderate, *CR-10: Measurement of pain construct

Figure 1. Selection of the studies for inclusion in the systematic review

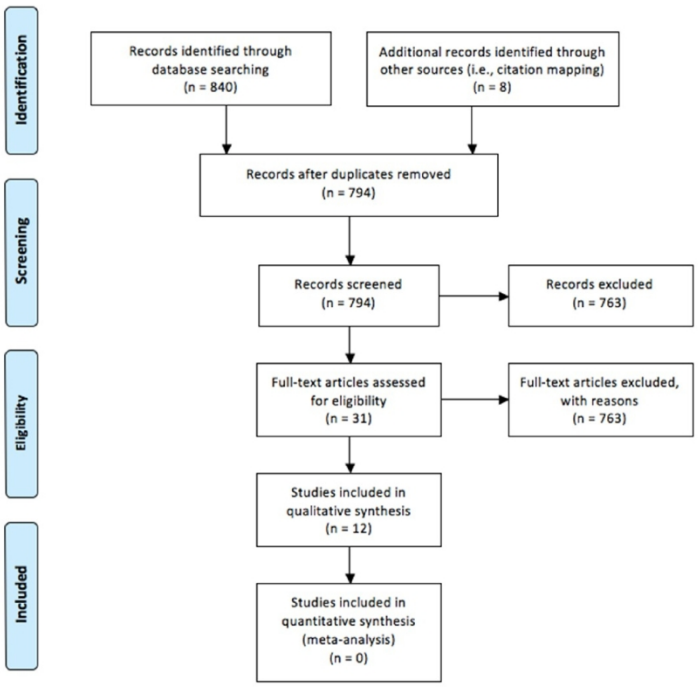

Figure 1. Selection of the studies for inclusion in the systematic review

215x279mm (300 x 300 DPI)

APPENDICES

APPENDIX A. Data extraction guide for studies evaluating the quality of studies evaluating the clinical measurement properties of outcome measures

Instructions

Clinical measurement studies may evaluate a wide spectrum of measurement properties; or evaluate aspects that relate to the implementability or interpretation of outcome measures. Individual clinical measurement studies cannot address every aspect of the measurement properties of an instrument. Ideally systematic reviews will synthesize the quality and content of research evidence addressing the clinical measurement properties of individual outcome measures. The summative knowledge about the measurement properties, cultural transferability, and utility across different contexts provides the scope of information needed to select an outcome measure for a specific patient (population), purpose and context.

This guide should facilitate extraction of data from individual clinical measurement studies. An explanation of the measurement property addressed in each item and how it might be measured within a given study is listed to facilitate finding and extracting that information. The accompanying extraction form can then be used to collect the specific information on these measurements or utility properties from specific studies.

The purpose of data extraction is to extract the specific information reported by authors within a study, not to evaluate the validity or value of that piece of information. Evaluation of the quality of the published version of the clinical measurement study (also called critical appraisal) is performed in a separate step. See the accompanying critical appraisal tool and guide. It is advisable to extract detailed specific information from the study; recognizing that this information may later be synthesized or subject to meta-analysis.

There is no standardized process for synthesizing clinical measurement information. Based on the findings of extraction you may elect to present the synthesize data in a descriptive way by creating a summary table of the data extracted in each category. If you find some studies with similar designs, you may be able to conduct a meta-analysis of some properties like clinically important difference (CID) or minimal detectable change (MDC); if appropriate given the sample and technique - this can be valuable as it may provide more stable estimates of these important properties.

Population studied		
Population	A description of the study population	Sample size, pathology/disorder, demographics, setting, acute vs. chronic, where subjects were chosen from. Report meaningful demographics and indicators of the population studied.
Intervention	Interventions (if applicable) applied during longitudinal studies	Description of the nature, frequency, intensity of the intervention and the follow-up interval.
Reliability		
Reliability Description	The extent to which the same results are obtained on repeated administrations of the same measure when no change in status has occurred (reliability) or the precision of the scores on repeated measurements (agreement).	Test procedures or measures are typically reapplied on repeated occasions in individuals considered to have a stable condition during that time frame which repeated testing occurs. Repeated testing may be performed on different occasions (test-retest) for self-report measures, OR by the same rater (intra-rater) or different raters (inter-rater) if it is an observer-based scale. In some cases different test instruments (inter-instrument) are evaluated. The most common statistic used is the intraclass correlation coefficient for quantitative data (Shrout & Fleiss, 1979) and kappa (Landis & Koch, 1977) for nominal data. Standard error of measurement is used to present a quantitative estimate of the reliability—in the original units of measure. Report the type of reliability evaluated and coefficients obtained.
Reliability (relative)	The relationship (ratio) between variability in test scores when repeating the test on the same person in comparison to the overall variability (including variation between people)—typically indicated by a reliability coefficient	ICCs (Shrout & Fleiss, 1979) or another reliability coefficient and their associated confidence intervals are extracted.
Reliability (absolute)	Absolute reliability is portrayed as the quantity of error that could be anticipated upon repeated testing - reported in the original units of measure.	This may be reported as 1. Standard error of measurement (in older articles you may see coefficient of variation);

		2. Altman and Bland graphical technique (Bland & Altman, 1990; Bland & Altman, 1987; Bland & Altman, 1986) where the difference on repeated tests for each individual (limits of agreement) is plotted versus their mean score. The mean difference and the boundaries of 2SD are shown to define the limits of agreement.
Minimum Detectable Change	Calculated from the reliability coefficient and the level of confidence specified for error margins. This indicator reflects the amount of change required before you can be confident that change exceeds the random error that occurs in stable patients.	Extract the number and level of confidence.
Content/structural validity		
Internal consistency	The extent to which items on a test or subscale are related (an indication of the consistency of the concept measured).	Cronbach's alpha is the inter-item correlation usually reported. Report alpha and whether it relates to the entire instrument or specific subscales.
Content Validity	The extent to which the conceptual domain or construct that a test is designed to measure is adequately reflected by the items in the measure. In assessing content validity, it is important to consider the population to whom the measure applies, the completeness of the content, the relevancy and emphasis of the content assessed.	A variety of techniques can be used to assess the extent to which items on a given measure reflected the necessary content to capture the concept of interest. Some of the techniques you will find are listed. Extract what was done to determine content validity and what was found. 1) Patients and experts were involved during item selection/reduction - report how they were used and key decisions 2) Patients were consulted for reading and comprehension - report key findings 3) Cognitive interviews (Cibelli, 1994; Ojanen & Gogates, 2006) were done with patients to determine how items were interpreted by respondents, their perceptions of the items - report key findings 4) Expert panels or Delphi procedures were used to select items or evaluate the validity of the instrument - report key findings and decisions

		5) During translation specific study, the meaning of the questions to another cultural or language group was studied - report key findings and decisions 6) ICF linking (Cieza et al., 2002) or other coding of content was performed - report the results which may include the distribution of content across ICF domains, or the distribution of specific codes
Floor-Ceiling Effects	The measure is unable to indicate a worsening score in patients who have clinically deteriorated and/or an improved score in patients who have clinically improved	There are a variety of potential methods; so the method and conclusion should be reported. Descriptive statistics of the distribution of scores that may be presented graphically or numerically may be used to indicate this. Other studies report the percentage of patients sustained a floor or ceiling effect defined by the number of people who fall in the extremes ranges. Note different studies may define the extreme ranges for floor/ceiling differently, so extract how it was defined and % of patients who obtained floor or ceiling category scores.
Factorial validity	The extent to which factor analysis supports assumptions surrounding constructs measured as defined by the measure or as indicated by subscale structure	Factor analysis may be reported as raw results; or compared to the inherent structure of the instrument or factor analysis upon which its construction was based. Report the type of factor analysis performed (exploratory or confirmatory), rotations used and the number of factors derived; specify whether this confirms the expected instrument structure or original factor structure.
Item response /Rasch Analyses	The extent to which items cross a range of difficulty, or a spectrum of the concept measured. The measurement scaling of the items.	Using item response theory or Rasch analysis, items are fit to a model to demonstrate interval scaling and determine item difficulty (Pallant & Tennant, 2007). Analyses might address item difficulty, person's ability curves, and comparison of ability estimation. Most commonly, the item difficulty and the composition of the test that fulfills interval scaling are defined. Data to be extracted include information on the scaling of the items, whether the interval scaling has been established; and the presence or absence of differential item functioning

36/bmjopen-2019-031424 on 4 December 2019.

		(DIF), where items perform differently on different types of respondents.
Construct Validity		
Construct Validity - correlational	Constructs are artificial frameworks that are not directly observable. Construct validity assesses the extent to which measures perform according to a priori defined constructs. Construct validity can be cross-sectional or longitudinal (predictive). Constructed hypotheses can assess convergent validity where measures are thought to represent similar constructs or divergent validity where it is assumed they measure different constructs. For cross-cultural validation, the expected relationships are those that have been reported in validation of the instrument in its original language/format.	When extracting data about correlational validity, the pre-constructed hypothesis and whether it is supported should be documented. For correlational construct validity, this will be the nature and strength of the prespecified relationship and the correlations that support that. Relation to other indices/constructs that are similar (convergent) or different (divergent) can be reported. Ideally, hypotheses are formulated/reported and supported by correlations that are in accordance with the hypotheses. Note that there is no consistent agreement on what subjective term should be applied to validity correlations. Note that there is no consistent agreement on what subjective term should be applied to validity correlations. Some authors use subjective terminology defined for reliability such as: strong (>0.70) and moderate (0.40-0.70) correlations; others use the correlations like effect size benchmarks that 0.4 indicates a moderate effect and 0.6 a large effect. For validity assessment is more important than correlations prespecified constructed hypotheses, although not all papers are written clearly with respect to this.
Convergent	The Relationship between similar scales/tests. Correlations are generally expected to be moderate to strong if the relationship is one where there is confidence that they measure a similar construct.	Extract test names, prespecified expected relationship and correlations observed.
Divergent	Divergent validity assesses the extent to which different scales/tests that are designed to	Extract test names, prespecified expected relationship and correlations observed.

	measure different constructs demonstrate that they are different by a lack of correlation between them.	
Construct validity - known groups	Known groups analysis supports the validity of a measure by demonstrating that the measurement is able to differentiate between groups that are prespecified and known to be different on the construct being assessed.	Data extraction should include the nature of the subgroups and the size of the difference observed between them (and its statistical significance). Typically, statistical tests of difference are performed. Since known groups analysis can provide data that is useful in clinical practice as benchmarks for comparing these known groups, it is a more practical form of construct validity than correlational. Data extraction/presentation should reflect this by presenting the group central tendency, their margins and statistical significance in an accessible manner.
Longitudinal Validity	This form of validity supports the validity of a measure by demonstrating that the change that occurs over time onto similar instruments is correlated in a manner consistent with the nature of the relationship between the scales. It is measured over a retest interval when clinically relevant change could be expected.	Extract test names and correlations Note: since longitudinal validity is based on four measures (pre-and post-test on two different measures), and since error tends to mitigate the strength of correlations, strong longitudinal correlations can be difficult to obtain.
Criterion validity Description	Criterion validation is determined by comparing a given outcome measure to an accepted standard of measure. For subjective constructs like pain and disability, it can be argued that there is no criterion since there is no external gold standard. Therefore, for self-report measures, validation focuses on construct validity. For performance measures, it is common to have a criterion measure that is considered to be highly precise and rigorous as the criterion comparator.	Authors will state that their measure is being compared against a specific instrument and report the correlation or agreement between the measures. Extract the test names and results: correlations or other as reported.
Concurrent criterion	Concurrent validity is assessed by comparing a scale and its criterion at a single point in time	Extract the test names and correlations.

Predictive criterion	Predictive validity is evaluated by determining the extent to which the results of administering an outcome measure at one point in time can accurately predict a future status or outcome.	Extract the test names and correlations and time interval. (and important cutoffs if those were established/reported), if diagnostic test methodology was used to examine prediction, and sensitivity specificity and other diagnostic criteria were reported, they should be extracted.
Responsiveness/Clinical Change		
Responsiveness	Does the instrument detect changes over time that matters to patients?	Extract indicators of responsiveness include: effect size, standard response mean and the method for assessing whether patients were improved, stable or worse. (Beaton, 2000)
Clinically Important Difference (CID)	CID is the difference in scores that patients find to be observable and clinically important. It is assessed by comparing scores to an external benchmark of clinical relevance such as a global rating of change or some other method. The terminology used to rate the nature of this difference will affect the estimation process. Differences in methods include how clinically importance is framed and the metrics/process by which that is determined.	Extract the MID or CID and note the method/cut-off used to define importance. Extract how the clinically important differences were framed to respondents; or determined. For example, minimal, moderate, extreme improvement or better/not better, etc.

APPENDIX B. Data extraction form for studies evaluating the clinical measurement properties of outcome measures

Authors: _____ Year:_____ Rater: _____

Instructions

When using the data extraction form, it is important to realize that the purpose of data extraction is to remove or extract the specific information reported by authors within a study, not to evaluate the validity or value of that piece of information. To make data extraction as useful as possible, and to avoid the need for repeated data extractions, it is advisable to read the accompanying guide and then be as specific as possible when extracting information.

	DATA EXTRACTED
Population studied	
Population	
Intervention	
Reliability	
Reliability (relative)	
Reliability (absolute)	
Minimum Detectable Change	
Content/structural validity	

Internal consistency	
Content Validity	
Floor-Ceiling Effects	
Factorial validity	
Item response /Rasch Analyses	
Construct/Criterion Validity	
Known groups	
Convergent	
Divergent	
Longitudinal Validity	
Concurrent criterion	
Predictive criterion	
Responsiveness/Clinical Change	

Responsiveness	
Minimally Clinical Important Difference	

APPENDIX C. Quality Appraisal for Clinical Measurement Research Reports Evaluation Form

Rater (Group)_____
Author(s) (Study Author(s) _____
Year (Year of publication)_____

1. Was the relevant background work cited to define what is currently known about the measurement properties of measures under study, and the potential contributions of the current research question to informing that knowledge base?
   2
   1
   0
2. Were appropriate inclusion/exclusion criteria defined? *
   2
   1
   0
3. Were specific clinical measurement questions/hypotheses identified?
   2
   1
   0
4. Was an appropriate scope of measurement properties considered?
   2
   1
   0
5. Was an appropriate sample size used?
   2
   1
   0
6. Was appropriate retention/follow-up obtained? (for studies involving retesting; otherwise n/a)

2
1
0

7. Were specific descriptions provided of the measure under study and the method(s) used to administer it?

2
1
0

8. Were standardized procedures used to administer all study measures in a manner that minimized potential sources of error/bias (including the study measure and its comparators)?

2
1
0

9. Were analyses conducted for each specific hypothesis or purpose?

2
1
0

10. Were appropriate statistical tests performed to obtain point estimates of the measurement properties?

2
1
0

11. Were appropriate ancillary analyses done to quantify the confidence in the estimates of the clinical measurement property (Precision/Confidence intervals; benchmark comparisons/ROC curves, alternate forms of analysis like SEM/MID, etc.)?

2
1
0

12. Were clear, specific and accurate conclusions made about the clinical measurement properties; that were associated with appropriate clinical measurement recommendations and supported by the study objectives, analysis and results?

2

1
0
Subtotals (of column 1 and 2) Total Score (sum of subtotals/24*100)

APPENDIX D. Description of each performance battery from selected articles

Battery	Description of Tasks
Relevant FCE Subtasks[25,26,27,28,29,30]	Material Handling Tasks: All lifting tests were executed with a wooden crate (40 × 30 × 26 cm) of 2.5 kg, and four to five weight increments of 2.5 kg or 5 kg each were used until the maximum amount of weight was reached. Maximum performance was recorded in kg. Lifting floor to waist: Measured after five lifts of crate from floor to table and vice versa (time limit < 90 s): hands remained on the crate during the test. Increase weight in 4-5 steps until maximum is reached Overhead lift test: Five lifts from waist to crown height and vice versa within 90 s in standing position. Increase weight in 4–5 steps until maximum is reached Two-handed carrying: Carrying of a crate for a short distance measured after five carries of 1.5 m distance at waist height. Hands remain on the crate during the test. One-handed carrying: Carrying wooden crate for 15 m within 90 s beginning with the right hand and thereafter the left hand. Overhead working: Standing with hands at crown height for manipulation of nuts and bolts. The time that the position was held is recorded (sec). Repetitive reaching: fast horizontal movements of the upper extremity in a sitting position. Marbles are removed from bowls at arm length distance at table height from left to right and vice versa, with right and then left arm. The time taken to remove 30 marbles is recorded (sec). Overhead lift test: Five lifts from waist to crown height and vice versa within 90 s in standing position. Increase weight in 4–5 steps until maximum is reached

	Repetitive bending and overhead reaching: 20 marbles in 2 bowls at table height and crown height. Standing in front of bowl of marbles and moving the marbles as fast as possible from table height to crown height.
A Physiotherapy Test Package[33,34,35,36]	**PILE Tests:** "The lifting tests were performed standing in front of bookshelves with shelves at 0.76m and 1.37 m from the floor. Subjects were asked to lift weights in a plastic box from floor to waist level (0–0.76 m) for the lumbar PILE test, or from waist to shoulder height (0.76–1.37 m) for the cervical PILE test. The initial weight was 3.6 kg for women and 5.9 kg for men. A 'lifting movement' involved a single transfer from one level to the next and back again. After every four such lifting movements (= 20 s), the weight was increased by 2.25 kg for women and 4.5 kg for men. The weight managed during the last lifting movement was recorded and used as a test result, as well as this maximum weight divided by the 'adjusted weight'". **2x20m WWB:** "Subjects were asked to walk 20 m at a comfortable speed along a corridor, to turn around where 20 m was marked and then to walk 20 m back to the starting point. In the first walking test they carried no extra weight, but in the second they carried one carrier bag in each hand, containing 4 kg each for the women, 8 kg each for the men. The time taken was recorded to get the walking speed. The tests were discontinued after 50 s".
BTEWS II[31]	"The protocol consisted of performing a series of shoulder functional tasks before and after a fatiguing activity. Functional tasks consisted of active shoulder range of motion (ROM) in both flexion and abduction and cumulative power output (PO) accumulated over 10s during a repetitive pushing/pulling task in a horizontal plane at shoulder level".
FIT - HaNSA[32]	"The FIT-HaNSA protocol consists of three timed tasks and each task is performed for a maximum of 300 seconds (s) with approximately 30 s pause between them (set-up time for next task). Task 1 (waist-up) requires the patient to alternately "grab, lift, move, and place" three 1000 g containers located on waist level and 25 cm above waist level shelves, using their affected arm, at a metronome pace of 60 beats per minute for 300 s or until they felt unable to continue. The time to complete Task 1 is measured using a stopwatch. Task 2 (eye-down) is identical to Task 1 except that the two shelves are placed at eye-level and 25 cm below. Task 3 (overhead work) requires a patient to repeatedly screw and unscrew bolts in a sagittal plane oriented plate

	positioned at eye-level using both arms". More complete description at https://srs-mcmaster.ca/wp-content/uploads/2015/04/FIT-HaNSAProtocol_April2007.pdf

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	2
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	3
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	3
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	4
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	4
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	3-4
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	3-4
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	4
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	4
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	5
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	NA
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	NA

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	NA
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	NA
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	6-7
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	6-7
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	6-10
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	6-10
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	6-10
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	6-10
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	NA
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	11-13
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	14-16
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	16
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	18

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.

BMJ Open

Systematic Review of the Measurement Properties of Performance-based Functional Tests in Patients with Neck Disorders

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-031242.R1
Article Type:	Original research
Date Submitted by the Author:	22-Aug-2019
Complete List of Authors:	McGee, Steven; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Sipos, Taylor; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Allin, Thomas; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Chen, Celia; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Greco, Alexandra; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Bobos, Pavlos; Western University, Health and Rehabilitation Sciences; University of Toronto, Dalla Lana School of Public Health, Institute of Health Policy Management and Evaluation MacDermid, Joy ; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Group, CATWAD; Michele Sterling, Anne Söderlund, Michele Curatolo, Jim Elliott, David M Walton, Helge Kasch, Linda Carroll, Hans Westergren, Samuel McLean, Gwendolen Jull, Genevieve Grant Luke Connelly, Joy C MacDermid, Mandy Nielsen, Pierre Côté, Tonny Elmose Andersen, Trudy Rebbeck Annick Maujean, Sarah Robins, Kenneth Chen, Julia Treleaven
Primary Subject Heading:	Rehabilitation medicine
Secondary Subject Heading:	Rehabilitation medicine
Keywords:	functional, psychometric properties, neck pain, cervical, outcome measures

Title: Systematic Review of the Measurement Properties of Performance-based Functional Tests in Patients with Neck Disorders

[1]Steven McGee, PT

[2]Taylor Sipos, PT

[3]Thomas Allin, PT

[4]Celia Chen, PT

[5]Alexandra Greco, PT

[6]Pavlos Bobos, PT, PhD(c) (corresponding author)

[7]Joy MacDermid, PT, PhD

[8]CATWAD

Authors' information

[1]Steven McGee PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (smcgee7@uwo.ca)

[2]Taylor Sipos PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (jsipos@uwo.ca)

[3]Thomas Allin PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (tallin@uwo.ca)

[4]Celia Chen PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (qchen224@uwo.ca)

[5]Alexandra Greco PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (agreco33@uwo.ca)

[6]Pavlos Bobos PT, PhD(c), (corresponding author) Doctoral Candidate, Western's Bone and Joint Institute, Department of Health and Rehabilitation Sciences, Western University, Elborn College, 1201 Western Road, N6G 1H1, London, Ontario, Dalla Lana School of Public Health, Institute of

Health Policy Management and Evaluation, Department of Clinical Epidemiology and Health Care Research, University of Toronto, Canada, (pbobos@uwo.ca), tel: +1 519 661 2111 x88912

[7]Joy C MacDermid BScPT, PhD, Professor, Physical Therapy and Surgery, Western University, London, ON and Co-director Clinical Research Lab, Hand and Upper Limb Centre, St. Joseph's Health Centre, London, Ontario; Professor Rehabilitation Science McMaster University, Hamilton, ON, Canada (jmacderm@uwo.ca)

[5]CATWAD Coauthors: Michele Sterling m.sterling@uq.edu.au, Anne Söderlund anne.soderlund@mdh.se, Michele Curatolo curatolo@uw.edu, Jim Elliott j-elliott@northwestern.edu, David M Walton dwalton5@uwo.ca, Helge Kasch helgkasc@rm.dk, Linda Carroll linda.carroll@ualberta.ca, Hans Westergren Hans.Westergren@skane.se, Samuel A McLean, Samuel_McLean@med.unc.edu, Gwendolen Jull g.jull@uq.edu.au, Genevieve Grant genevieve.grant@monash.edu, Luke Connelly l.connelly@uq.edu.au, Joy C MacDermid, jmacderm@uwo.ca, Mandy Nielsen mandy.nielsen@griffith.edu.au, Pierre Cote pierre.cote@uoit.ca, Tonny Elmose Andersen tandersen@health.sdu.dk, Trudy Rebbeck trudy.rebbeck@sydney.edu.au, Annick Maujean a.maujean@uq.edu.au, Sarah Robins s.robins1@uq.edu.au, Kenneth Chen k.chen8@uq.edu.au, Julia Treleaven j.treleaven@uq.edu.au

Key Words: functional, psychometric properties, neck, cervical, outcome measures

Word Count: 4509

Abstract

Objectives: The purpose of this systematic review is to identify and synthesize studies evaluating performance-based outcome measures designed to evaluate the functional abilities of patients with neck pain.

Design: Systematic review

Data Sources: A literature search using PubMed, Scopus, CINAHL, EMBASE, COCHRANE, Google Scholar, and a citation mapping strategy was conducted till July 2019

Eligibility criteria: More than half of the study's patient population had neck pain or a musculoskeletal neck disorder and completed a functional-based test. Clinimetric properties of at least one performance-based functional tests were reported. Both traumatic and non-traumatic origins of neck pain were considered.

Data extraction and synthesis: Relevant data were then extracted from selected articles using an extraction guide. Selected articles were appraised the Quality Appraisal for Clinical Measurement Research Reports Evaluation Form (QACMRR).

Results: The search obtained 12 articles which reported on 4 outcome measures (Functional Capacity Evaluations (FCE), Baltimore Therapeutic Equipment Work Simulator II (BTEWS II), Functional Impairment Test- Hand and Neck/Shoulder/Arm (FIT-HaNSA)) reporting to assess the functional abilities in patients with mechanical neck pathology. Of the selected papers: 1 reports content validity, 5 construct validity, 4 reliability, 1 sensitivity to change, and 1 both reliability and construct validity. QACMRR scores ranged from 68% to 95%.

Conclusions: This review found very good quality evidence that the FIT-HaNSA has excellent inter and intra-rater reliability and very weak to weak convergent validity. Excellent quality evidence of fair test-retest reliability, weak convergent validity, and very weak known

groups validity for the BTEWS II test was found. Good to excellent quality evidence exists that an FCE battery has poor to excellent reliability and very weak to strong validity. Good to excellent quality of weak to strong validity and trivial to strong effect sizes were found for a physiotherapy test package.

Prospero registration: CRD42018112358

Strengths and limitations of this study

- The psychometric properties of performance outcome measures for neck pain were synthesized and critically appraised
- This study assessed the risk of bias and the quality of measurements properties
- The feasibility or usability of these tools was not assessed

Introduction

Neck pain has been associated with high disability and is regarded as a substantial societal burden.[1] Approximately 70% of people experience neck pain within their lifetime and about 33% of adults experience neck pain every year.[2,3] Further concern is warranted as it has been suggested that the incidence of neck pain is increasing.[4–6] The economic burden due to neck disorders is high, including lost wages, costs of treatment, and compensation expenditures to injured people.[7,8] Neck pain is second only to low back pain in annual workers' compensation costs in the United States and has been associated with many other comorbidities such as headaches, anxiety, depression, back pain and arthralgias.[6,9,10]

Outcome measures are a crucial component in monitoring patients with neck pain to determine the effects of treatment[11,12], evaluation of interventions, guiding return to work, and justifying treatment.[13,14] Several self-reported outcome measures currently exist to assess disability and function in those with neck pain (e.g. the Neck Disability Index - NDI). [13] Evidence-based clinical practice guidelines suggest that measures assessing physical performance should also be used for people with neck pain.[15] Performance-based testing is where the assessment is based on actual performance of a task or activity. Physical performance can be assessed by testing a person's ability to execute a standardized activity in a standardized environment (i.e. clinical setting).[16] Time to complete the activity, number of repetitions performed, and weight lifted are frequently used to quantify the physical performance.[17] Conversely, self-report measures examine patients' perception and experience of their ability to perform functional tasks. [16] Previous research has demonstrated poor to fair relationships between physical performance and self-report measures of ability in patients with various musculoskeletal disorders suggesting that these measures assess different constructs of function. [17,18] Consequently, physical performance tests and self-report measures complement each other and may each contribute unique information about a patient's function. [19]

A fundamental component of monitoring outcomes is having reliable and valid tools with known measurement properties.[20,21] While recent research has investigated the psychometric properties of patient-reported outcomes in people with neck pain [21,22] there is a gap in knowledge with respect to performance-based functional outcomes. The purpose of this systematic review was to identify and synthesize clinical measurement studies that evaluate measurement properties of performance-based functional tests in patients with neck disorders.

METHODS

Patient and Public Involvement

There was no patient or public involvement in the design or planning of this study.

Study Design and Protocol Registration

We conducted a systematic review to evaluate the psychometric properties of performance-based functional tests for people with mechanical neck disorders. The protocol was registered in PROSPERO register with registration number CRD42018112358.

Search Strategy

A database search using CINAHL, PubMed, Scopus and Google Scholar was performed to identify articles published till July 2019. The following search strategy was used to search all databases for eligible studies: (Reliability OR validity OR responsiveness OR calibration OR validation) OR (minimal detectable change) OR (clinically important difference) OR (psychometric properties) AND cervical OR neck OR c-spine AND (performance measure) OR (functional test) OR (functional outcome) OR (performance outcome). MeSH terms were searched in PubMed. A citation map of articles and systematic reviews selected for the full-text review was performed. This strategy was included to minimize the risk of publication bias. The full search strategy is summarized in **APPENDIX 1.** The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) process[23] was followed to ensure all appropriate steps were taken in the selection process (**FIGURE 1**).

Inclusion Criteria

Articles were included in the final review if all of the following criteria were met:

- >50% of the study's patient population had neck pain or a musculoskeletal neck disorder (e.g. whiplash associated disorder (WAD II))
- Patients in the study completed a functional-based test
- Clinometric properties of at least one performance-based test were reported.

A test was considered functional-based if it met the following criteria:

- assessment of a patient's ability to execute a standardized activity in a standardized environment
- tests assessing muscular endurance (e.g. cervical flexion test) or proprioception were not deemed functional-based as they are often not reflective of physical working conditions.

Both traumatic and non-traumatic origins of neck pain were considered. Definitions for the properties can be found in **APPENDIX A.**

Article Selection

Titles and abstracts generated by the search strategy were screened by two authors (SM and PB) independently. Articles that met the inclusion criteria and selected for a full text review were also reviewed in pairs of authors. Disagreements were resolved by the most experienced author (JCM)

Data Extraction

Data extraction and critical appraisal was performed in pairs of two raters among the authors, after the completion of a calibration session in which the most experienced author (JCM) reviewed the data extraction tools with the authors that performed the data extraction. When reviewers disagreed

during data extraction and/or critical appraisal, and consensus could not be met, a third author arbitrated. A data extraction form [24] (**APPENDIX A and APPENDIX B**), developed by one of the authors (JCM.), was used to ensure systematicity. Authors extracted sample size, patient population characteristics, functional tests performed and reported psychometric properties. The ICC interpretation of ICC < 0.40 indicating poor, 0.40 ≤ ICC < 0.75 indicating fair-to-good and ICC ≥ 0.75 indicating excellent reliability were used as a common benchmark. For validity estimates, correlation coefficient (Pearson's/Spearman) and the 95% confidence intervals were extracted if were available. [24,25] Evan's guidelines to interpret the strength of the correlation was used which included: 0.00–0.19 "very weak", 0.20–0.39 "weak", 0.40–0.59 "moderate", 0.60–0.79 "strong", and 0.80–1.00 "very strong". To assist clinical decision making, standard benchmark scores of trivial (< 0.20), small (≥ 0.20 to < 0.50), moderate (≥ 0.50 to < 0.80) or large (≥ 0.80), as proposed by Cohen, were used. [26]

Quality Appraisal for Clinical Measurement Research Reports Evaluation Form

Pairs of authors critically appraised the quality of each study using a standardized 12-item evaluation tool (QACMRR) designed to assess the quality of studies determining measurement properties in outcome measures (**APPENDIX C**). If disagreement was present a third person (JM) assist in resolving the discrepancy. [24] This tool has been found to have good to excellent pre-consensus inter-rater reliability (ICC: 0.69-0.91) across a number of systematic reviews.[24,25,27] The evaluation criteria of this tool included twelve items: 1) Thorough literature review to define the research question; 2) Specific inclusion/exclusion criteria; 3) Specific hypotheses; 4) Appropriate scope of psychometric properties; 5) Sample size; 6) Follow-up; 7) The authors

referenced specific procedures for administration, scoring, and interpretation of procedures; 8) Measurement techniques were standardized; 9) Data were presented for each hypothesis; 10) Appropriate statistics-point estimates; 11) Appropriate statistical error estimates; and 12) Valid conclusions and recommendations. [24,25] Each item is scored from 0 to 2 with (score=2) is the best; (score=1) is acceptable but suboptimal; (score=0) is not done/documented, substantially inadequate or inappropriate. An article's total score – quality - was calculated by the sum of scores for each item, divided by the numbers of items and multiplied by 100%. [24,25] Overall, the quality summary of appraised articles ranges from (0%-30%) Poor, (31%-50%) Fair, (51%-70%) Good, (71%-90%) Very Good, and (>90%) Excellent

RESULTS

The search strategy resulted in 840 published articles. After duplications were removed, 31 articles were deemed relevant and were screened at full text. Overall, 12 articles met our inclusion criteria (**FIGURE 1**). The excluded articles were removed due to inappropriate patient populations, investigations into self-report measures or tests assessing proprioception/muscular endurance rather than functional-based measures, or because the articles were found to be systematic reviews. The characteristics of the included studies and the summary of psychometric properties are presented in **TABLE 1.** The quality assessment is summarized and presented in **TABLE 2**. Percent agreement was calculated for quality scores between the 2 raters and it was 90%.

Participants

Participants in the selected articles had various types of neck pain including subacute, chronic, and whiplash-associated disorder. The mean/median age of the samples of each study ranged from 30-48 years of age. The proportion of females in each article ranged from 34-78% of the study population. Two studies that had a mixed sample of subjects with various spinal pain did not report the demographics of the neck pain portion of their sample. One study did not contain any subjects and performed a review of epidemiological literature to establish content validity for work-related neck disorders **TABLE 1.**

Functional-Based Tests

The 12 articles that were included for review provided properties on the following functional based tests: Functional Capacity Evaluations (FCE)[28–33], The Baltimore Therapeutic Equipment Work Simulator II (BTEWS II) [34], Functional Impairment Test- Hand and Neck/Shoulder/Arm (FIT-HaNSA) [35], as well as items off of a physiotherapy test package including a cervical and lumbar Progressive Isoinertial Lifting Evaluation (PILE-C, PILE-L) test [36–39] and 2 x 20 m with burden walking test (2x20M-WWB) [36–39]. Descriptions of all functional-based tests and their relevant subtasks are provided in **APPENDIX D.**

Functional Capacity Evaluations (FCE)

Six articles reported measurement properties for an FCE battery. We identified multiple versions of the FCE in the literature with one article reporting properties on the Workwell FCE [29], two reporting on the Whiplash Associated Disorder (WAD) FCE [28,30] and three reporting on the neck-FCE.[31–33] These test batteries include various combinations of muscular strength,

endurance and functional based tests. The measurement properties of the functional based tests used by the FCE are outlined in **TABLE 3.**

Individuals with Sub-acute to chronic WAD

Trippolini et al. (2014)[29] evaluated the Workwell FCE test-retest reliability, measurement error, convergent validity and predictive criterion validity of future work capacity in workers diagnosed with WAD I or II. Interclass Correlation Coefficients (ICC) ranged from 0.66 to 0.96 (good to excellent). Limits of agreement relative to mean performance ranged from 21 to 57% for functional based sub-tests. Correlations between FCE sub scores and baseline work capacity were very weak to weak ranging between r=0.06 and r=0.39. FCE sub scores did not predict future work capacity at 1, 3, 6 and 12 months.

Trippolini et al. (2015)[28] assessed the WAD FCE (31) and evaluated convergent validity and known-groups validity. FCE subscales showed very weak to strong correlations (0.15-0.68) with each of: pain, self-reported functional ability, self-reported disability, anxiety and depression. It was found that the FCE had known-group sex validity (males vs females) for 1 of 3 functional subtests (lifting waist-overhead) and reported significant performance differences between culture groups (German vs non-German language groups).

Work-Related Neck Disorders

Reesink et al. (2007)[33] developed an independent FCE for patients with musculoskeletal neck disorders (neck FCE). They performed a review of epidemiological literature and identified four physical risk factors for work-related neck disorders and used that information to develop an FCE consisting of eight functional-based tests. Content validity was established by following

operational definitions of the risk factors when searching the literature and using current literature to provide a rationale to guide their development of the tasks comprising the FCE.

Chronic Neck Pain

Reneman et al. (2017)[31] measured test-retest reliability of the subscales of the neck FCE in patients with multifactorial neck pain. Test-retest ICC's ranged from poor to excellent (0.39-0.96). Limits of agreement relative to mean performance range from 32.0% to 56.5% for functional based sub tests. Convergent validity was performed against the Neck Disability Index (NDI) items and total score.[32] The authors found weak to strong Pearson correlations (0.39-0.70) for the FCE sub scores to both NDI individual items and the NDI total score.

The Baltimore Therapeutic Equipment Work Simulator II (BTEWS II)

Chronic Neck Pain

Lomond and Côté, (2011)[34] reported on the reliability, measurement error, minimum detectable change (MDC) and validity of the power output (PO) task during the BTEWS II test in patients with chronic neck and shoulder pain (**TABLE 4**). Test-retest reliability, measured with Spearman Rank correlations and ICC's was of fair and measured at ρ=0.37 and $ICC_{2,1}$ = 0.54, respectively. The standard error of measurement (SEM) and the minimal detectable change at 90% confidence (MDC_{90}) for the PO task were measured as 30.25 and 70.59, respectively. Weak Spearman Rank correlations between the PO task and the NDI, Shoulder Pain and Disability Index (SPADI) and Numeric Rating Scale (NRS) for pain tests were recorded. There were no significant performance differences between control and pain groups for the PO task.

Functional Impairment Test- Hand and Neck/Shoulder/Arm (Fit-HaNSA)

Sub-acute to chronic WAD

Pierrynowski et al. (2016)[35] reported on the reliability, measurement error, MDC and validity of the Fit-HaNSA test in a sample of people with WAD II following motor vehicle collision (MVC) (**TABLE 5**). Intra-rater reliability ICC's for patient subtask and total scores were good to excellent ranging between 0.70-0.78. [35] Inter-rater reliability ICC's for patient subtask and total scores were fair to excellent and ranged between 0.54-0.84. [35] The Bland and Altman plot for the patient group showed a 26 seconds (s) bias in terms of improved performance on the second test (possible learning effect). The standard deviation of difference was 124 s and 95% Limits of Agreement (LoA_{95}) was 248 seconds. [35] The SEM for people with WAD II was reported to be 76 s. The MDC_{90} was measured as 176 s. [35]

Spearman rank correlations were also calculated between the Fit-HANSA, Numeric Pain Rating Scale (NPRS), NDI, the disabilities of arm, hand and shoulder (DASH) and 6 cervical range of motion measures. Most (59 of 78) of the correlations between performance and comparator measures were very weak to weak (r=<0.4). [35] All correlations between total Fit-HaNSA scores and subtask scores had good correlations (r=<0.75), except for Task 1-Task 3. [35] Significant performance differences between WAD II and control groups (known group validity) were recorded for the total Fit-HaNSA score and all 3 subtask scores. [35]

Physiotherapy Test Package Subtests

Ljungquist et al. published a series of articles[36–39] which evaluated the clinimetric properties of a physiotherapy test package for patients with spinal pain (**TABLE 6**). This package included muscular strength & endurance tests, submaximal endurance tests, and three

functional tests. These functional tests included the PILE-C, PILE-L, and 2x20M-WWB test. Ljungquist's series of articles reported on convergent validity, known-groups validity, reliability, measurement error and sensitivity to change for these tests. [36–39]

Undetermined duration of neck pain

In a 1999 article [38], correlations between the tests of the package and pain (CR-10) and perceived exertion (Borg RPE) were determined. All correlations were very weak to moderate (0.10-0.48) except for moderate to strong correlations (0.55-0.65) between the PILE-C test and pain intensity and between 2x20M-WWB test and pain intensity.

In a 2003 article[36], the PILE-C, PILE-L and 2x20M-WWB tests were tested to determine their ability to discriminate between known-groups (neck pain vs back pain). Subjects with spinal pain completed the CR-10, the University of Alabama Pain Behavior scale (UAB) and the Borg RPE test. Specific cut points were used to distinguish patients with high vs. low pain intensity, high vs. low pain behavior, and high vs. low perceived exertion in patients, respectively. Participants then completed the test package and it was determined if each subtest could discriminate between participants with high vs. low pain intensity. The functional tests were able to discriminate between all 3 subgroups with the exception of the PILE-C being unable to discriminate between participants with high vs. low perceived exertion.

In a paper from 1999[38], the PILE-C, PILE-L and 2x20M-WWB tests were found to have significant discriminative abilities in distinguishing healthy subjects from patients with spinal pain. The sensitivity and specificity for this known group discrimination for the PILE-C test, were reported to be 0.93 (very strong) and 0.69 (strong), respectively. The sensitivity and specificity for the PILE-L test were reported to be 0.85 (very strong) and 0.65 (strong), respectively.

The inter and intra rater reliability were tested on participants with spinal pain.[37] Limits of agreement were used to measure inter rater reliability and repeatability, defined as 2x the within-subject standard deviation of each variable. Interrater agreement for 2 tests was deemed "acceptable", while all 3 functional tests had "clinically acceptable" intra-rater reliability.

Sensitivity-to-change was evaluated in the test package following 6 months of a physiotherapy intervention. Using ROC curves, Wilcoxon sign ranked tests and spearman correlation coefficients, only the 2x20m-WWB test and the PILE-C (women only) were deemed to be sensitive to change. [39] Additionally, moderate to large effect sizes were found for all test components.

DISCUSSION

This study synthesized 12 studies assessing clinometric properties of 4 different functional-based assessments. Given the limited number of studies, the substantial variation in the types of tests examined, the methods used to assess the clinical measurement properties, and the study populations, the current state of knowledge does not allow firm conclusions regarding recommendations for an optimal functional-based test at this time. Overall, the quality ranging from good to excellent (67-92%,) as determined by the QACMRR, for a range of properties of the 4 different assessments in patients with acute or chronic neck pain that is musculoskeletal in origin. Studies obtaining higher percentages indicate research that has been consistent with best practice where studies with lower percentages are more likely to be inadequate or inappropriate

FCE

The breadth of a functional-based test is variable and defined by the developers. An advantage of the functional assessment designed by Reesink et al.[33] is that they mapped the

eight subtests to risk factors identified in the literature for work-related neck disorders. The eight subtests consist of: material handling tasks, lifting floor to waist, overhead lift test, one-handed and two-handed carrying, overhead working, repetitive reaching, overhead lifting, and repetitive bending and overhead reaching. Given the systematic approach and rationale these authors used in developing the FCE and this approach being used in previous research [40], we suggest that this test has strong content validity.

Six articles address the clinical measurement properties of this FCE ranging from good to excellent quality (67-92%). There was evidence that the FCE was stable over test-retest time of 7-14 days. [30,31] These measures demonstrate longer stability over time compared to self-report measures such as the Neck Disability Index (NDI) which has demonstrated test-retest reliability within only a short period of 0-3 days. [27] Whether this longer-term stability is a characteristic of functional-based tests or reflects differences in study populations in context requires further testing. These two studies had relatively lower quality scores on the QACMRR (67-75%) compared to other studies in this review putting into question test-retest time. Although test-retest reliability has been assessed, inter-rater and intra-rater reliability has yet to be researched. Unlike self-report measures, we expect measurement error due to the evaluator and functional-based tests. Thus, future research should explore these aspects of reliability.

Convergent validity is often examined in clinical measurement studies. We suggest that this may be because these comparisons are easily performed by correlating different tests rather than providing strong confidence in the validity of the measurement. Often convenient comparisons are performed rather than those most relevant. Across many domains and measures it has become clear that the relationship between self-reported function and performance-based function or physical impairment is often very weak to moderate. Therefore, the value of assessment

of these relationships as a form of validation has limited value. Several studies of very good to excellent quality have reported on the convergent validity of the FCE. [28,29,32] The highest quality article determined by the QACMRR (92%) found the relationship between the FCE and work capacity to be poorly associated with one another. [29] The same study found that the ability of the FCE to predict future work capacity was poor. This may be considered a more important comparison since ideally functional-based tests would relate to important outcomes like return to work. No studies to our knowledge report the responsiveness or sensitivity to change of the FCE. This is an important gap since the focus of rehabilitation is often to remediate limitations in goal impairments or work capacity, and assessment of these changes is critical to clinical decision-making and reporting outcomes. Thus, future research should evaluate the responsiveness of the FCE to provide insight in the measure's ability to detect change after an intervention.

FIT-HaNSA

One study of very good quality (88%) assessed the FIT-HaNSA, a test consisting of two reaching tasks (waist and eye-level) and sustained overhead task performance. [35] Overall, the FIT-HaNSA demonstrated excellent inter-rater reliability (0.84) and intra-rater reliability (0.78). The specific subtests included within the FIT-HaNSA similarly demonstrate fair to excellent (0.54-0.80) and good (0.70-0.72) inter-rater and intra-rater reliability respectively. The FIT-HaNSA also demonstrated a clear ability to distinguish between people with WAD 2 and healthy controls. Correlations between the FIT-HaNSA and other patient self-report disability and functional outcome measures (NPRS, NDI, DASH, CROM and FIT-HaNSA) were generally very weak to weak ($\rho < 0.4$), consistent with other studies comparing performance and self-report. [17,18] The largest limitation in critically synthesizing information for this test is that only a single study was found that reported the measurement properties for people with neck disorders. It should be noted

however that it has been validated in other MSK disorders. [34,40] Although others have noted the lag in development of functional-based measures in comparison to self-report measures, FIT-HaNSA was recommended as a functional-based measure for people with shoulder disorders. [41]

BTEWS II

Another study of very good quality (88%) assessed the efficacy of the BTEWS II where the participants performed a dynamic pushing and pulling task in which power output was recorded over a 10 second sample.[34] While the convergent validity aspect of this paper was assessed as consistent with best practice through the critical appraisal process, the relationship between the power output on the BTEWS and measures of pain and disability (NDI, SPADI, NRS) were poorly associated with each other. In addition, the power output component was not found to be significantly different between people with neck pain and healthy controls which suggests it might not be discriminative. Discrimination between patients and those without any symptoms is a low benchmark, and tests that cannot fulfil this benchmark should be viewed with caution. Because of the weak measurement properties demonstrated by the power output component of the BTEWS II, it does not appear to be a desirable functional-based measure to assess function in people with neck pain. However, we acknowledge for all of the functional-based tests the evidence pool is so shallow that there is high potential that future studies might lead to different conclusions.

Physiotherapy Test Package Subtests

Four studies ranging from good to very good quality (68-82%) assessed relevant items from a physiotherapy test package, including a lift from floor-to-waist and a waist-to-shoulder task and a two-handed carrying task. The properties of these assessment items include weak to moderate correlations to pain, perceived exertion, and had “fair to good” reliability. The 2x20m-WWB and PILE-C tests were found to be sensitive-to-change which is valuable information as no

other study has assessed this property in functional-based measures in patients with neck disorders. Thus, this measure may be of value in clinical settings when assessing functional capacity before and after a treatment intervention. All tests had discriminative ability for detecting participants with spinal pain vs healthy controls. Most of the three tests demonstrated poor construct validity in that they were poorly related to pain and perceived exertion. Thus, further research is necessary to investigate these constructs.

Clinical Implications

This study confirms that functional-based tests have had far less development and evaluation than self-report measures. Limitations include the number of tests and insufficient body of evidence to make confident recommendations with respect to functional-based testing. It is clear that self-report and functional-based measures provide different perspectives. Theoretically, functional-based tests are important to inform our understanding about the mechanisms of intervention and how interventions increase capacity. Overall more work is required to further establish the psychometric properties of functional-based tests in persons with neck disorders, including sensitivity-to-change, responsiveness, and predictive validity.

The data presented suggest that the FIT-HaNSA has the strongest clinometric properties though this is based on a single higher quality paper specific to neck disorder. [35] Importantly, normative data have been published [42], it has been validated in multiple studies in patients with shoulder conditions [43–45] and has been recommended when compared to other measures [41]. The FCE has a limited evidence base from which to draw, though it was developed with strong content validity and further evaluation may demonstrate its usefulness.

Limitations

A challenge in synthesizing clinical measurement evidence is the wide range of properties and indicators that need to be considered. Unlike effectiveness studies where one can focus on the effect size of treatment there are many considerations that would affect the recommendations made about outcome measures. This is further complicated when the pool of evidence is shallow. Although the quality assessment tool (QACMRR) developed by one of the authors of this review which assess the quality of design of individual studies were useful for interpreting the evidentiary pool, there is no clear method to synthesize the extracted clinical measurement evidence. While some systematic reviews on treatment might only report findings from high-quality studies, it is important to see how outcome measures perform in different contexts. Further, the assessment of quality is complicated given that clinical measurement studies have so many dimensions. Therefore, exclusion of lower quality studies has questionable value. Thus, a more practical approach is to consider quality when interpreting the findings, rather than excluding studies.

The QACMRR focuses on whether the authors made appropriate decisions in selecting the scope and methods of their clinical measurement evaluations within a given study and provides descriptors of poor fair or good design options. Quality focuses on issues that might affect risk of bias or imprecision in estimates; whereas risk of bias assessments focusses on items that might result in a biased estimate. For example, insufficient power is a precision (quality) issue, not a risk of bias. Although it is difficult to interpret the meaning of the percentage of the QACMRR as there are no established cut-offs for distinguishing good and poor-quality studies, it provides one way of ranking the articles in order of quality. We did not use COSMIN checklist since it was developed for PROMS and some of the components/steps that involved are not applicable to performance-based tests.

Another limitation in this review was that the feasibility or usability of these tools was not assessed. While feasibility was not the focus of this review, information on the practical application of these functional-based measures provides valuable information to clinicians for determining whether these tests are appropriate to use in their given setting. Thus, future research should not only investigate further the psychometric properties of these tools, but also report the feasibility of using these tests so that they may be used in clinical settings and to identify limitations that restrict their application in practice.

CONCLUSION

This review found very good quality evidence that the FIT-HaNSA has excellent inter and intra-rater reliability and very weak to weak convergent validity. Excellent quality evidence of fair test-retest reliability, weak convergent validity, and very weak known groups validity for the BTEWS II test was found. Good to excellent quality evidence exists that an FCE battery has poor to excellent reliability and very weak to strong validity. Good to excellent quality of weak to strong validity and trivial to strong effect sizes were found for a physiotherapy test package. Functional-based evaluation in people with neck disorders is an area needing much research attention both to establish the measurement properties of existing measures, potentially to develop innovative new measures and to perform head-to-head comparisons of measures before an optimal functional-based test can be identified.

Authors' contributions

SM contributed significantly to conception and design of the study, data extraction, critical appraisal, interpretation of data and drafting of the manuscript. TS, TA, PB, and CC were involved in literature search, critical appraisal and interpretation of data and drafting. AG was involved in

critical appraisal and drafting. JM was also involved in the conception and design of the study, drafting, and revised the manuscript for important intellectual content. PB and CATWAD were involved in the drafting and review of the manuscript. All authors have given their final approval on the manuscript to be published

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study

Funding Statement

This work was supported by the Canadian Institutes of Health Research (CIHR) with funding reference number (FRN: SCA-145102).

Competing Interest Statement

None to report.

References

1 Carroll LJ, Hogg-Johnson S, van der Velde G, *et al.* Course and Prognostic Factors for Neck Pain in the General Population. Results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. *J Manipulative Physiol Ther* Published Online First: 2009. doi:10.1016/j.jmpt.2008.11.013

2 Croft PR, Lewis M, Papageorgiou AC, *et al.* Risk factors for neck pain: A longitudinal study in the general population. *Pain* Published Online First: 2001. doi:10.1016/S0304-3959(01)00334-7

3 Vos T, Allen C, Arora M, *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* Published Online First: 2016. doi:10.1016/S0140-6736(16)31678-6

4 Blanpied PR, Gross AR, Elliott JM, *et al.* Neck Pain: Revision 2017. *J Orthop Sport Phys Ther* Published Online First: 2017. doi:10.2519/jospt.2017.0302

5 Nygren A, Berglund A, Von Koch M. Neck-and-shoulder pain, an increasing problem. Strategies for using insurance material to follow trends. In: *Scandinavian Journal of Rehabilitation Medicine, Supplement*. 1995.

6 Wright A, Mayer TG, Gatchel RJ. Outcomes of disabling cervical spine disorders in compensation injuries: A prospective comparison to tertiary rehabilitation response for chronic lumbar spinal disorders. *Spine (Phila Pa 1976)* Published Online First: 1999. doi:10.1097/00007632-199901150-00020

7 Rempel DM, Harrison RJ, Barnhart S. Work-Related Cumulative Trauma Disorders of the Upper Extremity. *JAMA J Am Med Assoc* Published Online First: 1992. doi:10.1001/jama.1992.03480060084035

8 Borghouts JAJ, Koes BW, Vondeling H, *et al.* Cost-of-illness of neck pain in The Netherlands in 1996. *Pain* Published Online First: 1999. doi:10.1016/S0304-3959(98)00268-1

9 Hogg-Johnson S, van der Velde G, Carroll LJ, *et al.* The Burden and Determinants of

Neck Pain in the General Population. Results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. *J Manipulative Physiol Ther* Published Online First: 2009. doi:10.1016/j.jmpt.2008.11.010

10 Bobos P, Nazari G, Palimeris S, *et al.* The contribution of health and psychological factors in patients with chronic neck pain and disability: A cross-sectional study. *J Clin Diagnostic Res* Published Online First: 2018. doi:10.7860/JCDR/2018/31284.11203

11 Bobos P, Billis E, Papanikolaou D-T, *et al.* Does Deep Cervical Flexor Muscle Training Affect Pain Pressure Thresholds of Myofascial Trigger Points in Patients with Chronic Neck Pain? A Prospective Randomized Controlled Trial. *Rehabil Res Pract* Published Online First: 2016. doi:10.1155/2016/6480826

12 Nazari G, Bobos P, Billis E, *et al.* Cervical flexor muscle training reduces pain, anxiety, and depression levels in patients with chronic neck pain by a clinically important amount: A prospective cohort study. *Physiother Res Int* 2018;**23**. doi:10.1002/pri.1712

13 Bobos P, MacDermid JC, Walton DM, *et al.* Patient-Reported Outcome Measures Used for Neck Disorders: An Overview of Systematic Reviews. *J Orthop Sport Phys Ther* 2018;**48**:1–76. doi:10.2519/jospt.2018.8131

14 Macdermid JC, Walton DM, Bobos P, *et al.* The Open Orthopaedics Journal A Qualitative Description of Chronic Neck Pain has Implications for Outcome Assessment and Classification. *Open Orthop J* 2016;**10**:746–56. doi:10.2174/1874325001610010746

15 Childs JD, Cleland JA, Elliott JM, *et al.* Neck pain: Clinical practice guidelines linked to the international classification of functioning, disability, and health from the orthopaedic section of the american physical therapy association. J. Orthop. Sports Phys. Ther. 2008. doi:10.2519/jospt.2008.0303

16 Kay TM, Huijbregts M. Physical Rehabilitation Outcome Measures: A Guide to Enhanced Clinical Decision Making, Second Edition. *Physiother Canada* Published Online First: 2003. doi:10.2310/6640.2003.35271

17 Simmonds MJ, Olson SL, Jones S, *et al.* Psychometric characteristics and clinical usefulness of physical performance tests in patients with low back pain. *Spine (Phila Pa 1976)* Published Online First: 1998. doi:10.1097/00007632-199811150-00011

18 Stratford PW, Kennedy D, Pagura SMC, *et al.* The relationship between self-report and performance-related measures: Questioning the content validity of timed tests. *Arthritis Rheum* Published Online First: 2003. doi:10.1002/art.11196

19 Novy DM, Simmonds MJ, Lee CE. Physical performance tasks: What are the underlying constructs? *Arch Phys Med Rehabil* Published Online First: 2002. doi:10.1053/apmr.2002.27397

20 MacDermid JC, Stratford P. Applying evidence on outcome measures to hand therapy practice. *J Hand Ther* Published Online First: 2004. doi:10.1197/j.jht.2004.02.005

21 Bobos P, MacDermid JC, Walton DM, *et al.* Patient-Reported Outcome Measures Used for Neck Disorders: An Overview of Systematic Reviews. *J Orthop Sport Phys Ther* 2018;**48**:775–88. doi:10.2519/jospt.2018.8131

22 Alreni ASE, Harrop D, Lowe A, *et al.* Measures of upper limb function for people with neck pain. A systematic review of measurement and practical properties. *Musculoskelet Sci Pract* 2017;**29**:155–63. doi:10.1016/J.MSKSP.2017.02.004

23 Moher D, Shamseer L, Clarke M, *et al.* Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* Published Online First: 2015. doi:10.1186/2046-4053-4-1

24 Law MC, MacDermid J. *Evidence-based rehabilitation : a guide to practice*. Thorofare, NJ: : Slack Incorporated 2014.

25 Roy JS, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for musculoskeletal disorders: A systematic review. *J Rehabil Med* Published Online First: 2011. doi:10.2340/16501977-0643

26 Cohen J. Statistical power analysis for the behavioral sciences. Stat. Power Anal. Behav. Sci. 1988. doi:10.1234/12345678

27 Macdermid JC, Walton DM, Avery S, *et al.* Measurement properties of the neck disability index: A systematic review. In: *Journal of Orthopaedic and Sports Physical Therapy*. 2009. 400–16. doi:10.2519/jospt.2009.2930

28 Trippolini MA, Dijkstra PU, Geertzen JHB, *et al.* Construct Validity of Functional Capacity Evaluation in Patients with Whiplash-Associated Disorders. *J Occup Rehabil* Published Online First: 2015. doi:10.1007/s10926-014-9555-0

29 Trippolini MA, Dijkstra PU, Côté P, *et al.* Can functional capacity tests predict future work capacity in patients with whiplash-associated disorders? *Arch Phys Med Rehabil* Published Online First: 2014. doi:10.1016/j.apmr.2014.07.406

30 Trippolini MA, Reneman MF, Jansen B, *et al.* Reliability and safety of functional capacity evaluation in patients with whiplash associated disorders. *J Occup Rehabil* Published Online First: 2013. doi:10.1007/s10926-012-9403-z

31 Reneman MF, Roelofs M, Schiphorst Preuper HR. Reliability and Agreement of Neck Functional Capacity Evaluation Tests in Patients With Chronic Multifactorial Neck Pain. *Arch Phys Med Rehabil* Published Online First: 2017. doi:10.1016/j.apmr.2016.12.005

32 Van Der Meer S, Reneman MF, Verhoeven J, *et al.* Relationship between self-reported

disability and functional capacity in patients with Whiplash Associated Disorder. *J Occup Rehabil* Published Online First: 2014. doi:10.1007/s10926-013-9473-6

33 Reesink DD, Jorritsma W, Reneman MF. Basis for a functional capacity evaluation methodology for patients with work-related neck disorders. *J Occup Rehabil* Published Online First: 2007. doi:10.1007/s10926-007-9086-z

34 Lomond K V., Côté JN. Shoulder functional assessments in persons with chronic neck/shoulder pain and healthy subjects: Reliability and effects of movement repetition. In: *Work*. 2011. doi:10.3233/WOR-2011-1119

35 Pierrynowski M, McPhee C, P. Mehta S, *et al.* Intra and Inter-Rater Reliability and Convergent Validity of FIT-HaNSA in Individuals with Grade Π Whiplash Associated Disorder. *Open Orthop J* 2016;**10**:179–89. doi:10.2174/1874325001610010179

36 Ljungquist T, Jensen IB, Nygren Å, *et al.* Physical performance tests for people with long-term spinal pain: Aspects of construct validity. *J Rehabil Med* Published Online First: 2003. doi:10.1080/16501970306117

37 Ljungquist T, Harms-Ringdahl K, Nygren A, *et al.* Intra- and inter-rater reliability of an 11-test package for assessing dysfunction due to back or neck pain. *Physiother Res Int* Published Online First: 1999. doi:10.1002/pri.167

38 Ljungquist T, Fransson B, Harms-Ringdahl K, *et al.* A physiotherapy test package for assessing back and neck dysfunction--discriminative ability for patients versus healthy control subjects. *Physiother Res Int* Published Online First: 1999. doi:10.1002/pri.158

39 Ljungquist T, Nygren Å, Jensen I, *et al.* Physical performance tests for people with spinal pain - Sensitivity to change. *Disabil Rehabil* Published Online First: 2003. doi:10.1080/0963828031000090579

40 Reneman MF, Dijkstra PU, Westmaas M, *et al.* Test-retest reliability of lifting and carrying in a 2-day functional capacity evaluation. *J Occup Rehabil* Published Online First: 2002. doi:10.1023/A:1020274624791

41 Hegedus EJ, Vidt ME, Tarara DT. The best combination of physical performance and self-report measures to capture function in three patient groups. *Phys Ther Rev* 2014;**19**:196–203. doi:10.1179/1743288X13Y.0000000121

42 Roy J-S, Macdermid JC, Boyd KU, *et al.* Rotational strength, range of motion, and function in people with unaffected shoulders from various stages of life. *Sports Med Arthrosc Rehabil Ther Technol* 2009;**1**:4. doi:10.1186/1758-2555-1-4

43 Kumta P, MacDermid JC, Mehta SP, *et al.* The FIT-HaNSA demonstrates reliability and convergent validity of functional performance in patients with shoulder disorders. *J Orthop Sports Phys Ther* 2012;**42**:455–64. doi:10.2519/jospt.2012.3796

44 Macdermid JCJC, Ghobrial M, Badra Quirion K, *et al.* Validation of a new test that assesses functional performance of the upper extremity and neck (FIT-HaNSA) in patients with shoulder pathology. *BMC Musculoskelet Disord* 2007;**8**:42. doi:10.1186/1471-2474-8-42

45 Hawkes DH, Alizadehkhaiyat O, Fisher AC, *et al.* Normal shoulder muscular activation and co-ordination during a shoulder elevation task based on activities of daily living: An electromyographic study. *J Orthop Res* 2012;**30**:53–60. doi:10.1002/jor.21482

664
665
666
667
668
669

TABLE 1. Summary of Studies Reporting Psychometric Properties of Functional-based Tests in Neck Disorder Patients

Study	Population	Sample Size (n)	Functional Tests	Intervention/Test Interval	Quality
Ljungquist et al. 1999	Neck pain (55%), back pain, multiple pain sites,	53	PILE-C, PILE-L	N/A	Good (68%)
Ljungquist et al. 1999	Neck pain (50%), lumbar pain, thoracic pain, shoulder pain, multiple pain sites,	68	PILE-C, PILE-L, 2 x 20m WWB	8 days	Very Good (79%)
Ljungquist et al. 2003	Neck pain, lumbar pain, thoracic pain, shoulder pain, lower extremity pain, multiple pain sites,	235	PILE-C, PILE-L, 2 x 20m WWB	N/A	Very Good (82%)
Ljungquist et al. 2003	cervical pain (25%), lumbar pain, cervical (25%) and lumbar pain, multiple pain sites,	186	PILE-C, PILE-L, 2 x 20m WWB	6 months	Very Good (79%)
Lomond and Cote. 2011	Chronic neck and shoulder pain (100%)	32	BTEWS II	9.5 days	Very Good (88%)
Pierrynowski et al. 2016	Sub-acute and chronic WAD II	66	FIT-HaNSA	2-7 days	Very Good (88%)
Reesink et al. 2007	N/A	N/A	Neck-FCE	N/A	N/A
Reneman et al. 2017	Chronic multifactorial neck pain	18	Neck-FCE	2 weeks	Good (67%)
Trippolini et al. 2013	Sub acute and chronic WAD I and II	32	WAD FCE	7 days	Very Good (75%)
Trippolini et al. 2014	Sub acute and chronic WAD I and II	267	Workwell FCE	N/A	Excellent (92%)

Trippolini et al. 2015	Sub acute and chronic WAD I and II	314	WAD FCE	N/A	Very Good (86%)
Van der Meer et al. 2013	Chronic WAD I and II	40	Neck FCE	N/A	Very Good (86%)

PILE-C, Progressive Isoinertial Lifting Evaluation-Cervical; PILE-L, Progressive Isoinertial Lifting Evaluation; CBT, Cognitive-Behavioural Therapy; PT, Physical Therapy; NRPS, Numeric Pain Rating Scale; BTEWS II, Baltimore Therapeutic Equipment Work Simulator II; WAD, Whiplash Associated Disorder; MVA, Motor Vehicle Accident; FIT-HaNSA, Functional Impairment Test-Hand and Neck/Shoulder/Arm; FCE, Functional Capacity Evaluation; EXP, Experimental; M, Male; F, Female

TABLE 2. Quality of Studies on Psychometric Properties of Functional-based Tests Evaluated in Neck Disorder Patients

Study	Item Evaluation Criteria												
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Total (%)
Trippolini et al, 2014	2	2	2	2	1	2	2	2	2		1	2	92%
Lomond and Cote, 2011	2	2	1	2	0	2	2	2	2		2	2	88%
Pierrynowski et al, 2016	2	2	1	2	0	2	2	2	2		2	2	88%
Trippolini et al, 2015	2	2	2	0	1	N/A	2	2	2		2	2	86%
Van der Meer et al, 2013	2	1	2	1	2	N/A	2	1	2		1	2	86%
Ljungquist et al 2003 KGV	2	2	2	0	0	N/A	2	2	2		2	2	82%
Ljungquist et al 1999 Rel	2	1	1	2	0	2	2	2	2		1	2	79%
Ljungquist et al 2003 STC	1	1	1	2	1	1	2	2	2		2	2	79%
Trippolini et al, 2013	2	2	1	1	0	0	2	2	2		2	2	75%
Ljungquist et al 1999 KGV	2	1	1	2	0	N/A	2	1	2		1	2	68%
Reneman et al, 2017	1	2	1	1	1	0	1	2	2		2	1	67%

Reesink, 2007*	-	-	-	-	-	-	-	-	-	-	-	N/A

*Paper is not applicable for completion of study quality tool

TABLE 3. Psychometric Properties of the Functional Capacity Evaluation

FCE Battery	Type of Properties	Statistical Test	Value	Interpretation
Neck FCE	Test-retest	ICC	0.39-0.96	Poor-excellent
	Measurement Error	Ratio of LoA	32.0-56.5%	
	Convergent Validity	Pearson or Spearman correlation	NDI total: 0.39-0.62 NDI items: 0.03-0.63	Weak to moderate Very weak to strong
WAD FCE	Test-retest Reliability	ICC	0.66-0.96	Good-excellent
	Convergent Validity	Pearson Correlation	Pain* 0.31-0.39 SFS: 0.42-0.61 NDI: 0.34-0.45 HADS-A: 0.27-0.36 HADS-D: 0.30-0.41	Weak Moderate-strong Weak-moderate Weak Weak-moderate
	Known-groups Validity (German vs Non-German)	Linear Regression Analysis	p<0.001	Significant for All Tasks
	Known-groups Validity (sex)	t-test	p<0.001	Significant for Two Tasks
Workwell FCE	Convergent Validity	Pearson or Spearman Correlation	Work Capacity: 0.1-0.3	Very Weak – weak
	Predictive Validity	Pearson or Spearman Correlation Linear Mixed Model Regression of All Predictors	0.06-0.39 β=-0.04, 95% CI: -0.15 – 0.06 p=0.428 (task 6)	Very weak - Weak Not Significant

FCE, Functional Capacity Evaluation; ICC, Intraclass correlation coefficient; LoA, Limits of Agreement; NDI, Neck Disability Index; Mod., Moderate; Neg., Negligible; SFS, Spinal Function Sort; HADS-A, Hospital Anxiety and Depression Scale – Anxiety; HADS-D, Hospital Anxiety and Depression Scale – Depression; CI, Confidence Interval Sig., Significant
*Pain measured via Numeric Rating Scale

TABLE 4. Summary of Fit-HaNSA's psychometric properties in neck disorder patients

Test	Type of Property	Statistical Test	Value	Interpretation
Fit-HaNSA	Intra-rater Reliability	ICC	0.78	Excellent
Fit-HaNSA	Inter-rater Reliability	ICC	0.84	Excellent
Fit-HaNSA	Measurement Error	SEM LOA_{95} MDC_{90}	76 s 248 s 176 s	
Fit-HaNSA	Convergent Validity	Spearman Rank Correlation	<0.4 - >0.75	Weak – Strong
Fit-HaNSA	Known-groups Validity WAD II vs Control	F-test	62.6, <p,0.001	Significant
Fit-HaNSA Functional Sub-tasks	Intra-rater reliability	ICC	0.70-0.72	Good
	Inter-reliability	ICC	0.54-0.80	Fair - Excellent
	Convergent Validity	Spearman Rank Correlation	<0.4 - >0.75	Weak - Strong
	Known-groups Validity WAD II vs Control	F-test	42.0-53.3, p<0.001	Significant

Fit-HaNSA, Functional Impairment Test, Hand and Neck/Shoulder/Arm; ICC, Intraclass correlation coefficient; SEM, Standard Error of Measurement; LOA_{95}, 95% Limits of Agreement; MDC_{90}, 90% Minimal Detectable Change; WAD, Whiplash Associated Disorder; Mod, Moderate

*Correlations completed with Numeric Pain Rating Scale, Neck Disability Index, Disabilities of Arm, Shoulder, Hand and 6 cervical range of motion tests

TABLE 5. Psychometric Properties of Baltimore Therapeutic Equipment Work Simulator II – Power Output Task

Test	Type of Property	Statistical Test	Value	Interpretation
BTEWS II	Test-retest reliability	ICC	0.53	Fair
		Spearman	0.37	Poor
BTEWS II	Measurement Error	SEM	30.25	
		MDC_{90}	70.59	
BTEWS II	Convergent Validity*	Spearman	Not Reported	Weak
BTEWS II	Known-groups Validity (Pain vs Control)	Two-way Repeated Measures ANOVA	Not Reported	Non-significant

ICC, Intraclass correlation coefficient; SEM, Standard Error of Measurement; MDC_{90}, 90% Minimal Detectable Change; ANOVA, Analysis of Variance

*Spearman correlations completed with Numeric Rating Scale, Neck Disability Index and Shoulder Pain and Disability Index

TABLE 6. Psychometric Properties of performance-based tests included in physiotherapy test package

Test	Type of Property	Statistical Test	Value	Interpretation
PILE-C	Inter-rater Reliability	Mean Difference LoA	-0.24 -2.46 and 1.82	
PILE-C	Inter-rater Reliability	Repeatability (2X SD) % of Range	M=3.93; F=1.19 M=10.5%; F=6.1%	
PILE-C	Convergent Validity	Spearman Correlation	CR-10: 0.55-0.65* Borg RPE: 0.10 - 0.48	Moderate - Strong Very weak - moderate
PILE-C	KGV: spinal pain vs. control	Sensitivity and Specificity	0.93, 0.69	Strong – Very Strong
PILE-C	KGV: spinal pain vs. control	Wilcoxon Sign Ranked Test	p=0.008	Significant
PILE-C	KGV: High vs. low pain intensity	Mann-Whitney U	p=0.003	Significant
PILE-C	KGV: High vs. low Pain behavior	Mann-Whitney U	p=0.005	Significant
PILE-C	KGV: High vs. low perceived exertion	Mann-Whitney U	p=0.154	Non-significant
PILE-C	Sensitivity to Change	Effect Size	Subjects improving: 0.39 - 0.73 Subjects deteriorating: 0 – 0.4	Small – Moderate Trivial – Small
PILE-L	Inter-rater Reliability	Mean Difference LoA	-0.11 -2.33 and 2.11	
PILE-L	Intra-rater Reliability	Repeatability % of Range	M=4.0; F=3.59 M=10.7%; F=18.5%	
PILE-L	Convergent Validity	Spearman Correlation	CR-10: 0.11 – 0.45	Very weak – moderate Very weak – moderate

			Borg RPE: 0.10 - 0.48	
PILE-L	KGV: spinal pain vs no spinal pain	Sensitivity and Specificity	0.85, 0.65	Strong – Very Strong
PILE-L	KGV: spinal pain vs control	Wilcoxon Sign Ranked Test	p=0.002	Significant
PILE-L	KGV: High vs. low pain intensity	Mann-Whitney U	p=0.001	Significant
PILE-L	KGV: High vs. low pain behaviour	Mann-Whitney U	p<0.001	Significant
PILE-L	KGV: High vs. low perceived exertion	Mann-Whitney U	p<0.001	Significant
PILE-L	Sensitivity to change	Effect Size	Subjects improving: 0.02 – 1.08 Subjects deteriorating 0.42-0.81	Trivial – Large Small – Large
2 x 20m WWB	Inter-rater Reliability	Mean Difference LoA	0.05 -1.33 and 1.43	
2 x 20m WWB	Intra-rater Reliability	Repeatability % of Range	3.2 10.7%	
2 x 20m WWB	Convergent Validity	Spearman Correlation	CR-10: 0.55 - 0.65Borg RPE: 0.10 - 0.48	Moderate - Strong very weak – moderate
2 x 20m WWB	KGV: spinal pain vs control	Wilcoxon Sign Ranked Test	p=0.014	Significant
2 x 20m WWB	KGV: High vs. low pain intensity	Mann Whitney U	p<0.001	Significant
2 x 20m WWB	KGV: High vs. low pain behaviour	Mann Whitney U	p<0.001	Significant
2 x 20m WWB	KGV: High vs. low perceived exertion	Mann Whitney U	p<0.001	Significant
2 x 20m WWB	Sensitivity to change	Effect Size	Subjects improving: 0.38-0.78	Small – Moderate Trivial – Moderate

		Subjects deteriorating: 0.13-0.62

PILE-C, Progressive Iso-intertial Lifting Evaluation – Cervical; PILE-L, Progressive Iso-intertial Lifting Evaluation – Lumbar; LoA, Limits of Agreement; SD, Standard Deviation; M, Male; F, Female; RPE, Rating of perceived exertion; KGV, Known-groups Validity; Neg., Negligible; Mod., Moderate, *CR-10: Measurement of pain construct

Figure 1. Selection of the studies for inclusion in the systematic review

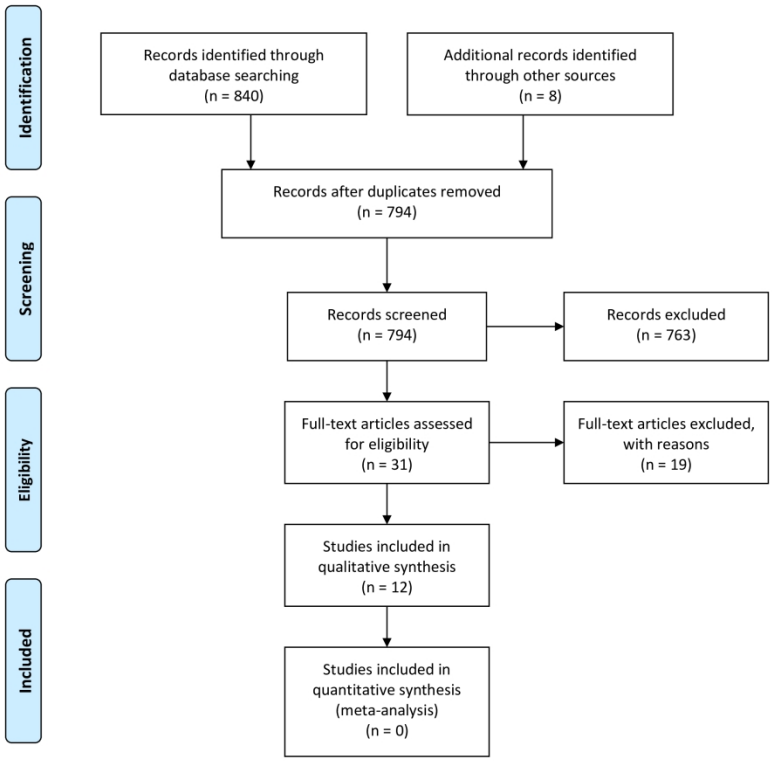

Figure 1

215x279mm (300 x 300 DPI)

Appendix 1: Search terms

EMBASE-OVID

1. exp "outcome and process assessment (health care)"/ or "outcome assessment (health care)"/ or treatment outcome/
2. outcome?.ti.
3. exp "Range of Motion, Articular"/
4. Pain Measurement/
5. exp disability evaluation/
6. "Recovery of Function"/
7. Questionnaires/
8. self-report.tw.
9. ((impairment or disability or function) adj2 (measure? or scale? or evaluation?)).tw.
10. range of motion.tw.
11. (strength adj2 (measure? or scale? or evaluation?)).tw.
12. (outcome? adj2 (measure* or scale? or indicator?)).tw.
13. or/1-12
14. "reproducibility of results"/
15. exp "Sensitivity and Specificity"/
16. reliability.mp.
17. validity.mp.
18. responsiveness.mp.
19. Psychometrics/
20. rasch.mp.
21. factor analysis, statistical/
22. factor analysis.tw.
23. differential functioning.mp.
24. (validity or validation).mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier]
25. (validity or validation).mp.
26. item difficulty.mp.
27. translation.tw.
28. or/14-27
29. 13 and 28
30. Neck Pain/
31. exp Brachial Plexus Neuropathies/
32. exp neck injuries/ or exp whiplash injuries/
33. cervical pain.mp.
34. neckache.mp.
35. whiplash.mp.
36. cervicodynia.mp.
37. cervicalgia.mp.
38. brachialgia.mp.
39. brachial neuritis.mp.
40. brachial neuralgia.mp.
41. neck pain.mp.

42. neck injur*.mp.
43. brachial plexus neuropath*.mp.
44. brachial plexus neuritis.mp.
45. thoracic outlet syndrome/ or cervical rib syndrome/
46. Torticollis/
47. exp brachial plexus neuropathies/ or exp brachial plexus neuritis/
48. cervico brachial neuralgia.ti,ab.
49. cervicobrachial neuralgia.ti,ab.
50. (monoradicul* or monoradicl*).tw.
51. or/30-50
52. exp headache/ and cervic*.tw.
53. exp genital diseases, female/
54. genital disease*.mp.
55. or/53-54
56. 52 not 55
57. 51 or 56
58. neck/
59. neck muscles/
60. exp cervical plexus/
61. exp cervical vertebrae/
62. atlanto-axial joint/
63. atlanto-occipital joint/
64. Cervical Atlas/
65. spinal nerve roots/
66. exp brachial plexus/
67. (odontoid* or cervical or occip* or atlant*).tw.
68. axis/ or odontoid process/
69. Thoracic Vertebrae/
70. cervical vertebrae.mp.
71. cervical plexus.mp.
72. cervical spine.mp.
73. (neck adj3 muscles).mp.
74. (brachial adj3 plexus).mp.
75. (thoracic adj3 vertebrae).mp.
76. neck.mp.
77. (thoracic adj3 spine).mp.
78. (thoracic adj3 outlet).mp.
79. trapezius.mp.
80. cervical.mp.
81. cervico*.mp.
82. 80 or 81
83. exp genital diseases, female/
84. genital disease*.mp.
85. exp *Uterus/
86. 83 or 84 or 85
87. 82 not 86

88. 58 or 59 or 60 or 61 or 62 or 63 or 64 or 65 or 66 or 67 or 68 or 69 or 70 or 71 or 72 or 73 or 74 or 75 or 76 or 77 or 78 or 79 or 87
89. exp pain/
90. exp injuries/
91. pain.mp.
92. ache.mp.
93. sore.mp.
94. stiff.mp.
95. discomfort.mp.
96. injur*.mp.
97. neuropath*.mp.
98. or/89-97
99. 88 and 98
100. Radiculopathy/
101. exp temporomandibular joint disorders/ or exp temporomandibular joint dysfunction syndrome/
102. myofascial pain syndromes/
103. exp "Sprains and Strains"/
104. exp Spinal Osteophytosis/
105. exp Neuritis/
106. Polyradiculopathy/
107. exp Arthritis/
108. Fibromyalgia/
109. spondylitis/ or discitis/
110. spondylosis/ or spondylolysis/ or spondylolisthesis/
111. radiculopathy.mp.
112. radiculitis.mp.
113. temporomandibular.mp.
114. myofascial pain syndrome*.mp.
115. thoracic outlet syndrome*.mp.
116. spinal osteophytosis.mp.
117. neuritis.mp.
118. spondylosis.mp.
119. spondylitis.mp.
120. spondylolisthesis.mp.
121. or/100-120
122. 88 and 121
123. exp neck/
124. exp cervical vertebrae/
125. Thoracic Vertebrae/
126. neck.mp.
127. (thoracic adj3 vertebrae).mp.
128. cervical.mp.
129. cervico*.mp.
130. 128 or 129
131. exp genital diseases, female/

132. genital disease*.mp.
133. exp *Uterus/
134. or/131-133
135. 130 not 134
136. (thoracic adj3 spine).mp.
137. cervical spine.mp.
138. 123 or 124 or 125 or 126 or 127 or 135 or 136 or 137
139. Intervertebral Disk/
140. (disc or discs).mp.
141. (disk or disks).mp.
142. 139 or 140 or 141
143. 138 and 142
144. herniat*.mp.
145. slipped.mp.
146. prolapse*.mp.
147. displace*.mp.
148. degenerat*.mp.
149. (bulge or bulged or bulging).mp.
150. 144 or 145 or 146 or 147 or 148 or 149
151. 143 and 150
152. intervertebral disk degeneration/ or intervertebral disk displacement/
153. intervertebral disk displacement.mp.
154. intervertebral disc displacement.mp.
155. intervertebral disk degeneration.mp.
156. intervertebral disc degeneration.mp.
157. 152 or 153 or 154 or 155 or 156
158. 138 and 157
159. 57 or 99 or 122 or 151 or 158
160. animals/ not (animals/ and humans/)
161. 159 not 160
162. exp *neoplasms/
163. exp *wounds, penetrating/
164. 162 or 163
165. 161 not 164
166. 29 and 165
167. guidelines as topic/
168. practice guidelines as topic/
169. guideline.pt.
170. practice guideline.pt.
171. (guideline? or guidance or recommendations).ti.
172. consensus.ti.
173. or/167-172
174. meta-analysis/
175. exp meta-analysis as topic/
176. (meta analy* or metaanaly* or met analy* or metanaly*).tw.
177. review literature as topic/

178. (collaborative research or collaborative review* or collaborative overview*).tw.
179. (integrative research or integrative review* or intergrative overview*).tw.
180. (quantitative adj3 (research or review* or overview*)).tw.
181. (research integration or research overview*).tw.
182. (systematic* adj3 (review* or overview*)).tw.
183. (methodologic* adj3 (review* or overview*)).tw.
184. exp technology assessment biomedical/
185. (hta or thas or technology assessment*).tw.
186. ((hand adj2 search*) or (manual* adj search*)).tw.
187. ((electronic adj database*) or (bibliographic* adj database*)).tw.
188. ((data adj2 abstract*) or (data adj2 extract*)).tw.
189. (analys* adj3 (pool or pooled or pooling)).tw.
190. mantel haenszel.tw.
191. (cohrane or pubmed or pub med or medline or embase or psycinfo or psyclit or psychinfo or psychlit or cinahl or science citation indes).ab.
192. or/174-191
193. 173 or 192
194. 166 and 193

APPENDICES

APPENDIX A. Data extraction guide for studies evaluating the quality of studies evaluating the clinical measurement properties of outcome measures

Instructions

Clinical measurement studies may evaluate a wide spectrum of measurement properties; or evaluate aspects that relate to the implementability or interpretation of outcome measures. Individual clinical measurement studies cannot address every aspect of the measurement properties of an instrument. Ideally systematic reviews will synthesize the quality and content of research evidence addressing the clinical measurement properties of individual outcome measures. The summative knowledge about the measurement properties, cultural transferability, and utility across different contexts provides the scope of information needed to select an outcome measure for a specific patient (population), purpose and context.

This guide should facilitate extraction of data from individual clinical measurement studies. An explanation of the measurement property addressed in each item and how it might be measured within a given study is listed to facilitate finding and extracting that information. The accompanying extraction form can then be used to collect the specific information on these measurements or utility properties from specific studies.

The purpose of data extraction is to extract the specific information reported by authors within a study, not to evaluate the validity or value of that piece of information. Evaluation of the quality of the published version of the clinical measurement study (also called critical appraisal) is performed in a separate step. See the accompanying critical appraisal tool and guide. It is advisable to extract detailed specific information from the study; recognizing that this information may later be synthesized or subject to meta-analysis.

There is no standardized process for synthesizing clinical measurement information. Based on the findings of extraction you may elect to present the synthesize data in a descriptive way by creating a summary table of the data extracted in each category. If you find some studies with similar designs, you may be able to conduct a meta-analysis of some properties like clinically important difference (CID) or minimal detectable change (MDC); if appropriate given the sample and technique - this can be valuable as it may provide more stable estimates of these important properties.

Population studied		
Population	A description of the study population	Sample size, pathology/disorder, demographics, setting, acute vs. chronic, where subjects were chosen from. Report meaningful demographics and indicators of the population studied.
Intervention	Interventions (if applicable) applied during longitudinal studies	Description of the nature, frequency, intensity of the intervention and the follow-up interval.
Reliability		
Reliability Description	The extent to which the same results are obtained on repeated administrations of the same measure when no change in status has occurred (reliability) or the precision of the scores on repeated measurements (agreement).	Test procedures or measures are typically reapplied on repeated occasions in individuals considered to have a stable condition during that time frame which repeated testing occurs. Repeated testing may be performed on different occasions (test-retest) for self-report measures, OR by the same rater (intra-rater) or different raters (inter-rater) if it is an observer-based scale. In some cases different test instruments (inter-instrument) are evaluated. The most common statistic used is the intraclass correlation coefficient for quantitative data (Shrout & Fleiss, 1979) and kappa (Landis & Koch, 1977) for nominal data. Standard error of measurement is used to present a quantitative estimate of the reliability—in the original units of measure. Report the type of reliability evaluated and coefficients obtained.
Reliability (relative)	The relationship (ratio) between variability in test scores when repeating the test on the same person in comparison to the overall variability (including variation between people)—typically indicated by a reliability coefficient	ICCs (Shrout & Fleiss, 1979) or another reliability coefficient and their associated confidence intervals are extracted.
Reliability (absolute)	Absolute reliability is portrayed as the quantity of error that could be anticipated upon repeated testing - reported in the original units of measure.	This may be reported as 1. Standard error of measurement (in older articles you may see coefficient of variation);

		2. Altman and Bland graphical technique (Bland & Altman, 1990; Bland & Altman, 1987; Bland & Altman, 1986) where the difference on repeated tests for each individual (limits of agreement) is plotted versus their mean score. The mean difference and the boundaries of 2SD are shown to define the limits of agreement.
Minimum Detectable Change	Calculated from the reliability coefficient and the level of confidence specified for error margins. This indicator reflects the amount of change required before you can be confident that change exceeds the random error that occurs in stable patients.	Extract the number and level of confidence.
Content/structural validity		
Internal consistency	The extent to which items on a test or subscale are related (an indication of the consistency of the concept measured).	Cronbach's alpha is the inter-item correlation usually reported. Report alpha and whether it relates to the entire instrument or specific subscales.
Content Validity	The extent to which the conceptual domain or construct that a test is designed to measure is adequately reflected by the items in the measure. In assessing content validity, it is important to consider the population to whom the measure applies, the completeness of the content, the relevancy and emphasis of the content assessed.	A variety of techniques can be used to assess the extent to which items on a given measure reflected the necessary content to capture the concept of interest. Some of the techniques you will find are listed. Extract what was done to determine content validity and what was found. 1) Patients and experts were involved during item selection/reduction - report how they were used and key decisions 2) Patients were consulted for reading and comprehension - report key findings 3) Cognitive interviews (Cibelli, 1994; Ojanen & Gogates, 2006) were done with patients to determine how items were interpreted by respondents, their perceptions of the items - report key findings 4) Expert panels or Delphi procedures were used to select items or evaluate the validity of the instrument - report key findings and decisions

		5) During translation specific study, the meaning of the questions to another cultural or language group was studied - report key findings and decisions 6) ICF linking (Cieza et al., 2002) or other coding of content was performed - report the results which may include the distribution of content across ICF domains, or the distribution of specific codes
Floor-Ceiling Effects	The measure is unable to indicate a worsening score in patients who have clinically deteriorated and/or an improved score in patients who have clinically improved	There are a variety of potential methods; so the method and conclusion should be reported. Descriptive statistics of the distribution of scores that may be presented graphically or numerically may be used to indicate this. Other studies report the percentage of patients sustained a floor or ceiling effect defined by the number of people who fall in the extremes ranges. Note different studies may define the extreme ranges for floor/ceiling differently, so extract how it was defined and % of patients who obtained floor or ceiling category scores.
Factorial validity	The extent to which factor analysis supports assumptions surrounding constructs measured as defined by the measure or as indicated by subscale structure	Factor analysis may be reported as raw results; or compared to the inherent structure of the instrument or factor analysis upon which its construction was based. Report the type of factor analysis performed (exploratory or confirmatory), rotations used and the number of factors derived; specify whether this confirms the expected instrument structure or original factor structure.
Item response /Rasch Analyses	The extent to which items cross a range of difficulty, or a spectrum of the concept measured. The measurement scaling of the items.	Using item response theory or Rasch analysis, items are fit to a model to demonstrate interval scaling and determine item difficulty (Pallant & Tennant, 2007). Analyses might address item difficulty, person's ability curves, and comparison of ability estimation. Most commonly, the item difficulty and the composition of the test that fulfills interval scaling are defined. Data to be extracted include information on the scaling of the items, whether the interval scaling has been established; and the presence or absence of differential item functioning

		(DIF), where items perform differently on different types of respondents.
Construct Validity		
Construct Validity - correlational	Constructs are artificial frameworks that are not directly observable. Construct validity assesses the extent to which measures perform according to a priori defined constructs. Construct validity can be cross-sectional or longitudinal (predictive). Constructed hypotheses can assess convergent validity where measures are thought to represent similar constructs or divergent validity where it is assumed they measure different constructs. For cross-cultural validation, the expected relationships are those that have been reported in validation of the instrument in its original language/format.	When extracting data about correlational validity, the pre-constructed hypothesis and whether it is supported should be documented. For correlational construct validity, this will be the nature and strength of the prespecified relationship and the correlations that support that. Relation to other indices/constructs that are similar (convergent) or different (divergent) can be reported. Ideally, hypotheses are formulated/reported and supported by correlations that are in accordance with the hypotheses. Note that there is no consistent agreement on what subjective term should be applied to validity correlations. Note that there is no consistent agreement on what subjective term should be applied to validity correlations. Some authors use subjective terminology defined for reliability such as: strong (>0.70) and moderate (0.40-0.70) correlations; others use the correlations like effect size benchmarks that 0.4 indicates a moderate effect and 0.6 a large effect. For validity assessment is more important than correlations prespecified constructed hypotheses, although not all papers are written clearly with respect to this.
Convergent	The Relationship between similar scales/tests. Correlations are generally expected to be moderate to strong if the relationship is one where there is confidence that they measure a similar construct.	Extract test names, prespecified expected relationship and correlations observed.
Divergent	Divergent validity assesses the extent to which different scales/tests that are designed to	Extract test names, prespecified expected relationship and correlations observed.

	measure different constructs demonstrate that they are different by a lack of correlation between them.	
Construct validity - known groups	Known groups analysis supports the validity of a measure by demonstrating that the measurement is able to differentiate between groups that are prespecified and known to be different on the construct being assessed.	Data extraction should include the nature of the subgroups and the size of the difference observed between them (and its statistical significance). Typically, statistical tests of difference are performed. Since known groups analysis can provide data that is useful in clinical practice as benchmarks for comparing these known groups, it is a more practical form of construct validity than correlational. Data extraction/presentation should reflect this by presenting the group central tendency, their margins and statistical significance in an accessible manner.
Longitudinal Validity	This form of validity supports the validity of a measure by demonstrating that the change that occurs over time onto similar instruments is correlated in a manner consistent with the nature of the relationship between the scales. It is measured over a retest interval when clinically relevant change could be expected.	Extract test names and correlations Note: since longitudinal validity is based on four measures (pre-and post-test on two different measures), and since error tends to mitigate the strength of correlations, strong longitudinal correlations can be difficult to obtain.
Criterion validity Description	Criterion validation is determined by comparing a given outcome measure to an accepted standard of measure. For subjective constructs like pain and disability, it can be argued that there is no criterion since there is no external gold standard. Therefore, for self-report measures, validation focuses on construct validity. For performance measures, it is common to have a criterion measure that is considered to be highly precise and rigorous as the criterion comparator.	Authors will state that their measure is being compared against a specific instrument and report the correlation or agreement between the measures. Extract the test names and results: correlations or other as reported.
Concurrent criterion	Concurrent validity is assessed by comparing a scale and its criterion at a single point in time	Extract the test names and correlations.

Predictive criterion	Predictive validity is evaluated by determining the extent to which the results of administering an outcome measure at one point in time can accurately predict a future status or outcome.	Extract the test names and correlations and time interval. (and important cutoffs if those were established/reported), if diagnostic test methodology was used to examine prediction, and sensitivity specificity and other diagnostic criteria were reported, they should be extracted.
Responsiveness/Clinical Change		
Responsiveness	Does the instrument detect changes over time that matters to patients?	Extract indicators of responsiveness include: effect size, standard response mean and the method for assessing whether patients were improved, stable or worse. (Beaton, 2000)
Clinically Important Difference (CID)	CID is the difference in scores that patients find to be observable and clinically important. It is assessed by comparing scores to an external benchmark of clinical relevance such as a global rating of change or some other method. The terminology used to rate the nature of this difference will affect the estimation process. Differences in methods include how clinically importance is framed and the metrics/process by which that is determined.	Extract the MID or CID and note the method/cut-off used to define importance. Extract how the clinically important differences were framed to respondents; or determined. For example, minimal, moderate, extreme improvement or better/not better, etc.

APPENDIX B. Data extraction form for studies evaluating the clinical measurement properties of outcome measures

Authors: _____ Year:_____ Rater: _____

Instructions

When using the data extraction form, it is important to realize that the purpose of data extraction is to remove or extract the specific information reported by authors within a study, not to evaluate the validity or value of that piece of information. To make data extraction as useful as possible, and to avoid the need for repeated data extractions, it is advisable to read the accompanying guide and then be as specific as possible when extracting information.

	DATA EXTRACTED
Population studied	
Population	
Intervention	
Reliability	
Reliability (relative)	
Reliability (absolute)	
Minimum Detectable Change	
Content/structural validity	

Internal consistency	
Content Validity	
Floor-Ceiling Effects	
Factorial validity	
Item response /Rasch Analyses	
Construct/Criterion Validity	
Known groups	
Convergent	
Divergent	
Longitudinal Validity	
Concurrent criterion	
Predictive criterion	
Responsiveness/Clinical Change	

Responsiveness	
Minimally Clinical Important Difference	

APPENDIX C. Quality Appraisal for Clinical Measurement Research Reports Evaluation Form

Rater (Group)_____
Author(s) (Study Author(s) _____
Year (Year of publication)_____

1. Was the relevant background work cited to define what is currently known about the measurement properties of measures under study, and the potential contributions of the current research question to informing that knowledge base?
   2
   1
   0
2. Were appropriate inclusion/exclusion criteria defined? *
   2
   1
   0
3. Were specific clinical measurement questions/hypotheses identified?
   2
   1
   0
4. Was an appropriate scope of measurement properties considered?
   2
   1
   0
5. Was an appropriate sample size used?
   2
   1
   0
6. Was appropriate retention/follow-up obtained? (for studies involving retesting; otherwise n/a)

2
1
0

7. Were specific descriptions provided of the measure under study and the method(s) used to administer it?

2
1
0

8. Were standardized procedures used to administer all study measures in a manner that minimized potential sources of error/bias (including the study measure and its comparators)?

2
1
0

9. Were analyses conducted for each specific hypothesis or purpose?

2
1
0

10. Were appropriate statistical tests performed to obtain point estimates of the measurement properties?

2
1
0

11. Were appropriate ancillary analyses done to quantify the confidence in the estimates of the clinical measurement property (Precision/Confidence intervals; benchmark comparisons/ROC curves, alternate forms of analysis like SEM/MID, etc.)?

2
1
0

12. Were clear, specific and accurate conclusions made about the clinical measurement properties; that were associated with appropriate clinical measurement recommendations and supported by the study objectives, analysis and results?

2

1
0

Subtotals (of column 1 and 2) Total Score (sum of subtotals/24*100)

APPENDIX D. Description of each performance battery from selected articles

Battery	Description of Tasks
Relevant FCE Subtasks[25,26,27,28,29,30]	Material Handling Tasks: All lifting tests were executed with a wooden crate (40 × 30 × 26 cm) of 2.5 kg, and four to five weight increments of 2.5 kg or 5 kg each were used until the maximum amount of weight was reached. Maximum performance was recorded in kg. Lifting floor to waist: Measured after five lifts of crate from floor to table and vice versa (time limit < 90 s): hands remained on the crate during the test. Increase weight in 4-5 steps until maximum is reached Overhead lift test: Five lifts from waist to crown height and vice versa within 90 s in standing position. Increase weight in 4–5 steps until maximum is reached Two-handed carrying: Carrying of a crate for a short distance measured after five carries of 1.5 m distance at waist height. Hands remain on the crate during the test. One-handed carrying: Carrying wooden crate for 15 m within 90 s beginning with the right hand and thereafter the left hand. Overhead working: Standing with hands at crown height for manipulation of nuts and bolts. The time that the position was held is recorded (sec). Repetitive reaching: fast horizontal movements of the upper extremity in a sitting position. Marbles are removed from bowls at arm length distance at table height from left to right and vice versa, with right and then left arm. The time taken to remove 30 marbles is recorded (sec). Overhead lift test: Five lifts from waist to crown height and vice versa within 90 s in standing position. Increase weight in 4–5 steps until maximum is reached

	Repetitive bending and overhead reaching: 20 marbles in 2 bowls at table height and crown height. Standing in front of bowl of marbles and moving the marbles as fast as possible from table height to crown height.
A Physiotherapy Test Package[33,34,35,36]	**PILE Tests:** "The lifting tests were performed standing in front of bookshelves with shelves at 0.76m and 1.37 m from the floor. Subjects were asked to lift weights in a plastic box from floor to waist level (0–0.76 m) for the lumbar PILE test, or from waist to shoulder height (0.76–1.37 m) for the cervical PILE test. The initial weight was 3.6 kg for women and 5.9 kg for men. A 'lifting movement' involved a single transfer from one level to the next and back again. After every four such lifting movements (= 20 s), the weight was increased by 2.25 kg for women and 4.5 kg for men. The weight managed during the last lifting movement was recorded and used as a test result, as well as this maximum weight divided by the 'adjusted weight'". **2x20m WWB:** "Subjects were asked to walk 20 m at a comfortable speed along a corridor, to turn around where 20 m was marked and then to walk 20 m back to the starting point. In the first walking test they carried no extra weight, but in the second they carried one carrier bag in each hand, containing 4 kg each for the women, 8 kg each for the men. The time taken was recorded to get the walking speed. The tests were discontinued after 50 s".
BTEWS II[31]	"The protocol consisted of performing a series of shoulder functional tasks before and after a fatiguing activity. Functional tasks consisted of active shoulder range of motion (ROM) in both flexion and abduction and cumulative power output (PO) accumulated over 10s during a repetitive pushing/pulling task in a horizontal plane at shoulder level".
FIT - HaNSA[32]	"The FIT-HaNSA protocol consists of three timed tasks and each task is performed for a maximum of 300 seconds (s) with approximately 30 s pause between them (set-up time for next task). Task 1 (waist-up) requires the patient to alternately "grab, lift, move, and place" three 1000 g containers located on waist level and 25 cm above waist level shelves, using their affected arm, at a metronome pace of 60 beats per minute for 300 s or until they felt unable to continue. The time to complete Task 1 is measured using a stopwatch. Task 2 (eye-down) is identical to Task 1 except that the two shelves are placed at eye-level and 25 cm below. Task 3 (overhead work) requires a patient to repeatedly screw and unscrew bolts in a sagittal plane oriented plate

	positioned at eye-level using both arms". More complete description at https://srs-mcmaster.ca/wp-content/uploads/2015/04/FIT-HaNSAProtocol_April2007.pdf

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	2
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	3
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	3
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	4
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	4
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	3-4
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	3-4
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	4
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	4
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	5
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	NA
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	NA

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	NA
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	NA
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	6-7
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	6-7
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	6-10
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	6-10
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	6-10
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	6-10
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	NA
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	11-13
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	14-16
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	16
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	18

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.

BMJ Open

Systematic Review of the Measurement Properties of Performance-based Functional Tests in Patients with Neck Disorders

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-031242.R2
Article Type:	Original research
Date Submitted by the Author:	08-Oct-2019
Complete List of Authors:	McGee, Steven; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Sipos, Taylor; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Allin, Thomas; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Chen, Celia; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Greco, Alexandra; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Bobos, Pavlos; Western University, Health and Rehabilitation Sciences; University of Toronto, Dalla Lana School of Public Health, Institute of Health Policy Management and Evaluation MacDermid, Joy ; Western University, School of Physical Therapy, Health and Rehabilitation Sciences Group, CATWAD; Michele Sterling, Anne Söderlund, Michele Curatolo, Jim Elliott, David M Walton, Helge Kasch, Linda Carroll, Hans Westergren, Samuel McLean, Gwendolen Jull, Genevieve Grant Luke Connelly, Joy C MacDermid, Mandy Nielsen, Pierre Côté, Tonny Elmose Andersen, Trudy Rebbeck Annick Maujean, Sarah Robins, Kenneth Chen, Julia Treleaven
Primary Subject Heading:	Rehabilitation medicine
Secondary Subject Heading:	Rehabilitation medicine
Keywords:	functional, psychometric properties, neck pain, cervical, outcome measures

Title: Systematic Review of the Measurement Properties of Performance-based Functional Tests in Patients with Neck Disorders

[1]Steven McGee, PT

[2]Taylor Sipos, PT

[3]Thomas Allin, PT

[4]Celia Chen, PT

[5]Alexandra Greco, PT

[6]Pavlos Bobos, PT, PhD(c) (corresponding author)

[7]Joy MacDermid, PT, PhD

[8]CATWAD

Authors' information

[1]Steven McGee PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (smcgee7@uwo.ca)

[2]Taylor Sipos PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (jsipos@uwo.ca)

[3]Thomas Allin PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (tallin@uwo.ca)

[4]Celia Chen PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (qchen224@uwo.ca)

[5]Alexandra Greco PT, School of Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada, (agreco33@uwo.ca)

[6]Pavlos Bobos PT, PhD(c), (corresponding author) Doctoral Candidate, Western's Bone and Joint Institute, Department of Health and Rehabilitation Sciences, Western University, Elborn College, 1201 Western Road, N6G 1H1, London, Ontario, Dalla Lana School of Public Health, Institute of

Health Policy Management and Evaluation, Department of Clinical Epidemiology and Health Care Research, University of Toronto, Canada, (pbobos@uwo.ca), tel: +1 519 661 2111 x88912

[7]Joy C MacDermid BScPT, PhD, Professor, Physical Therapy and Surgery, Western University, London, ON and Co-director Clinical Research Lab, Hand and Upper Limb Centre, St. Joseph's Health Centre, London, Ontario; Professor Rehabilitation Science McMaster University, Hamilton, ON, Canada (jmacderm@uwo.ca)

[5]CATWAD Coauthors: Michele Sterling m.sterling@uq.edu.au, Anne Söderlund anne.soderlund@mdh.se, Michele Curatolo curatolo@uw.edu, Jim Elliott j-elliott@northwestern.edu, David M Walton dwalton5@uwo.ca, Helge Kasch helgkasc@rm.dk, Linda Carroll linda.carroll@ualberta.ca, Hans Westergren Hans.Westergren@skane.se, Samuel A McLean, Samuel_McLean@med.unc.edu, Gwendolen Jull g.jull@uq.edu.au, Genevieve Grant genevieve.grant@monash.edu, Luke Connelly l.connelly@uq.edu.au, Joy C MacDermid, jmacderm@uwo.ca, Mandy Nielsen mandy.nielsen@griffith.edu.au, Pierre Cote pierre.cote@uoit.ca, Tonny Elmose Andersen tandersen@health.sdu.dk, Trudy Rebbeck trudy.rebbeck@sydney.edu.au, Annick Maujean a.maujean@uq.edu.au, Sarah Robins s.robins1@uq.edu.au, Kenneth Chen k.chen8@uq.edu.au, Julia Treleaven j.treleaven@uq.edu.au

Key Words: functional, psychometric properties, neck, cervical, outcome measures

Word Count: 4509

Abstract

Objectives: The purpose of this systematic review is to identify and synthesize studies evaluating performance-based functional outcome measures designed to evaluate the functional abilities of patients with neck pain.

Design: Systematic review

Data Sources: A literature search using PubMed, Scopus, CINAHL, EMBASE, COCHRANE, Google Scholar, and a citation mapping strategy was conducted till July 2019

Eligibility criteria: More than half of the study's patient population had neck pain or a musculoskeletal neck disorder and completed a functional-based test. Clinimetric properties of at least one performance-based functional tests were reported. Both traumatic and non-traumatic origins of neck pain were considered.

Data extraction and synthesis: Relevant data were then extracted from selected articles using an extraction guide. Selected articles were appraised using the Quality Appraisal for Clinical Measurement Research Reports Evaluation Form (QACMRR).

Results: The search obtained 12 articles which reported on 4 outcome measures (Functional Capacity Evaluations (FCE), Baltimore Therapeutic Equipment Work Simulator II (BTEWS II), Functional Impairment Test- Hand and Neck/Shoulder/Arm (FIT-HaNSA)) and a physiotherapy test package, to assess the functional abilities in patients with mechanical neck pain. Of the selected papers: 1 reports content validity, 5 construct validity, 4 reliability, 1 sensitivity to change, and 1 both reliability and construct validity. QACMRR scores ranged from 68% to 95%.

Conclusions: This review found very good quality evidence that the FIT-HaNSA has excellent inter and intra-rater reliability and very weak to weak convergent validity. Excellent quality evidence of fair test-retest reliability, weak convergent validity, and very weak known

groups validity for the BTEWS II test was found. Good to excellent quality evidence exists that an FCE battery has poor to excellent reliability and very weak to strong validity. Good to excellent quality of weak to strong validity and trivial to strong effect sizes were found for a physiotherapy test package.

Prospero registration: CRD42018112358

Strengths and limitations of this study

- The psychometric properties of performance outcome measures for neck pain were synthesized and critically appraised
- This study assessed the risk of bias and the quality of measurements properties
- The feasibility or usability of these tools was not assessed

Introduction

Neck pain has been associated with high disability and is regarded as a substantial societal burden.[1] Approximately 70% of people experience neck pain within their lifetime and about 33% of adults experience neck pain every year.[2,3] Further concern is warranted as it has been suggested that the incidence of neck pain is increasing.[4–6] The economic burden due to neck disorders is high, including lost wages, costs of treatment, and compensation expenditures to injured people.[7,8] Neck pain is second only to low back pain in annual workers' compensation costs in the United States and has been associated with many other comorbidities such as headaches, anxiety, depression, back pain and arthralgias.[6,9,10]

Outcome measures are a crucial component in monitoring patients with neck pain to determine the effects of treatment[11,12], evaluation of interventions, guiding return to work, and justifying treatment.[13,14] Several self-reported outcome measures currently exist to assess disability and function in those with neck pain (e.g. the Neck Disability Index - NDI). [13] Evidence-based clinical practice guidelines suggest that measures assessing physical performance should also be used for people with neck pain.[15] Performance-based testing is where the assessment is based on actual performance of a task or activity. Physical performance can be assessed by testing a person's ability to execute a standardized activity in a standardized environment (i.e. clinical setting).[16] Time to complete the activity, number of repetitions performed, and weight lifted are frequently used to quantify the physical performance.[17] Conversely, self-report measures examine patients' perception and experience of their ability to perform functional tasks. [16] Previous research has demonstrated poor to fair relationships between physical performance and self-report measures of ability in patients with various musculoskeletal disorders suggesting that these measures assess different constructs of function. [17,18] Consequently, physical performance tests and self-report measures complement each other and may each contribute unique information about a patient's function. [19]

A fundamental component of monitoring outcomes is having reliable and valid tools with known measurement properties.[13,20] While recent research has investigated the psychometric properties of patient-reported outcomes in people with neck pain [13,21] there is a gap in knowledge with respect to performance-based functional outcomes. The purpose of this systematic review was to identify and synthesize clinical measurement studies that evaluate measurement properties of performance-based functional tests in patients with neck disorders.

METHODS

Patient and Public Involvement

There was no patient or public involvement in the design or planning of this study.

Study Design and Protocol Registration

We conducted a systematic review to evaluate the psychometric properties of performance-based functional tests for people with mechanical neck disorders. The protocol was registered in PROSPERO register with registration number CRD42018112358.

Search Strategy

A database search using CINAHL, PubMed, Scopus and Google Scholar was performed to identify articles published till July 2019. The following search strategy was used to search all databases for eligible studies: (Reliability OR validity OR responsiveness OR calibration OR validation) OR (minimal detectable change) OR (clinically important difference) OR (psychometric properties) AND cervical OR neck OR c-spine AND (performance measure) OR (functional test) OR (functional outcome) OR (performance outcome). MeSH terms were searched in PubMed. A citation map of articles and systematic reviews selected for the full-text review was performed. This strategy was included to minimize the risk of publication bias. The full search strategy is summarized in **APPENDIX 1.** The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) process[22] was followed to ensure all appropriate steps were taken in the selection process (**FIGURE 1**).

Inclusion Criteria

Articles were included in the final review if all of the following criteria were met:

- >50% of the study's patient population had neck pain or a musculoskeletal neck disorder (e.g. whiplash associated disorder (WAD II))
- Patients in the study completed a functional-based test
- Clinometric properties of at least one performance-based test were reported.

A test was considered functional-based if it met the following criteria:

- assessment of a patient's ability to execute a standardized activity in a standardized environment
- tests assessing muscular endurance (e.g. cervical flexion test) or proprioception were not deemed functional-based as they are often not reflective of physical working conditions.

Both traumatic and non-traumatic origins of neck pain were considered. Definitions for the properties can be found in **APPENDIX A.**

Article Selection

Titles and abstracts generated by the search strategy were screened by two authors (SM and PB) independently. Articles that met the inclusion criteria and selected for a full text review were also reviewed in pairs of authors. Disagreements were resolved by the most experienced author (JCM)

Data Extraction

Data extraction and critical appraisal was performed in pairs of two raters among the authors, after the completion of a calibration session in which the most experienced author (JCM) reviewed the data extraction tools with the authors that performed the data extraction. When reviewers disagreed

during data extraction and/or critical appraisal, and consensus could not be met, a third author arbitrated. A data extraction form [23] (**APPENDIX A and APPENDIX B**), developed by one of the authors (JCM.), was used to ensure systematicity. Authors extracted sample size, patient population characteristics, functional tests performed and reported psychometric properties. The interpretation of ICC was as follows: ICC < 0.50 indicating poor, 0.50 ≤ ICC < 0.75 indicating moderate, 0.75≤ICC <0.9 indicating good, and ICC ≥ 0.9 indicating excellent reliability were used as a common benchmark. [24] For validity estimates, correlation coefficient (Pearson's/Spearman) and the 95% confidence intervals were extracted if were available. [23,25] Evan's guidelines to interpret the strength of the correlation was used which included: 0.00–0.19 "very weak", 0.20–0.39 "weak", 0.40–0.59 "moderate", 0.60–0.79 "strong", and 0.80–1.00 "very strong".[26] To assist clinical decision making, standard benchmark scores of trivial (< 0.20), small (≥ 0.20 to < 0.50), moderate (≥ 0.50 to < 0.80) or large (≥ 0.80), as proposed by Cohen, were used. [27] For studies assessing construct validity specifically, results in accordance with pre-defined hypotheses were evaluated to interpret the findings.

Quality Appraisal for Clinical Measurement Research Reports Evaluation Form

Pairs of authors critically appraised the quality of each study using a standardized 12-item evaluation tool (QACMRR) designed to assess the quality of studies determining measurement properties in outcome measures (**APPENDIX C**). If disagreement was present a third person (JM) assist in resolving the discrepancy. [23] This tool has been found to have moderate to excellent pre-consensus inter-rater reliability (ICC: 0.69-0.91, κ = 0.62–1.00) across a number of systematic reviews.[23,25,28] The evaluation criteria of this tool included twelve items: 1) Thorough

literature review to define the research question; 2) Specific inclusion/exclusion criteria; 3) Specific hypotheses; 4) Appropriate scope of psychometric properties; 5) Sample size; 6) Follow-up; 7) The authors referenced specific procedures for administration, scoring, and interpretation of procedures; 8) Measurement techniques were standardized; 9) Data were presented for each hypothesis; 10) Appropriate statistics-point estimates; 11) Appropriate statistical error estimates; and 12) Valid conclusions and recommendations. [23,25] Each item is scored from 0 to 2 with (score=2) is the best; (score=1) is acceptable but suboptimal; (score=0) is not done/documented, substantially inadequate or inappropriate. An article's total score – quality - was calculated by the sum of scores for each item, divided by the numbers of items and multiplied by 100%. [23,25] Overall, the quality summary of appraised articles ranges from (0%-30%) Poor, (31%-50%) Fair, (51%-70%) Good, (71%-90%) Very Good, and (>90%) Excellent

RESULTS

The search strategy resulted in 840 published articles. After duplications were removed, 31 articles were deemed relevant and were screened at full text. Overall, 12 articles met our inclusion criteria (**FIGURE 1**). The excluded articles were removed due to inappropriate patient populations, investigations into self-report measures or tests assessing proprioception/muscular endurance rather than functional-based measures, or because the articles were found to be systematic reviews. The characteristics of the included studies and the summary of psychometric properties are presented in **TABLE 1.** The quality assessment is summarized and presented in **TABLE 2**. Percent agreement was calculated for quality scores between the 2 raters and it was 90%.

Participants

Participants in the selected articles had various types of neck pain including subacute, chronic, and whiplash-associated disorder. The mean/median age of the samples of each study ranged from 30 to 48 years of age. The proportion of females in each article ranged from 34-78% of the study population. Two studies that had a mixed sample of subjects with various spinal pain did not report the demographics of the neck pain portion of their sample. One study did not contain any subjects and performed a review of epidemiological literature to establish content validity for work-related neck disorders **TABLE 1.**

Functional-Based Tests

The 12 articles that were included for review provided properties on the following functional based tests: Functional Capacity Evaluations (FCE)[29–34], The Baltimore Therapeutic Equipment Work Simulator II (BTEWS II) [35], Functional Impairment Test- Hand and Neck/Shoulder/Arm (FIT-HaNSA) [36], as well as items off of a physiotherapy test package including a cervical and lumbar Progressive Isoinertial Lifting Evaluation (PILE-C, PILE-L) test [37–40] and 2 x 20 m with burden walking test (2x20M-WWB) [37–40]. Descriptions of all functional-based tests and their relevant subtasks are provided in **APPENDIX D.**

Functional Capacity Evaluations (FCE)

Six articles reported measurement properties for an FCE battery. We identified multiple versions of the FCE in the literature with one article reporting properties on the Workwell FCE [30], two reporting on the Whiplash Associated Disorder (WAD) FCE [29,31] and three reporting

on the neck-FCE.[32–34] These test batteries include various combinations of muscular strength, endurance and functional based tests. The measurement properties of the functional based tests used by the FCE are outlined in **TABLE 3.**

Individuals with Sub-acute to chronic WAD

Trippolini et al. (2014)[30] evaluated the Workwell FCE test-retest reliability, measurement error, convergent validity and predictive criterion validity of future work capacity in workers diagnosed with WAD I or II. Interclass Correlation Coefficients (ICC) ranged from 0.66 to 0.96 (moderate to excellent). Limits of agreement relative to mean performance ranged from 21 to 57% for functional based sub-tests. Correlations between FCE sub scores and baseline work capacity were very weak to weak ranging between r=0.06 and r=0.39. FCE sub scores did not predict future work capacity at 1, 3, 6 and 12 months.

Trippolini et al. (2015)[29] assessed the WAD FCE (31) and evaluated convergent validity and known-groups validity. FCE subscales showed very weak to strong correlations (0.15-0.68) with each of: pain, self-reported functional ability, self-reported disability, anxiety and depression. It was found that the FCE had known-group sex validity (males vs females) for 1 of 3 functional subtests (lifting waist-overhead) and reported significant performance differences between culture groups (German vs non-German language groups). To test construct validity, 29 a priori formulated hypotheses were tested, 4 related to gender differences, 20 related associations with other constructs, 5 related to cultural differences. In total 23 out of 29 hypotheses were confirmed (79 %).

Work-Related Neck Disorders

Reesink et al. (2007)[34] developed an independent FCE for patients with musculoskeletal neck disorders (neck FCE). They performed a review of epidemiological literature and identified four physical risk factors for work-related neck disorders and used that information to develop an FCE consisting of eight functional-based tests. Content validity was established by following operational definitions of the risk factors when searching the literature and using current literature to provide a rationale to guide their development of the tasks comprising the FCE.

Chronic Neck Pain

Reneman et al. (2017)[32] measured test-retest reliability of the subscales of the neck FCE in patients with multifactorial neck pain. Test-retest ICC's ranged from poor to excellent (0.39-0.96). Limits of agreement relative to mean performance range from 32.0% to 56.5% for functional based sub tests. Convergent validity was performed against the Neck Disability Index (NDI) items and total score.[33] The authors found weak to strong Pearson correlations (0.39-0.70) for the FCE sub scores to both NDI individual items and the NDI total score.

The Baltimore Therapeutic Equipment Work Simulator II (BTEWS II)

Chronic Neck Pain

Lomond and Côté, (2011)[35] reported on the reliability, measurement error, minimum detectable change (MDC) and validity of the power output (PO) task during the BTEWS II test in patients with chronic neck and shoulder pain (**TABLE 4**). Test-retest reliability, measured with Spearman Rank correlations and ICC's was moderate and measured at ρ=0.37 and $ICC_{2,1}$ = 0.54, respectively. The standard error of measurement (SEM) and the minimal detectable change at 90% confidence (MDC_{90}) for the PO task were measured as 30.25 and 70.59, respectively. Weak

Spearman Rank correlations between the PO task and the NDI, Shoulder Pain and Disability Index (SPADI) and Numeric Rating Scale (NRS) for pain tests were recorded. There were no significant performance differences between control and pain groups for the PO task.

Functional Impairment Test- Hand and Neck/Shoulder/Arm (Fit-HaNSA)

Sub-acute to chronic WAD

Pierrynowski et al. (2016)[36] reported on the reliability, measurement error, MDC and validity of the Fit-HaNSA test in a sample of people with WAD II following motor vehicle collision (MVC) (**TABLE 5**). Intra-rater reliability ICC's for patient subtask and total scores were moderate to good ranging between 0.70-0.78. [36] Inter-rater reliability ICC's for patient subtask and total scores were moderate to good and ranged between 0.54-0.84. [36] The Bland and Altman plot for the patient group showed a 26 seconds (s) bias in terms of improved performance on the second test (possible learning effect). The standard deviation of difference was 124 s and 95% Limits of Agreement (LoA_{95}) was 248 seconds. [36] The SEM for people with WAD II was reported to be 76 s. The MDC_{90} was measured as 176 s. [36]

Spearman rank correlations were also calculated between the Fit-HANSA, Numeric Pain Rating Scale (NPRS), NDI, the disabilities of arm, hand and shoulder (DASH) and 6 cervical range of motion measures. Most (59 of 78) of the correlations between performance and comparator measures were very weak to weak (r=<0.4). [36] All correlations between total Fit-HaNSA scores and subtask scores had good correlations (r=<0.75), except for Task 1-Task 3. [36] Significant performance differences between WAD II and control groups (known group validity) were recorded for the total Fit-HaNSA score and all 3 subtask scores. [36]

Physiotherapy Test Package Subtests

Ljungquist et al. published a series of articles[37–40] which evaluated the clinimetric properties of a physiotherapy test package for patients with spinal pain (**TABLE 6**). This package included muscular strength & endurance tests, submaximal endurance tests, and three functional tests. These functional tests included the PILE-C, PILE-L, and 2x20M-WWB test. Ljungquist's series of articles reported on convergent validity, known-groups validity, reliability, measurement error and sensitivity to change for these tests. [37–40]

Undetermined duration of neck pain

In a 1999 article [39], correlations between the tests of the package and pain (CR-10) and perceived exertion (Borg RPE) were determined. All correlations were very weak to moderate (0.10-0.48) except for moderate to strong correlations (0.55-0.65) between the PILE-C test and pain intensity and between 2x20M-WWB test and pain intensity.

In a 2003 article[37], the PILE-C, PILE-L and 2x20M-WWB tests were tested to determine their ability to discriminate between known-groups (neck pain vs back pain). Subjects with spinal pain completed the CR-10, the University of Alabama Pain Behavior scale (UAB) and the Borg RPE test. Specific cut points were used to distinguish patients with high vs. low pain intensity, high vs. low pain behavior, and high vs. low perceived exertion in patients, respectively. Participants then completed the test package and it was determined if each subtest could discriminate between participants with high vs. low pain intensity. The PILE-C and the 2x20M-WWB tests were hypothesized to be more difficult for persons with neck pain and the PILE-L was hypothesized to be more difficult for persons with back pain. Subjects with neck pain performed worse on the PILE-C test compared to those with back pain. Subjects with back pain did not

perform worse than those with neck pain on the PILE-L test and subjects with back pain performed worse on the 2x20M-WWB test.

The functional tests were able to discriminate between all 3 subgroups with the exception of the PILE-C being unable to discriminate between participants with high vs. low perceived exertion.

In a paper from 1999[39], the PILE-C, PILE-L and 2x20M-WWB tests were found to have significant discriminative abilities in distinguishing healthy subjects from patients with spinal pain. The sensitivity and specificity for this known group discrimination for the PILE-C test, were reported to be 0.93 (very strong) and 0.69 (strong), respectively. The sensitivity and specificity for the PILE-L test were reported to be 0.85 (very strong) and 0.65 (strong), respectively.

The inter and intra rater reliability were tested on participants with spinal pain.[38] Limits of agreement were used to measure inter rater reliability and repeatability, defined as 2x the within-subject standard deviation of each variable. Interrater agreement for 2 tests was deemed "acceptable", while all 3 functional tests had "clinically acceptable" intra-rater reliability.

Sensitivity-to-change was evaluated in the test package following 6 months of a physiotherapy intervention. Using ROC curves, Wilcoxon sign ranked tests and spearman correlation coefficients, only the 2x20m-WWB test and the PILE-C (women only) were deemed to be sensitive to change. [40] Additionally, moderate to large effect sizes were found for all test components.

DISCUSSION

This study synthesized 12 studies assessing clinometric properties of 4 different functional-based assessments. Given the limited number of studies, the substantial variation in the types of tests examined, the methods used to assess the clinical measurement properties, and the study

populations, the current state of knowledge does not allow firm conclusions regarding recommendations for an optimal functional-based test at this time. Overall, the quality ranging from good to excellent (67-92%,) as determined by the QACMRR, for a range of properties of the 4 different assessments in patients with acute or chronic neck pain that is musculoskeletal in origin. Studies obtaining higher percentages indicate research that has been consistent with best practice where studies with lower percentages are more likely to be inadequate or inappropriate

FCE

The breadth of a functional-based test is variable and defined by the developers. An advantage of the functional assessment designed by Reesink et al.[34] is that they mapped the eight subtests to risk factors identified in the literature for work-related neck disorders. The eight subtests consist of: material handling tasks, lifting floor to waist, overhead lift test, one-handed and two-handed carrying, overhead working, repetitive reaching, overhead lifting, and repetitive bending and overhead reaching. Given the systematic approach and rationale these authors used in developing the FCE and this approach being used in previous research [41], we suggest that this test has strong content validity.

Six articles address the clinical measurement properties of this FCE ranging from good to excellent quality (67-92%). There was evidence that the FCE was stable over test-retest time of 7-14 days. [31,32] These measures demonstrate longer stability over time compared to self-report measures such as the Neck Disability Index (NDI) which has demonstrated test-retest reliability within only a short period of 0-3 days. [28] Whether this longer-term stability is a characteristic of functional-based tests or reflects differences in study populations in context requires further testing. These two studies had relatively lower quality scores on the QACMRR (67-75%) compared to other studies in this review putting into question test-retest time. Although test-retest

reliability has been assessed, inter-rater and intra-rater reliability has yet to be researched. Unlike self-report measures, we expect measurement error due to the evaluator and functional-based tests. Thus, future research should explore these aspects of reliability.

Convergent validity is often examined in clinical measurement studies. We suggest that this may be because these comparisons are easily performed by correlating different tests rather than providing strong confidence in the validity of the measurement. Often convenient comparisons are performed rather than those most relevant. Across many domains and measures it has become clear that the relationship between self-reported function and performance-based function or physical impairment is often very weak to moderate. Therefore, the value of assessment of these relationships as a form of validation has limited value. Several studies of very good to excellent quality have reported on the convergent validity of the FCE. [29,30,33] The highest quality article determined by the QACMRR (92%) found the relationship between the FCE and work capacity to be poorly associated with one another. [30] The same study found that the ability of the FCE to predict future work capacity was poor. This may be considered a more important comparison since ideally functional-based tests would relate to important outcomes like return to work. No studies to our knowledge report the responsiveness or sensitivity to change of the FCE. This is an important gap since the focus of rehabilitation is often to remediate limitations in goal impairments or work capacity, and assessment of these changes is critical to clinical decision-making and reporting outcomes. Thus, future research should evaluate the responsiveness of the FCE to provide insight in the measure's ability to detect change after an intervention.

FIT-HaNSA

One study of very good quality (88%) assessed the FIT-HaNSA, a test consisting of two reaching tasks (waist and eye-level) and sustained overhead task performance. [36] Overall, the

FIT-HaNSA demonstrated excellent inter-rater reliability (0.84) and intra-rater reliability (0.78). The specific subtests included within the FIT-HaNSA similarly demonstrate fair to excellent (0.54-0.80) and good (0.70-0.72) inter-rater and intra-rater reliability respectively. The FIT-HaNSA also demonstrated a clear ability to distinguish between people with WAD 2 and healthy controls. Correlations between the FIT-HaNSA and other patient self-report disability and functional outcome measures (NPRS, NDI, DASH, CROM and FIT-HaNSA) were generally very weak to weak ($\rho < 0.4$), consistent with other studies comparing performance and self-report. [17,18] The largest limitation in critically synthesizing information for this test is that only a single study was found that reported the measurement properties for people with neck disorders. It should be noted however that it has been validated in other MSK disorders. [35,41] Although others have noted the lag in development of functional-based measures in comparison to self-report measures, FIT-HaNSA was recommended as a functional-based measure for people with shoulder disorders. [42] Further research is necessary to investigate the responsiveness of the FIT-HaNSA.

BTEWS II

Another study of very good quality (88%) assessed the efficacy of the BTEWS II where the participants performed a dynamic pushing and pulling task in which power output was recorded over a 10 second sample.[35] While the convergent validity aspect of this paper was assessed as consistent with best practice through the critical appraisal process, the relationship between the power output on the BTEWS and measures of pain and disability (NDI, SPADI, NRS) were poorly associated with each other. In addition, the power output component was not found to be significantly different between people with neck pain and healthy controls which suggests it might not be discriminative. Discrimination between patients and healthy controls is a low standard for an outcome measure, and tests that cannot fulfil this benchmark should be viewed with caution.

Because of the weak measurement properties demonstrated by the power output component of the BTEWS II, it does not appear to be a desirable functional-based measure to assess function in people with neck pain. However, we acknowledge for all of the functional-based tests the evidence pool is so shallow that there is high potential that future studies might lead to different conclusions. Future research should also investigate the reliability and responsiveness of the BTEWS II.

Physiotherapy Test Package Subtests

Four studies ranging from good to very good quality (68-82%) assessed relevant items from a physiotherapy test package, including a lift from floor-to-waist and a waist-to-shoulder task and a two-handed carrying task. The properties of these assessment items include weak to moderate correlations to pain, perceived exertion, and had "fair to good" reliability. The 2x20m-WWB and PILE-C tests were found to be sensitive-to-change which is valuable information as no other study has assessed this property in functional-based measures in patients with neck disorders. Thus, this measure may be of value in clinical settings when assessing functional capacity before and after a treatment intervention. All tests had discriminative ability for detecting participants with spinal pain vs healthy controls. Most of the three tests demonstrated poor construct validity in that they were poorly related to pain and perceived exertion and the results were not in accordance with pre-defined hypotheses. Thus, further research is necessary to investigate these constructs. Three of the four results from the studies assessing the physiotherapy test package had a mixed sample of patients with various pain sites including back pain. While the majority of each cohort in these studies had neck pain, careful consideration should be taken to apply these tests to a neck pain specific population.

Clinical Implications

This study confirms that functional-based tests have had far less development and evaluation than self-report measures. Limitations include the number of tests and insufficient body of evidence to make confident recommendations with respect to functional-based testing. It is clear that self-report and functional-based measures provide different perspectives. Theoretically, functional-based tests are important to inform our understanding about the mechanisms of intervention and how interventions increase capacity. Future research may benefit by also comparing results from a functional-based measure to work capacity to when assessing construct validity. Overall more work is required to further establish the psychometric properties of functional-based tests in persons with neck disorders, including sensitivity-to-change, responsiveness, and predictive validity.

The FCE evaluated patients with neck pain of varying origin including WAD, work-related neck disorders, and chronic idiopathic neck pain. The BTEWs II evaluated functional capacity in patients with chronic neck pain, the FIT-HaNSA evaluated patients with WAD, and the physiotherapy test package did not specify the origin of musculoskeletal neck pain in their cohort. Thus, specific functional-based measures may be more applicable depending on the origin of the musculoskeletal neck pain being assessed.

The data presented suggest that the FIT-HaNSA has the strongest clinometric properties though this is based on a single higher quality paper specific to neck disorder. [36] Importantly, normative data have been published [43], it has been validated in multiple studies in patients with shoulder conditions [44–46] and has been recommended when compared to other measures [42]. The FCE has a limited evidence base from which to draw, though it was developed with strong content validity and further evaluation may demonstrate its usefulness.

Limitations

A challenge in synthesizing clinical measurement evidence is the wide range of properties and indicators that need to be considered. Unlike effectiveness studies where one can focus on the effect size of treatment there are many considerations that would affect the recommendations made about outcome measures. This is further complicated when the pool of evidence is shallow. Although the quality assessment tool (QACMRR) developed by one of the authors of this review which assess the quality of design of individual studies were useful for interpreting the evidentiary pool, there is no clear method to synthesize the extracted clinical measurement evidence. While some systematic reviews on treatment might only report findings from high-quality studies, it is important to see how outcome measures perform in different contexts. Further, the assessment of quality is complicated given that clinical measurement studies have so many dimensions. Therefore, exclusion of lower quality studies has questionable value. Thus, a more practical approach is to consider quality when interpreting the findings, rather than excluding studies.

The QACMRR focuses on whether the authors made appropriate decisions in selecting the scope and methods of their clinical measurement evaluations within a given study and provides descriptors of poor fair or good design options. Quality focuses on issues that might affect risk of bias or imprecision in estimates; whereas risk of bias assessments focusses on items that might result in a biased estimate. For example, insufficient power is a precision (quality) issue, not a risk of bias. Although it is difficult to interpret the meaning of the percentage of the QACMRR as there are no established cut-offs for distinguishing good and poor-quality studies, it provides one way of ranking the articles in order of quality. We did not use COSMIN checklist since it was developed for PROMS and some of the components/steps that involved are not applicable to performance-based tests.

Another limitation in this review was that the feasibility or usability of these tools was not assessed. While feasibility was not the focus of this review, information on the practical application of these functional-based measures provides valuable information to clinicians for determining whether these tests are appropriate to use in their given setting. Thus, future research should not only investigate further the psychometric properties of these tools, but also report the feasibility of using these tests so that they may be used in clinical settings and to identify limitations that restrict their application in practice.

CONCLUSION

This review found very good quality evidence that the FIT-HaNSA has excellent inter and intra-rater reliability and very weak to weak convergent validity. Excellent quality evidence of fair test-retest reliability, weak convergent validity, and very weak known groups validity for the BTEWS II test was found. Good to excellent quality evidence exists that an FCE battery has poor to excellent reliability and very weak to strong validity. Good to excellent quality of weak to strong validity and trivial to strong effect sizes were found for a physiotherapy test package. Functional-based evaluation in people with neck disorders is an area needing much research attention both to establish the measurement properties of existing measures, potentially to develop innovative new measures and to perform head-to-head comparisons of measures before an optimal functional-based test can be identified.

Authors' contributions

SM contributed significantly to conception and design of the study, data extraction, critical appraisal, interpretation of data and drafting of the manuscript. TS, TA, PB, and CC were involved in literature search, critical appraisal and interpretation of data and drafting. AG was involved in

critical appraisal and drafting. JM was also involved in the conception and design of the study, drafting, and revised the manuscript for important intellectual content. PB and CATWAD were involved in the drafting and review of the manuscript. All authors have given their final approval on the manuscript to be published

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study

Funding Statement

This work was supported by the Canadian Institutes of Health Research (CIHR) with funding reference number (FRN: SCA-145102).

Competing Interest Statement

None to report.

References

1 Carroll LJ, Hogg-Johnson S, van der Velde G, *et al.* Course and Prognostic Factors for Neck Pain in the General Population. Results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. *J Manipulative Physiol Ther* Published Online First: 2009. doi:10.1016/j.jmpt.2008.11.013

2 Croft PR, Lewis M, Papageorgiou AC, *et al.* Risk factors for neck pain: A longitudinal study in the general population. *Pain* Published Online First: 2001. doi:10.1016/S0304-3959(01)00334-7

3 Vos T, Allen C, Arora M, *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet* Published Online First: 2016. doi:10.1016/S0140-6736(16)31678-6

4 Blanpied PR, Gross AR, Elliott JM, *et al.* Neck Pain: Revision 2017. *J Orthop Sport Phys Ther* Published Online First: 2017. doi:10.2519/jospt.2017.0302

5 Nygren A, Berglund A, von Koch M. Neck-and-shoulder pain, an increasing problem. Strategies for using insurance material to follow trends. *Scand J Rehabil Med Suppl* 1995;**32**:107–12.http://www.ncbi.nlm.nih.gov/pubmed/7784832 (accessed 19 Jul 2018).

6 Wright A, Mayer TG, Gatchel RJ. Outcomes of disabling cervical spine disorders in compensation injuries: A prospective comparison to tertiary rehabilitation response for chronic lumbar spinal disorders. *Spine (Phila Pa 1976)* Published Online First: 1999. doi:10.1097/00007632-199901150-00020

7 Rempel DM, Harrison RJ, Barnhart S. Work-Related Cumulative Trauma Disorders of the Upper Extremity. *JAMA J Am Med Assoc* Published Online First: 1992. doi:10.1001/jama.1992.03480060084035

8 Borghouts JAJ, Koes BW, Vondeling H, *et al.* Cost-of-illness of neck pain in The Netherlands in 1996. *Pain* Published Online First: 1999. doi:10.1016/S0304-3959(98)00268-1

9 Hogg-Johnson S, van der Velde G, Carroll LJ, *et al.* The Burden and Determinants of

Neck Pain in the General Population. Results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. *J Manipulative Physiol Ther* Published Online First: 2009. doi:10.1016/j.jmpt.2008.11.010

10 Bobos P, Nazari G, Palimeris S, *et al.* The contribution of health and psychological factors in patients with chronic neck pain and disability: A cross-sectional study. *J Clin Diagnostic Res* Published Online First: 2018. doi:10.7860/JCDR/2018/31284.11203

11 Bobos P, Billis E, Papanikolaou D-T, *et al.* Does Deep Cervical Flexor Muscle Training Affect Pain Pressure Thresholds of Myofascial Trigger Points in Patients with Chronic Neck Pain? A Prospective Randomized Controlled Trial. *Rehabil Res Pract* Published Online First: 2016. doi:10.1155/2016/6480826

12 Nazari G, Bobos P, Billis E, *et al.* Cervical flexor muscle training reduces pain, anxiety, and depression levels in patients with chronic neck pain by a clinically important amount: A prospective cohort study. *Physiother Res Int* 2018;**23**. doi:10.1002/pri.1712

13 Bobos P, MacDermid JC, Walton DM, *et al.* Patient-Reported Outcome Measures Used for Neck Disorders: An Overview of Systematic Reviews. *J Orthop Sport Phys Ther* 2018;**48**:1–76. doi:10.2519/jospt.2018.8131

14 Macdermid JC, Walton DM, Bobos P, *et al.* The Open Orthopaedics Journal A Qualitative Description of Chronic Neck Pain has Implications for Outcome Assessment and Classification. *Open Orthop J* 2016;**10**:746–56. doi:10.2174/1874325001610010746

15 Childs JD, Cleland JA, Elliott JM, *et al.* Neck pain: Clinical practice guidelines linked to the international classification of functioning, disability, and health from the orthopaedic section of the american physical therapy association. J. Orthop. Sports Phys. Ther. 2008. doi:10.2519/jospt.2008.0303

16 Kay TM, Huijbregts M. Physical Rehabilitation Outcome Measures: A Guide to Enhanced Clinical Decision Making, Second Edition. *Physiother Canada* Published Online First: 2003. doi:10.2310/6640.2003.35271

17 Simmonds MJ, Olson SL, Jones S, *et al.* Psychometric characteristics and clinical usefulness of physical performance tests in patients with low back pain. *Spine (Phila Pa 1976)* Published Online First: 1998. doi:10.1097/00007632-199811150-00011

18 Stratford PW, Kennedy D, Pagura SMC, *et al.* The relationship between self-report and performance-related measures: questioning the content validity of timed tests. *Arthritis Rheum* 2003;**49**:535–40. doi:10.1002/art.11196

19 Novy DM, Simmonds MJ, Lee CE. Physical performance tasks: what are the underlying constructs? *Arch Phys Med Rehabil* 2002;**83**:44–7.http://www.ncbi.nlm.nih.gov/pubmed/11782832 (accessed 19 Jul 2018).

20 MacDermid JC, Stratford P. Applying evidence on outcome measures to hand therapy practice. *J Hand Ther* Published Online First: 2004. doi:10.1197/j.jht.2004.02.005

21 Alreni ASE, Harrop D, Lowe A, *et al.* Measures of upper limb function for people with neck pain. A systematic review of measurement and practical properties. *Musculoskelet Sci Pract* 2017;**29**:155–63. doi:10.1016/j.msksp.2017.02.004

22 Moher D, Shamseer L, Clarke M, *et al.* Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* Published Online First: 2015. doi:10.1186/2046-4053-4-1

23 Law MC, MacDermid J. *Evidence-based rehabilitation : a guide to practice*. Thorofare, NJ: : Slack Incorporated 2014.

24 Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation

Coefficients for Reliability Research. *J Chiropr Med* 2016;**15**:155–63. doi:10.1016/j.jcm.2016.02.012

25 Roy JS, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for musculoskeletal disorders: A systematic review. *J Rehabil Med* Published Online First: 2011. doi:10.2340/16501977-0643

26 Divaris K, Vann WF, Baker AD, *et al.* Examining the accuracy of caregivers' assessments of young children's oral health status. *J Am Dent Assoc* 2012;**143**:1237–47. doi:10.14219/jada.archive.2012.0071

27 Cohen J. Statistical power analysis for the behavioral sciences. Stat. Power Anal. Behav. Sci. 1988. doi:10.1234/12345678

28 MacDermid JC, Walton DM, Avery S, *et al.* Measurement properties of the neck disability index: a systematic review. *J Orthop Sports Phys Ther* 2009;**39**:400–17. doi:10.2519/jospt.2009.2930

29 Trippolini MA, Dijkstra PU, Geertzen JHB, *et al.* Construct Validity of Functional Capacity Evaluation in Patients with Whiplash-Associated Disorders. *J Occup Rehabil* 2015;**25**:481–92. doi:10.1007/s10926-014-9555-0

30 Trippolini MA, Dijkstra PU, Côté P, *et al.* Can functional capacity tests predict future work capacity in patients with whiplash-associated disorders? *Arch Phys Med Rehabil* 2014;**95**:2357–66. doi:10.1016/j.apmr.2014.07.406

31 Trippolini MA, Reneman MF, Jansen B, *et al.* Reliability and safety of functional capacity evaluation in patients with whiplash associated disorders. *J Occup Rehabil* 2013;**23**:381–90. doi:10.1007/s10926-012-9403-z

32 Reneman MF, Roelofs M, Schiphorst Preuper HR. Reliability and Agreement of Neck

Functional Capacity Evaluation Tests in Patients With Chronic Multifactorial Neck Pain. *Arch Phys Med Rehabil* 2017;**98**:1476–9. doi:10.1016/j.apmr.2016.12.005

33 van der Meer S, Reneman MF, Verhoeven J, *et al.* Relationship between self-reported disability and functional capacity in patients with whiplash associated disorder. *J Occup Rehabil* 2014;**24**:419–24. doi:10.1007/s10926-013-9473-6

34 Reesink DD, Jorritsma W, Reneman MF. Basis for a functional capacity evaluation methodology for patients with work-related neck disorders. *J Occup Rehabil* 2007;**17**:436–49. doi:10.1007/s10926-007-9086-z

35 Lomond K V, Côté JN. Shoulder functional assessments in persons with chronic neck/shoulder pain and healthy subjects: Reliability and effects of movement repetition. *Work* 2011;**38**:169–80. doi:10.3233/WOR-2011-1119

36 Pierrynowski M, McPhee C, P Mehta S, *et al.* Intra and Inter-Rater Reliability and Convergent Validity of FIT-HaNSA in Individuals with Grade Π Whiplash Associated Disorder. *Open Orthop J* 2016;**10**:179–89. doi:10.2174/1874325001610010179

37 Ljungquist T, Jensen IB, Nygren A, *et al.* Physical performance tests for people with long-term spinal pain: aspects of construct validity. *J Rehabil Med* 2003;**35**:69–75.http://www.ncbi.nlm.nih.gov/pubmed/12691336 (accessed 11 Jul 2018).

38 Ljungquist T, Harms-Ringdahl K, Nygren A, *et al.* Intra- and inter-rater reliability of an 11-test package for assessing dysfunction due to back or neck pain. *Physiother Res Int* 1999;**4**:214–32.http://www.ncbi.nlm.nih.gov/pubmed/10581627 (accessed 11 Jul 2018).

39 Ljungquist T, Fransson B, Harms-Ringdahl K, *et al.* A physiotherapy test package for assessing back and neck dysfunction--discriminative ability for patients versus healthy control subjects. *Physiother Res Int* Published Online First: 1999. doi:10.1002/pri.158

40 Ljungquist T, Nygren Å, Jensen I, *et al.* Physical performance tests for people with spinal pain - Sensitivity to change. *Disabil Rehabil* Published Online First: 2003. doi:10.1080/0963828031000090579

41 Reneman MF, Dijkstra PU, Westmaas M, *et al.* Test-retest reliability of lifting and carrying in a 2-day functional capacity evaluation. *J Occup Rehabil* 2002;**12**:269–75.http://www.ncbi.nlm.nih.gov/pubmed/12389478 (accessed 19 Jul 2018).

42 Hegedus EJ, Vidt ME, Tarara DT. The best combination of physical performance and self-report measures to capture function in three patient groups. *Phys Ther Rev* 2014;**19**:196–203. doi:10.1179/1743288X13Y.0000000121

43 Roy J-S, Macdermid JC, Boyd KU, *et al.* Rotational strength, range of motion, and function in people with unaffected shoulders from various stages of life. *Sports Med Arthrosc Rehabil Ther Technol* 2009;**1**:4. doi:10.1186/1758-2555-1-4

44 Kumta P, MacDermid JC, Mehta SP, *et al.* The FIT-HaNSA Demonstrates Reliability and Convergent Validity of Functional Performance in Patients With Shoulder Disorders. *J Orthop Sport Phys Ther* 2012;**42**:455–64. doi:10.2519/jospt.2012.3796

45 Macdermid JCJC, Ghobrial M, Badra Quirion K, *et al.* Validation of a new test that assesses functional performance of the upper extremity and neck (FIT-HaNSA) in patients with shoulder pathology. *BMC Musculoskelet Disord* 2007;**8**:42. doi:10.1186/1471-2474-8-42

46 Hawkes DH, Alizadehkhaiyat O, Fisher AC, *et al.* Normal shoulder muscular activation and co-ordination during a shoulder elevation task based on activities of daily living: An electromyographic study. *J Orthop Res* 2012;**30**:53–60. doi:10.1002/jor.21482

TABLE 1. Summary of Studies Reporting Psychometric Properties of Functional-based Tests in Neck Disorder Patients

Study	Population	Sample Size (n)	Functional Tests	Intervention/Test Interval	Quality
Ljungquist et al. 1999	Neck pain (55%), back pain, multiple pain sites,	53	PILE-C, PILE-L	N/A	Good (68%)
Ljungquist et al. 1999	Neck pain (50%), lumbar pain, thoracic pain, shoulder pain, multiple pain sites,	68	PILE-C, PILE-L, 2 x 20m WWB	8 days	Very Good (79%)
Ljungquist et al. 2003	Neck pain, lumbar pain, thoracic pain, shoulder pain, lower extremity pain, multiple pain sites,	235	PILE-C, PILE-L, 2 x 20m WWB	N/A	Very Good (82%)
Ljungquist et al. 2003	cervical pain (25%), lumbar pain, cervical (25%) and lumbar pain, multiple pain sites,	186	PILE-C, PILE-L, 2 x 20m WWB	6 months	Very Good (79%)
Lomond and Cote. 2011	Chronic neck and shoulder pain (100%)	32	BTEWS II	9.5 days	Very Good (88%)
Pierrynowski et al. 2016	Sub-acute and chronic WAD II	66	FIT-HaNSA	2-7 days	Very Good (88%)
Reesink et al. 2007	N/A	N/A	Neck-FCE	N/A	N/A
Reneman et al. 2017	Chronic multifactorial neck pain	18	Neck-FCE	2 weeks	Good (67%)
Trippolini et al. 2013	Sub acute and chronic WAD I and II	32	WAD FCE	7 days	Very Good (75%)
Trippolini et al. 2014	Sub acute and chronic WAD I and II	267	Workwell FCE	N/A	Excellent (92%)

Trippolini et al. 2015	Sub acute and chronic WAD I and II	314	WAD FCE	N/A	Very Good (86%)
Van der Meer et al. 2013	Chronic WAD I and II	40	Neck FCE	N/A	Very Good (86%)

PILE-C, Progressive Isoinertial Lifting Evaluation-Cervical; PILE-L, Progressive Isoinertial Lifting Evaluation; CBT, Cognitive-Behavioural Therapy; PT, Physical Therapy; NRPS, Numeric Pain Rating Scale; BTEWS II, Baltimore Therapeutic Equipment Work Simulator II; WAD, Whiplash Associated Disorder; MVA, Motor Vehicle Accident; FIT-HaNSA, Functional Impairment Test-Hand and Neck/Shoulder/Arm; FCE, Functional Capacity Evaluation; EXP, Experimental; M, Male; F, Female; N/A, not applicable

TABLE 2. Quality of Studies on Psychometric Properties of Functional-based Tests Evaluated in Neck Disorder Patients

Study	Item Evaluation Criteria												
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Total (%)
Trippolini et al, 2014	2	2	2	2	1	2	2	2	2	1	1	2	92%
Lomond and Cote, 2011	2	2	1	2	0	2	2	2	2	2	2	2	88%
Pierrynowski et al, 2016	2	2	1	2	0	2	2	2	2	2	2	2	88%
Trippolini et al, 2015	2	2	2	0	1	N/A	2	2	2	2	2	2	86%
Van der Meer et al, 2013	2	1	2	1	2	N/A	2	1	2	2	1	2	86%
Ljungquist et al 2003 KGV**	2	2	2	0	0	N/A	2	2	2	2	2	2	82%
Ljungquist et al 1999 Rel****	2	1	1	2	0	2	2	2	2	2	1	2	79%
Ljungquist et al 2003 STC***	1	1	1	2	1	1	2	2	2	2	2	2	79%
Trippolini et al, 2013	2	2	1	1	0	0	2	2	2	2	2	2	75%
Ljungquist et al 1999 KGV**	2	1	1	2	0	N/A	2	1	2	2	1	2	68%
Reneman et al, 2017	1	2	1	1	1	0	1	2	2	2	2	1	67%
Reesink, 2007*	-	-	-	-	-	-	-	-	-	-	-	-	N/A

12-item evaluation tool (QACMRR) designed to assess the quality of studies determining measurement properties in outcome measures. Questions 1-12 in the tool evaluate aspects of study question, study design, measurements, analyses, and study recommendations.

KGV, known-groups validity; rel, reliability; STC, sensitivity-to-change

*Paper is not applicable for completion of study quality tool

TABLE 3. Psychometric Properties of the Functional Capacity Evaluation

FCE Battery	Type of Properties	Statistical Test	Value	Interpretation
Neck FCE	Test-retest	ICC	0.39-0.96	Poor-excellent
	Measurement Error	Ratio of LoA	32.0-56.5%	
	Convergent Validity	Pearson or Spearman correlation	NDI total: 0.39-0.62 NDI items: 0.03-0.63	Weak to moderate very weak to strong
WAD FCE	Test-retest Reliability	ICC	0.66-0.96	moderate-excellent
	Convergent Validity	Pearson Correlation	Pain* 0.31-0.39 SFS: 0.42-0.61 NDI: 0.34-0.45 HADS-A: 0.27-0.36 HADS-D: 0.30-0.41	Weak Moderate-strong Weak-moderate weak Weak-moderate
	Discriminative Validity (German vs Non-German)	Linear Regression Analysis	p<0.001	Significant for All Tasks
	Discriminative Validity (sex)	t-test	p<0.001	Significant for Two Tasks
Workwell FCE	Convergent Validity	Pearson or Spearman Correlation	Work Capacity: 0.1-0.3	Very Weak – weak
	Predictive Validity	Pearson or Spearman Correlation	0.06-0.39	Very weak - Weak
		Linear Mixed Model Regression of All Predictors	β=-0.04, 95% CI: -0.15 – 0.06 p=0.428 (task 6)	Not Significant

FCE, Functional Capacity Evaluation; ICC, Intraclass correlation coefficient; LoA, Limits of Agreement; NDI, Neck Disability Index; Mod., Moderate; Neg., Negligible; SFS, Spinal Function Sort; HADS-A, Hospital Anxiety and Depression Scale – Anxiety; HADS-D, Hospital Anxiety and Depression Scale – Depression; CI, Confidence Interval Sig., Significant
*Pain measured via Numeric Rating Scale

TABLE 4. Summary of Fit-HaNSA's psychometric properties in neck disorder patients

Test	Type of Property	Statistical Test	Value	Interpretation
Fit-HaNSA	Intra-rater Reliability	ICC	0.78	good
Fit-HaNSA	Inter-rater Reliability	ICC	0.84	good
Fit-HaNSA	Measurement Error	SEM LOA_{95} MDC_{90}	76 s 248 s 176 s	
Fit-HaNSA	Convergent Validity	Spearman Rank Correlation	<0.4 - >0.75	Weak – Strong
Fit-HaNSA	Discriminative WAD II vs Control	F-test	62.6, <p,0.001	Significant
Fit-HaNSA Functional Sub-tasks	Intra-rater reliability	ICC	0.70-0.72	moderate
	Inter-reliability	ICC	0.54-0.80	moderate – good
	Convergent Validity	Spearman Rank Correlation	<0.4 - >0.75	Weak - Strong
	Discriminative Validity WAD II vs Control	F-test	42.0-53.3, p<0.001	Significant

Fit-HaNSA, Functional Impairment Test, Hand and Neck/Shoulder/Arm; ICC, Intraclass correlation coefficient; SEM, Standard Error of Measurement; LOA_{95}, 95% Limits of Agreement; MDC_{90}, 90% Minimal Detectable Change; WAD, Whiplash Associated Disorder; Mod, Moderate

*Correlations completed with Numeric Pain Rating Scale, Neck Disability Index, Disabilities of Arm, Shoulder, Hand and 6 cervical range of motion tests

TABLE 5. Psychometric Properties of Baltimore Therapeutic Equipment Work Simulator II – Power Output Task

Test	Type of Property	Statistical Test	Value	Interpretation
BTEWS II	Test-retest reliability	ICC	0.53	moderate
		Spearman	0.37	Poor
BTEWS II	Measurement Error	SEM	30.25	
		MDC_{90}	70.59	
BTEWS II	Convergent Validity*	Spearman	Not Reported	Weak
BTEWS II	Discriminative Validity (Pain vs Control)	Two-way Repeated Measures ANOVA	Not Reported	Non-significant

ICC, Intraclass correlation coefficient; SEM, Standard Error of Measurement; MDC_{90}, 90% Minimal Detectable Change; ANOVA, Analysis of Variance

*Spearman correlations completed with Numeric Rating Scale, Neck Disability Index and Shoulder Pain and Disability Index

TABLE 6. Psychometric Properties of performance-based tests included in physiotherapy test package

Test	Type of Property	Statistical Test	Value	Interpretation
PILE-C	Inter-rater Reliability	Mean Difference LoA	-0.24 -2.46 and 1.82	
PILE-C	Inter-rater Reliability	Repeatability (2X SD) % of Range	M=3.93; F=1.19 M=10.5%; F=6.1%	
PILE-C	Convergent Validity	Spearman Correlation	CR-10: 0.55-0.65* Borg RPE: 0.10 - 0.48	Moderate - Strong Very weak - moderate
PILE-C	Discriminative: spinal pain vs. control	Sensitivity and Specificity	0.93, 0.69	Strong – Very Strong
PILE-C	Discriminative: spinal pain vs. control	Wilcoxon Sign Ranked Test	p=0.008	Significant
PILE-C	Discriminative: High vs. low pain intensity	Mann-Whitney U	p=0.003	Significant
PILE-C	Discriminative: High vs. low Pain behavior	Mann-Whitney U	p=0.005	Significant
PILE-C	Discriminative: High vs. low perceived exertion	Mann-Whitney U	p=0.154	Non-significant
PILE-C	Sensitivity to Change	Effect Size	Subjects improving: 0.39 - 0.73 Subjects deteriorating: 0 – 0.4	Small – Moderate Trivial – Small
PILE-L	Inter-rater Reliability	Mean Difference LoA	-0.11 -2.33 and 2.11	
PILE-L	Intra-rater Reliability	Repeatability % of Range	M=4.0; F=3.59 M=10.7%; F=18.5%	
PILE-L	Convergent Validity	Spearman Correlation	CR-10: 0.11 – 0.45 Borg RPE: 0.10 - 0.48	very weak – moderate very weak – moderate
PILE-L	Discriminative: spinal pain vs no spinal pain	Sensitivity and Specificity	0.85, 0.65	Strong – Very Strong

PILE-L	Discriminative: spinal pain vs control	Wilcoxon Sign Ranked Test	p=0.002	Significant
PILE-L	Discriminative: High vs. low pain intensity	Mann-Whitney U	p=0.001	Significant
PILE-L	Discriminative: High vs. low pain behaviour	Mann-Whitney U	p<0.001	Significant
PILE-L	Discriminative: High vs. low perceived exertion	Mann-Whitney U	p<0.001	Significant
PILE-L	Sensitivity to change	Effect Size	Subjects improving: 0.02 – 1.08 Subjects deteriorating 0.42-0.81	Trivial – Large Small – Large
2 x 20m WWB	Inter-rater Reliability	Mean Difference LoA	0.05 -1.33 and 1.43	
2 x 20m WWB	Intra-rater Reliability	Repeatability % of Range	3.2 10.7%	
2 x 20m WWB	Convergent Validity	Spearman Correlation	CR-10: 0.55 - 0.65Borg RPE: 0.10 - 0.48	Moderate - Strong very weak – moderate
2 x 20m WWB	Discriminative: spinal pain vs control	Wilcoxon Sign Ranked Test	p=0.014	Significant
2 x 20m WWB	Discriminative: High vs. low pain intensity	Mann Whitney U	p<0.001	Significant
2 x 20m WWB	Discriminative: High vs. low pain behaviour	Mann Whitney U	p<0.001	Significant
2 x 20m WWB	Discriminative: High vs. low perceived exertion	Mann Whitney U	p<0.001	Significant
2 x 20m WWB	Sensitivity to change	Effect Size	Subjects improving: 0.38-0.78 Subjects deteriorating: 0.13-0.62	Small – Moderate Trivial – Moderate

PILE-C, Progressive Iso-intertial Lifting Evaluation – Cervical; PILE-L, Progressive Iso-intertial Lifting Evaluation – Lumbar; LoA, Limits of Agreement; SD, Standard Deviation; M, Male; F, Female; RPE, Rating of perceived exertion; KGV, Known-groups Validity; Neg., Negligible; Mod., Moderate, *CR-10: Measurement of pain construct

Figure 1. Selection of the studies for inclusion in the systematic review

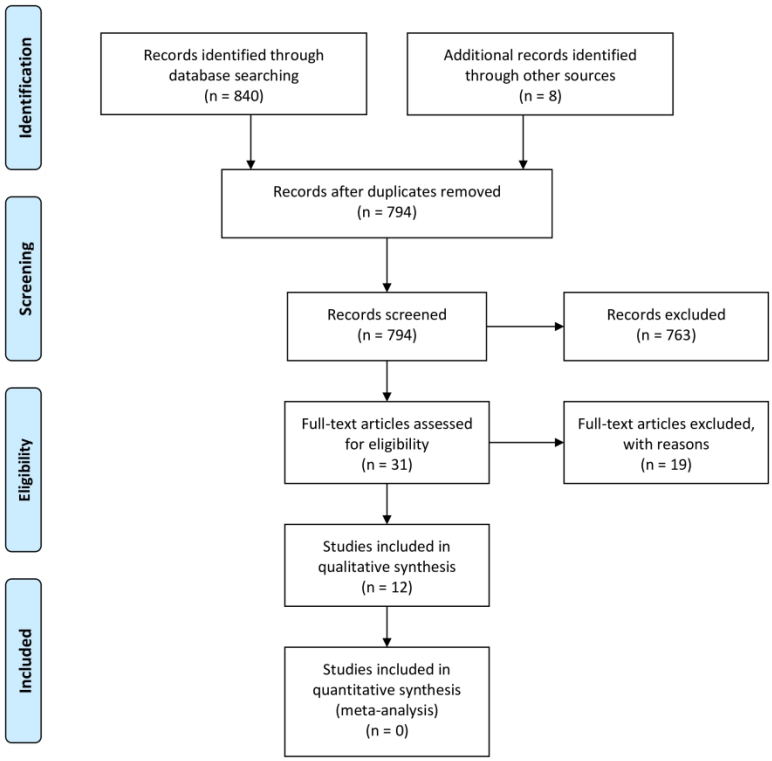

Figure 1

215x279mm (300 x 300 DPI)

Appendix 1: Search terms

EMBASE-OVID

1. exp "outcome and process assessment (health care)"/ or "outcome assessment (health care)"/ or treatment outcome/
2. outcome?.ti.
3. exp "Range of Motion, Articular"/
4. Pain Measurement/
5. exp disability evaluation/
6. "Recovery of Function"/
7. Questionnaires/
8. self-report.tw.
9. ((impairment or disability or function) adj2 (measure? or scale? or evaluation?)).tw.
10. range of motion.tw.
11. (strength adj2 (measure? or scale? or evaluation?)).tw.
12. (outcome? adj2 (measure* or scale? or indicator?)).tw.
13. or/1-12
14. "reproducibility of results"/
15. exp "Sensitivity and Specificity"/
16. reliability.mp.
17. validity.mp.
18. responsiveness.mp.
19. Psychometrics/
20. rasch.mp.
21. factor analysis, statistical/
22. factor analysis.tw.
23. differential functioning.mp.
24. (validity or validation).mp. [mp=title, original title, abstract, name of substance word, subject heading word, unique identifier]
25. (validity or validation).mp.
26. item difficulty.mp.
27. translation.tw.
28. or/14-27
29. 13 and 28
30. Neck Pain/
31. exp Brachial Plexus Neuropathies/
32. exp neck injuries/ or exp whiplash injuries/
33. cervical pain.mp.
34. neckache.mp.
35. whiplash.mp.
36. cervicodynia.mp.
37. cervicalgia.mp.
38. brachialgia.mp.
39. brachial neuritis.mp.
40. brachial neuralgia.mp.
41. neck pain.mp.

42. neck injur*.mp.
43. brachial plexus neuropath*.mp.
44. brachial plexus neuritis.mp.
45. thoracic outlet syndrome/ or cervical rib syndrome/
46. Torticollis/
47. exp brachial plexus neuropathies/ or exp brachial plexus neuritis/
48. cervico brachial neuralgia.ti,ab.
49. cervicobrachial neuralgia.ti,ab.
50. (monoradicul* or monoradicl*).tw.
51. or/30-50
52. exp headache/ and cervic*.tw.
53. exp genital diseases, female/
54. genital disease*.mp.
55. or/53-54
56. 52 not 55
57. 51 or 56
58. neck/
59. neck muscles/
60. exp cervical plexus/
61. exp cervical vertebrae/
62. atlanto-axial joint/
63. atlanto-occipital joint/
64. Cervical Atlas/
65. spinal nerve roots/
66. exp brachial plexus/
67. (odontoid* or cervical or occip* or atlant*).tw.
68. axis/ or odontoid process/
69. Thoracic Vertebrae/
70. cervical vertebrae.mp.
71. cervical plexus.mp.
72. cervical spine.mp.
73. (neck adj3 muscles).mp.
74. (brachial adj3 plexus).mp.
75. (thoracic adj3 vertebrae).mp.
76. neck.mp.
77. (thoracic adj3 spine).mp.
78. (thoracic adj3 outlet).mp.
79. trapezius.mp.
80. cervical.mp.
81. cervico*.mp.
82. 80 or 81
83. exp genital diseases, female/
84. genital disease*.mp.
85. exp *Uterus/
86. 83 or 84 or 85
87. 82 not 86

88. 58 or 59 or 60 or 61 or 62 or 63 or 64 or 65 or 66 or 67 or 68 or 69 or 70 or 71 or 72 or 73 or 74 or 75 or 76 or 77 or 78 or 79 or 87
89. exp pain/
90. exp injuries/
91. pain.mp.
92. ache.mp.
93. sore.mp.
94. stiff.mp.
95. discomfort.mp.
96. injur*.mp.
97. neuropath*.mp.
98. or/89-97
99. 88 and 98
100. Radiculopathy/
101. exp temporomandibular joint disorders/ or exp temporomandibular joint dysfunction syndrome/
102. myofascial pain syndromes/
103. exp "Sprains and Strains"/
104. exp Spinal Osteophytosis/
105. exp Neuritis/
106. Polyradiculopathy/
107. exp Arthritis/
108. Fibromyalgia/
109. spondylitis/ or discitis/
110. spondylosis/ or spondylolysis/ or spondylolisthesis/
111. radiculopathy.mp.
112. radiculitis.mp.
113. temporomandibular.mp.
114. myofascial pain syndrome*.mp.
115. thoracic outlet syndrome*.mp.
116. spinal osteophytosis.mp.
117. neuritis.mp.
118. spondylosis.mp.
119. spondylitis.mp.
120. spondylolisthesis.mp.
121. or/100-120
122. 88 and 121
123. exp neck/
124. exp cervical vertebrae/
125. Thoracic Vertebrae/
126. neck.mp.
127. (thoracic adj3 vertebrae).mp.
128. cervical.mp.
129. cervico*.mp.
130. 128 or 129
131. exp genital diseases, female/

132. genital disease*.mp.
133. exp *Uterus/
134. or/131-133
135. 130 not 134
136. (thoracic adj3 spine).mp.
137. cervical spine.mp.
138. 123 or 124 or 125 or 126 or 127 or 135 or 136 or 137
139. Intervertebral Disk/
140. (disc or discs).mp.
141. (disk or disks).mp.
142. 139 or 140 or 141
143. 138 and 142
144. herniat*.mp.
145. slipped.mp.
146. prolapse*.mp.
147. displace*.mp.
148. degenerat*.mp.
149. (bulge or bulged or bulging).mp.
150. 144 or 145 or 146 or 147 or 148 or 149
151. 143 and 150
152. intervertebral disk degeneration/ or intervertebral disk displacement/
153. intervertebral disk displacement.mp.
154. intervertebral disc displacement.mp.
155. intervertebral disk degeneration.mp.
156. intervertebral disc degeneration.mp.
157. 152 or 153 or 154 or 155 or 156
158. 138 and 157
159. 57 or 99 or 122 or 151 or 158
160. animals/ not (animals/ and humans/)
161. 159 not 160
162. exp *neoplasms/
163. exp *wounds, penetrating/
164. 162 or 163
165. 161 not 164
166. 29 and 165
167. guidelines as topic/
168. practice guidelines as topic/
169. guideline.pt.
170. practice guideline.pt.
171. (guideline? or guidance or recommendations).ti.
172. consensus.ti.
173. or/167-172
174. meta-analysis/
175. exp meta-analysis as topic/
176. (meta analy* or metaanaly* or met analy* or metanaly*).tw.
177. review literature as topic/

178. (collaborative research or collaborative review* or collaborative overview*).tw.
179. (integrative research or integrative review* or intergrative overview*).tw.
180. (quantitative adj3 (research or review* or overview*)).tw.
181. (research integration or research overview*).tw.
182. (systematic* adj3 (review* or overview*)).tw.
183. (methodologic* adj3 (review* or overview*)).tw.
184. exp technology assessment biomedical/
185. (hta or thas or technology assessment*).tw.
186. ((hand adj2 search*) or (manual* adj search*)).tw.
187. ((electronic adj database*) or (bibliographic* adj database*)).tw.
188. ((data adj2 abstract*) or (data adj2 extract*)).tw.
189. (analys* adj3 (pool or pooled or pooling)).tw.
190. mantel haenszel.tw.
191. (cohrane or pubmed or pub med or medline or embase or psycinfo or psyclit or psychinfo or psychlit or cinahl or science citation indes).ab.
192. or/174-191
193. 173 or 192
194. 166 and 193

APPENDICES

APPENDIX A. Data extraction guide for studies evaluating the quality of studies evaluating the clinical measurement properties of outcome measures

Instructions

Clinical measurement studies may evaluate a wide spectrum of measurement properties; or evaluate aspects that relate to the implementability or interpretation of outcome measures. Individual clinical measurement studies cannot address every aspect of the measurement properties of an instrument. Ideally systematic reviews will synthesize the quality and content of research evidence addressing the clinical measurement properties of individual outcome measures. The summative knowledge about the measurement properties, cultural transferability, and utility across different contexts provides the scope of information needed to select an outcome measure for a specific patient (population), purpose and context.

This guide should facilitate extraction of data from individual clinical measurement studies. An explanation of the measurement property addressed in each item and how it might be measured within a given study is listed to facilitate finding and extracting that information. The accompanying extraction form can then be used to collect the specific information on these measurements or utility properties from specific studies.

The purpose of data extraction is to extract the specific information reported by authors within a study, not to evaluate the validity or value of that piece of information. Evaluation of the quality of the published version of the clinical measurement study (also called critical appraisal) is performed in a separate step. See the accompanying critical appraisal tool and guide. It is advisable to extract detailed specific information from the study; recognizing that this information may later be synthesized or subject to meta-analysis.

There is no standardized process for synthesizing clinical measurement information. Based on the findings of extraction you may elect to present the synthesize data in a descriptive way by creating a summary table of the data extracted in each category. If you find some studies with similar designs, you may be able to conduct a meta-analysis of some properties like clinically important difference (CID) or minimal detectable change (MDC); if appropriate given the sample and technique - this can be valuable as it may provide more stable estimates of these important properties.

Population studied		
Population	A description of the study population	Sample size, pathology/disorder, demographics, setting, acute vs. chronic, where subjects were chosen from. Report meaningful demographics and indicators of the population studied.
Intervention	Interventions (if applicable) applied during longitudinal studies	Description of the nature, frequency, intensity of the intervention and the follow-up interval.
Reliability		
Reliability Description	The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: for example, using different sets of items from the same health-related instrument (internal consistency), over time (test retest) by different persons on the same occasion (interrater) or by the same persons (i.e., raters or responders) on different occasions (intra-rater)	Test procedures or measures are typically reapplied on repeated occasions in individuals considered to have a stable condition during that time frame which repeated testing occurs. Repeated testing may be performed on different occasions (test-retest) for self-report measures, OR by the same rater (intra-rater) or different raters (inter-rater) if it is an observer-based scale. In some cases different test instruments (inter-instrument) are evaluated. The most common statistic used is the intraclass correlation coefficient for quantitative data (Shrout & Fleiss, 1979) and kappa(Landis & Koch, 1977) for nominal data. Standard error of measurement is used to present a quantitative estimate of the reliability—in the original units of measure. Report the type of reliability evaluated and coefficients obtained.
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	This may be reported as 1. Standard error of measurement (in older articles you may see coefficient of variation); 2. Altman and Bland graphical technique (Bland & Altman, 1990; Bland & Altman, 1987; Bland & Altman, 1986) where the difference on repeated tests for each individual (limits of agreement) is plotted versus their

		mean score. The mean difference and the boundaries of 2SD are shown to define the limits of agreement.
Internal consistency	The extent to which items on a test or subscale are related (an indication of the consistency of the concept measured).	Cronbach's alpha is the inter-item correlation usually reported. Report alpha and whether it relates to the entire instrument or specific subscales.
Validity		
Content Validity	The degree to which the content of a health-related instrument is an adequate reflection of the construct to be measured	A variety of techniques can be used to assess the extent to which items on a given measure reflected the necessary content to capture the concept of interest. Some of the techniques you will find are listed. Extract what was done to determine content validity and what was found. 1) Patients and experts were involved during item selection/reduction - report how they were used and key decisions 2) Patients were consulted for reading and comprehension - report key findings 3) Cognitive interviews (Cibelli, 1994; Ojanen & Gogates, 2006) were done with patients to determine how items were interpreted by respondents; their perceptions of the items - report key findings 4) Expert panels or Delphi procedures were used to select items or evaluate the validity of the instrument - report key findings and decisions 5) During translation specific study, the meaning of the questions to another cultural or language group was studied - report key findings and decisions 6) ICF linking (Cieza et al., 2002) or other coding of content was performed - report the results which may include the distribution of content across ICF domains, or the distribution of specific codes
Construct Validity	The degree to which the scores of a health-related instrument are consistent with hypotheses (for instance with regard to internal	When extracting data about correlational validity, the pre-constructed hypothesis and whether it is supported should be documented. For correlational construct

	relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the health-related instrument validly measures the construct to be measured	validity, this will be the nature and strength of the prespecified relationship and the correlations that support that. Relation to other indices/constructs that are similar (convergent) or different (divergent) can be reported. Ideally, hypotheses are formulated/reported and supported by correlations that are in accordance with the hypotheses. Note that there is no consistent agreement on what subjective term should be applied to validity correlations. Note that there is no consistent agreement on what subjective term should be applied to validity correlations. Some authors use subjective terminology defined for reliability such as: strong (>0.70) and moderate (0.40-0.70) correlations; others use the correlations like effect size benchmarks that 0.4 indicates a moderate effect and 0.6 a large effect. For validity assessment is more important than correlations prespecified constructed hypotheses, although not all papers are written clearly with respect to this.
Structural Validity/Hypothesis Testing	The degree to which the scores of a health-related instrument are an adequate reflection of the dimensionality of the construct to be measured	Extract test names, prespecified expected relationship and correlations observed.
Structural validity - discriminative	discriminative analysis supports the validity of a measure by demonstrating that the measurement is able to differentiate between groups that are prespecified and known to be different on the construct being assessed.	Data extraction should include the nature of the subgroups and the size of the difference observed between them (and its statistical significance). Typically, statistical tests of difference are performed. Since known groups analysis can provide data that is useful in clinical practice as benchmarks for comparing these known groups, it is a more practical form of construct validity than correlational. Data extraction/presentation should reflect this by presenting the group central tendency, their margins and statistical significance in an accessible manner.

Criterion validity	Criterion validation is determined by comparing a given outcome measure to an accepted standard of measure. For subjective constructs like pain and disability, it can be argued that there is no criterion since there is no external gold standard. Therefore, for self-report measures, validation focuses on construct validity. For performance measures, it is common to have a criterion measure that is considered to be highly precise and rigorous as the criterion comparator.	Authors will state that their measure is being compared against a specific instrument and report the correlation or agreement between the measures. Extract the test names and results: correlations or other as reported.
Responsiveness/Clinical Change		
Responsiveness	The ability of a health-related instrument to detect change over time in the construct to be measured	Extract indicators of responsiveness include: effect size, standard response mean and the method for assessing whether patients were improved, stable or worse. (Beaton, 2000)
Interpretability		
Interpretability	The degree to which one can assign qualitative meaning that is, clinical or commonly understood connotations to an instrument's quantitative scores or change in scores.	

APPENDIX B. Data extraction form for studies evaluating the clinical measurement properties of outcome measures

Authors: _____ Year:_____ Rater: _____

Instructions

When using the data extraction form, it is important to realize that the purpose of data extraction is to remove or extract the specific information reported by authors within a study, not to evaluate the validity or value of that piece of information. To make data extraction as useful as possible, and to avoid the need for repeated data extractions, it is advisable to read the accompanying guide and then be as specific as possible when extracting information.

	DATA EXTRACTED
Population studied	
Population	
Intervention	
Reliability	
Reliability (relative)	
Reliability (absolute)	
Minimum Detectable Change	
Content/structural validity	
Internal consistency	
Content Validity	

Floor-Ceiling Effects	
Factorial validity	
Item response /Rasch Analyses	
Construct/Criterion Validity	
Known groups	
Convergent	
Divergent	
Longitudinal Validity	
Concurrent criterion	
Predictive criterion	
Responsiveness/Clinical Change	
Responsiveness	

Minimally Clinical Important Difference	

APPENDIX C. Quality Appraisal for Clinical Measurement Research Reports Evaluation Form

Rater (Group)_____
Author(s) (Study Author(s) _____
Year (Year of publication)_____

1. Was the relevant background work cited to define what is currently known about the measurement properties of measures under study, and the potential contributions of the current research question to informing that knowledge base?
   2
   1
   0
2. Were appropriate inclusion/exclusion criteria defined? *
   2
   1
   0
3. Were specific clinical measurement questions/hypotheses identified?
   2
   1
   0
4. Was an appropriate scope of measurement properties considered?
   2
   1
   0
5. Was an appropriate sample size used?
   2
   1
   0
6. Was appropriate retention/follow-up obtained? (for studies involving retesting; otherwise n/a)
   2
   1
   0

7. Were specific descriptions provided of the measure under study and the method(s) used to administer it?

2
1
0

8. Were standardized procedures used to administer all study measures in a manner that minimized potential sources of error/bias (including the study measure and its comparators)?

2
1
0

9. Were analyses conducted for each specific hypothesis or purpose?

2
1
0

10. Were appropriate statistical tests performed to obtain point estimates of the measurement properties?

2
1
0

11. Were appropriate ancillary analyses done to quantify the confidence in the estimates of the clinical measurement property (Precision/Confidence intervals; benchmark comparisons/ROC curves, alternate forms of analysis like SEM/MID, etc.)?

2
1
0

12. Were clear, specific and accurate conclusions made about the clinical measurement properties; that were associated with appropriate clinical measurement recommendations and supported by the study objectives, analysis and results?

2
1
0

Subtotals (of column 1 and 2) Total Score (sum of subtotals/24*100)

APPENDIX D. Description of each performance battery from selected articles

Battery	Description of Tasks
Relevant FCE Subtasks[25,26,27,28,29,30]	Material Handling Tasks: All lifting tests were executed with a wooden crate (40 × 30 × 26 cm) of 2.5 kg, and four to five weight increments of 2.5 kg or 5 kg each were used until the maximum amount of weight was reached. Maximum performance was recorded in kg. Lifting floor to waist: Measured after five lifts of crate from floor to table and vice versa (time limit < 90 s): hands remained on the crate during the test. Increase weight in 4-5 steps until maximum is reached Overhead lift test: Five lifts from waist to crown height and vice versa within 90 s in standing position. Increase weight in 4–5 steps until maximum is reached Two-handed carrying: Carrying of a crate for a short distance measured after five carries of 1.5 m distance at waist height. Hands remain on the crate during the test. One-handed carrying: Carrying wooden crate for 15 m within 90 s beginning with the right hand and thereafter the left hand. Overhead working: Standing with hands at crown height for manipulation of nuts and bolts. The time that the position was held is recorded (sec). Repetitive reaching: fast horizontal movements of the upper extremity in a sitting position. Marbles are removed from bowls at arm length distance at table height from left to right and vice versa, with right and then left arm. The time taken to remove 30 marbles is recorded (sec). Overhead lift test: Five lifts from waist to crown height and vice versa within 90 s in standing position. Increase weight in 4–5 steps until maximum is reached Repetitive bending and overhead reaching: 20 marbles in 2 bowls at table height and crown height. Standing in front of bowl of marbles and moving the marbles as fast as possible from table height to crown height.

A Physiotherapy Test Package[33,34,35,36]	**PILE Tests:** "The lifting tests were performed standing in front of bookshelves with shelves at 0.76m and 1.37 m from the floor. Subjects were asked to lift weights in a plastic box from floor to waist level (0–0.76 m) for the lumbar PILE test, or from waist to shoulder height (0.76–1.37 m) for the cervical PILE test. The initial weight was 3.6 kg for women and 5.9 kg for men. A 'lifting movement' involved a single transfer from one level to the next and back again. After every four such lifting movements (= 20 s), the weight was increased by 2.25 kg for women and 4.5 kg for men. The weight managed during the last lifting movement was recorded and used as a test result, as well as this maximum weight divided by the 'adjusted weight'". **2x20m WWB:** "Subjects were asked to walk 20 m at a comfortable speed along a corridor, to turn around where 20 m was marked and then to walk 20 m back to the starting point. In the first walking test they carried no extra weight, but in the second they carried one carrier bag in each hand, containing 4 kg each for the women, 8 kg each for the men. The time taken was recorded to get the walking speed. The tests were discontinued after 50 s".
BTEWS II[31]	"The protocol consisted of performing a series of shoulder functional tasks before and after a fatiguing activity. Functional tasks consisted of active shoulder range of motion (ROM) in both flexion and abduction and cumulative power output (PO) accumulated over 10s during a repetitive pushing/pulling task in a horizontal plane at shoulder level".
FIT - HaNSA[32]	"The FIT-HaNSA protocol consists of three timed tasks and each task is performed for a maximum of 300 seconds (s) with approximately 30 s pause between them (set-up time for next task). Task 1 (waist-up) requires the patient to alternately "grab, lift, move and place" three 1000 g containers located on waist level and 25 cm above waist level shelves, using their affected arm, at a metronome pace of 60 beats per minute for 300 s or until they feel unable to continue. The time to complete Task 1 is measured using a stopwatch. Task 2 (eye-down) is identical to Task 1 except that the two shelves are placed at eye-level and 25 cm below. Task 3 (overhead work) requires a patient to repeatedly screw and unscrew bolts in a sagittal plane oriented plate positioned at eye-level using both arms". More complete description at https://srs-mcmaster.ca/wp-content/uploads/2015/04/FIT-HaNSAProtocol_April2007.pdf

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	2
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	3
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	3
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	4
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	4
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	3-4
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	3-4
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	4
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	4
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	5
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	NA
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	NA

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	NA
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	NA
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	6-7
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	6-7
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	6-10
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	6-10
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	6-10
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	6-10
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	NA
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	11-13
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	14-16
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	16
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	18

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: www.prisma-statement.org.