BMJ Open

Psychometric properties of the global rating of change scales in patients with neck disorders: a systematic review with meta-analysis and meta-regression

Pavlos Bobos ,[1,2] Joy MacDermid,[1,3] Goris Nazari,[1] Rochelle Furtado,[1] CATWAD

To cite: Bobos P, MacDermid J, Nazari G, *et al*. Psychometric properties of the global rating of change scales in patients with neck disorders: a systematic review with meta-analysis and meta-regression. *BMJ Open* 2019;**9**:e033909. doi:10.1136/bmjopen-2019-033909

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2019-033909).

Received 27 August 2019
Revised 28 October 2019
Accepted 30 October 2019

Check for updates

© Author(s) (or their employer(s)) 2019. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

[1]Western's Bone and Joint Institute, School of Physical Physical Therapy, Department of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada
[2]Institute of Health Policy Management and Evaluation, Department of Clinical Epidemiology and Health Care Research, University of Toronto, Toronto, Ontario, Canada
[3]School of Rehabilitation Science, McMaster University, Hamilton, Ontario, Canada

Correspondence to
Dr Pavlos Bobos;
pbobos@uwo.ca

ABSTRACT

Objective The purpose of this systematic review was to critically appraise and synthesise the psychometric properties of Global Rating of Change (GROC) scales for assessment of patients with neck pain.

Design Systematic review.

Data sources A search was performed in four databases (MEDLINE, EMBASE, CINAHL, SCOPUS) until February 2019.

Data extraction and synthesis Eligible articles were appraised using Consensus-based Standards for the selection of health Measurement Instruments checklist and the Quality Appraisal for Clinical Measurement Research Reports Evaluation Form.

Results The search obtained 16 eligible studies and included in total 1533 patients with neck pain. Test–retest reliability of global perceived effect (GPE) was very high (intraclass correlation coefficient=0.80 to 0.92) for patients with whiplash. Pooled data of Pearson's r indicated that GROC scores were moderately correlated with neck disability change scores (0.53, 95% CI: 0.47 to 0.59). Pooled data of Spearman's correlations indicated that GROC scores were moderately correlated with neck disability change scores (0.56, 95% CI: 0.41 to 0.68).

Conclusions This study found excellent quality evidence of very good-to-excellent test–retest reliability of GPE for patients with whiplash-associated disorders. Evidence from very good-to-excellent quality studies found that GROC scores are moderately correlated to an external criterion patient-reported outcome measure evaluated pre-post treatment in patients with neck pain. No studies were found that addressed the optimal form of GROC scales for patients with neck disorders or compared the GROC to other options for single-item global assessment.

PROSPERO registration number CRD42018117874.

Strengths and limitations of this study

- We rated the quality of individual studies and the overall risk of bias using two standardised approaches.
- Our focus on neck pain increased the specificity of results but are not necessarily applicable to other musculoskeletal conditions.
- Conceptual concerns about global ratings of change being affected by recall bias are not adequately addressed by psychometric evidence.
- No studies addressing the optimal form of global rating were found.

INTRODUCTION

Neck pain is the fourth leading cause of disability and approximately half of the adult population with neck pain will experience a clinically important episode once in their lifetime.[1–3] The annual prevalence of neck pain is estimated between 15% and 50%, with women having a higher prevalence rate than men.[2 3] Neck pain has been associated with many other comorbidities such as headaches, dizziness, anxiety, depression, back pain and arthralgias.[3–6] Several different methods for classifying neck pain have been described, using indicators such as duration (acute, subacute or chronic), degree of interference (low, moderate, severe) or most likely structure at fault (eg, neuropathy vs mechanical).[7]

As part of a patient-centric approach to care, clinicians will commonly evaluate response to intervention by asking the patient directly whether they feel better, worse or the same since the prior encounter. While direct questioning can provide a qualitative indicator of change in status, many best practice guidelines endorse use of some form of quantified patient-reported outcome (PRO) as an adjunct to oral self-report. PROs are available to quantify several different constructs in people with neck pain, including pain severity, disability and neck function.[8] Any PRO intended to provide an estimate of change over time should be responsive to subtle shifts in the patient's condition. To facilitate interpretation of change scores, a common property of many such tools is the minimum clinically important difference (MCID), which is a change threshold that corresponds to the minimum shift in scale values that most patients would indicate

BMJ Open: first published as 10.1136/bmjopen-2019-033909 on 25 November 2019. Downloaded from http://bmjopen.bmj.com/ on May 13, 2025 at Department GEZ-LTA Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

corresponds to an important change in their overall condition. A well-recognised approach to establishing an MCID for a PRO is to compare the magnitude of change against an anchor, most commonly a Global Rating of Change (GROC) scale. These scales allow patients or study participants to indicate whether their condition has gotten worse, better or stayed the same and to quantify the magnitude of that change. As they have been adopted as a sort of 'standard' against which change in other tools is compared, the GROC can also be used on its own as an omnibus generic indicator of change.[8]

Despite being accepted as a standard measure, there is considerable variation in how the GROC has been constructed and implemented in research in neck pain. GROC scales consist of ordered categories which may have different ranked levels (some have 15 levels, some 11 levels and others have 7 levels). The common structure across these is the use of a middle '0' score corresponding to 'no change', with negative values indicating magnitudes of worsening while positive values indicate improvement.[9] Variations of the GROC (in name or structure) include the Global Perceived Effect (GPE), Patient Global Impression of Change (PGIC), Transition Ratings and Global Scale.[9]

A well-established component of health outcomes is having a tool with strong psychometric properties of validity, reliability and responsiveness to be able to monitor change. While recent research[8] has examined the psychometric properties of the most commonly reported PROs for neck disorders, to date there has been no systematic review to summarise the measurement properties of GROC scales themselves in patients with neck disorders. Therefore, this systematic review aims to critically appraise and synthesise the psychometric properties of the GROC scales 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 GROC scales in patients with neck disorders.

Eligibility criteria

We included studies in this systematic review if the following criteria were met[10–12]:

1. Design: psychometric testing, randomised/ cohort studies.
2. Participants:>50% of the study's patient population with neck conditions/disorders.
3. Intervention/comparison: studies that reported on the psychometric properties (reliability, validity, responsiveness) of GROC, GPE and PGIC.
4. Outcomes: GROC, GPE and PGIC.
5. Articles were written in English language only.

Studies with no data on the GROC scale's psychometric properties, and conference abstract/posters were excluded from this systematic review.

Information sources

To identify studies on the psychometric properties (reliability, validity, responsiveness) of the GROC, GPE and PGIC, we searched the MEDLINE, EMBASE, SCOPUS and CINAHL databases from inception till February 2019, using a combination of keywords. Furthermore, we identified additional studies by examining the reference list of each of the selected studies. The full list with keyword strategy is presented in online supplementary appendix 1.

Study selection

Two investigators (PB and GN) performed the systematic electronic searches independently in each database. The same investigators then proceeded to identify and remove the duplicate studies. In the next stage, we performed the independent screening of the titles and abstracts and any full-text article marked as include or uncertain were obtained. In the final stage, the same two independent authors performed the full-text reviews independently to assess final article eligibility. In case of disagreement, a third reviewer, the most experienced member (JM), facilitated a consensus through discussion.

Data extraction

The fourth author (RF) performed the data extractions. The extracted data were then crosschecked by another author (PB). Data extraction included the author, year, study population/condition, setting, sample size, age, properties evaluated, retest-interval and the intervention protocol (if used to assess responsiveness parameters).[13 14] For reliability estimates, standard error of measurement (SEM), intraclass correlation coefficient (ICC), minimal detectable change and 95% CIs were extracted.[13 14] 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.[15] For validity estimates, correlation coefficient (Pearson's/Spearman) and the 95% CIs were extracted.[13 14] 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'.[16] For responsiveness estimates, the effect size, standardised response mean, clinically important difference and/or MCID including the method of MCID estimation-based, anchor-based or distribution-based methods and 95% CIs were extracted.[13 14] 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.[17] When insufficient data were presented, PB contacted the authors by email and requested further data.

BMJ Open: first published as 10.1136/bmjopen-2019-033909 on 25 November 2019. Downloaded from http://bmjopen.bmj.com/ on May 13, 2025 at Department GEZ-LTA Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

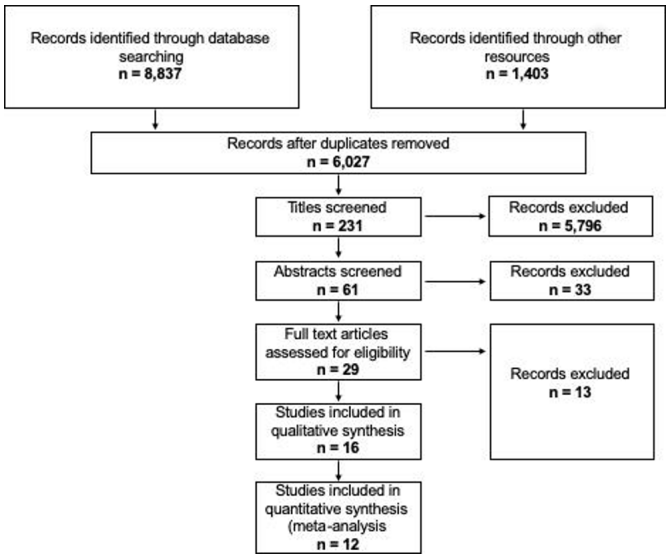

Figure 1 Flow diagram of the included studies.

Consensus-based standards for the selection of health measurement instruments (COSMIN)

COSMIN assesses the risk of bias for the psychometric properties reported on a property-by-property basis. A score for the risk of bias in estimates of psychometric properties was assessed by two authors (PB) and (RF) using the new (COSMIN) checklist.[18] If disagreement was present, a third person (JM) assist in resolving the discrepancy. Each study was assessed by COSMIN on the 4-point scale as 'very good', 'adequate', 'doubtful' or 'inadequate' for each of the checklist criteria for relevant measurement properties (eg, reliability, responsiveness, and so on). According to COSMIN, when determining 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.[19] We then assessed the result of individual studies on a measurement property against the updated criteria for good measurement properties. This involved the evaluation of results of the included studies as either sufficient (+), insufficient (−), or indeterminate (?).[18]

Quality appraisal for clinical measurement research reports evaluation form

A summary score for the overall quality of individual studies was appraised independently by the authors (PB) and (RF) using a structured clinical measurement-specific appraisal tool.[13 14] In case of disagreement, a third person was consulted (JM) to resolve the conflict. The evaluation criteria of this tool included 12 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 standardised; (9) data were presented for each hypothesis; (10) appropriate statistics-point estimates; (11) appropriate statistical error estimates; and (12) valid conclusions and recommendations.[13 14] 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%.[13 14] Overall, the quality summary of appraised articles range from poor (0%–30%), fair (31%–50%), good (51%–70%), very good (71%–90%) and excellent (>90%).[13 14]

Synthesis of results

A qualitative synthesis was conducted to report findings on test–retest reliability statistics. A meta-analysis of Pearson's and Spearman's correlation was performed in R V.3.6.1 with metaphor package.[20] The meta-analyses were conducted using a random effect model and the correlation coefficients were converted to z values. Heterogeneity was deemed substantial if I^2 values were more than 50%.[21] A meta-regression was planned to explore the sources of unexplained heterogeneity by considering the following factors: (1) neck pain with or without radicular symptoms, (2) acute or chronic, (3) age and (4) sex. Forest plots were created using means and 95% CIs for correlation coefficients. We summarise the main results of the included articles based on the neck disorders, reported psychometric estimate and the study quality ratings.

RESULTS

Study selection

Our search yielded 8837 articles. After removal of duplicates, 6027 studies remained and were screened using their title and abstract; leaving 29 articles selected for full-text review. Of these, 16 studies were considered eligible.[22–37] The flow of the study selection process is presented in figure 1.

Study characteristics

The 16 eligible studies were conducted between 2006 and 2017 and included 1533 participants with neck pain/disorders (mean of 96 participants per study).[22–32 34–37] Study size ranged from 29 to 200 participants. A summary description of all the studies included is displayed in table 1. Concurrent validity was evaluated in 14 studies by comparing the difference of pain intensity, disability and function scores with the score of GROC scales. Two studies[26 31] examined the test–retest reliability of a 7-point and an 11-point GPE scale for patients with whiplash-associated disorders (WADs). One study[24] examined whether occurrences of within-session and between-session changes were significantly associated with functional outcomes, pain and self-report of recovery in patients at discharge who were treated with manual therapy for mechanical neck pain.

BMJ Open: first published as 10.1136/bmjopen-2019-033909 on 25 November 2019. Downloaded from http://bmjopen.bmj.com/ on May 13, 2025 at Department GEZ-LTA Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Table 1 Study characteristics

Study	Population	Setting	Sample size	Properties evaluated	GROC evaluated (ranked categories)	Interval
Bjorklund *et al*[32]	Women with non-specific neck–shoulder pain	Not specified	104	Validity (correlation) between NDI and GROC	GROC (7) 1. Very much worse; 2. Much worse; 3. Minimally worse; 4. No change; 5. Minimally improved; 6. Much improved; 7. Very much improved	GROC scale administered only after intervention at one time point (1 week)
Cleland *et al*[22]	Patients with cervical radiculopathy	Hospital	38	Validity (correlation) between NDI and GROC between PSFS and GROC	GROC (15) −7 (a very great deal worse) to 0 (about the same) to +7 (a very great deal better)	GROC was completed at follow-up. Within a week over the period of 7 weeks.
Cleland *et al*[23]	Patients with neck pain only	Five outpatient physical therapy clinics	137	Validity (correlation) between NDI and GROC between NPRS and GROC	GROC (15) −7 (a very great deal worse) to 0 (about the same) to +7 (a very great deal better)	GROC was completed at follow-up. Within a week
Cook *et al*[24]	Patients with any neck pain	Academic locations in Northeast Ohio	56	ROC curves and AUC to measure sensitivity and specificity. Binomial logistic regression analysis was also calculated to determine overall effect	GROC (15) −7 (a very great deal worse) to 0 (about the same) to +7 (a very great deal better)	Baseline and at follow-up 48 and 96 hours post baseline
Farooq *et al*[38]	Patients with neck pain	Physical therapy clinics	106	Validity (correlation) between NDI-U and GROC	GROC (15) −7 (a very great deal worse) to 0 (about the same) to +7 (a very great deal better)	GROC was completed at 3 weeks after intervention
Guzy *et al*[25]	Patients with neck pain	Outpatient rehabilitation clinic	95	Validity (correlation) between NDI-P and GROC	GROC (7) 'complete recovery' over 'no change' to 'my complaints are worse than ever'	GROC scale was completed at 2 weeks and at 4 weeks
Jorritsma *et al*[34]	Patients with chronic non-specific neck pain	Tertiary university centre for rehabilitation	76	Validity (correlation) between NDI and GROC between NPAD and GROC	GPE (7) 3 (completely recovered) to 0 (no change) to −3 (worse than ever)	After completion of the programme varying from 3 to 5 months patients filled the GPE
Kamper *et al*[26]	Patients with any WAD	Physical therapy clinics	134	Test–retest reliability	GPE (11) −5 (vastly worse) to 0 (unchanged) to +5 (completely recovered)	Baseline, 6 weeks, and 12 months
Monticone *et al*[35]	Patients with chronic neck pain	Outpatient Rehabilitation Unit	153	Validity (correlation) between NeckPix and GPE	GPE (5) (helped a lot=1, helped=2), one no change level (helped only a little=3), and two worsening levels (did not help=4, made things worse=5)	At the end of treatment (8 weeks) and 1 year before follow-up
Monticone *et al*[36]	Patients with chronic neck pain	Outpatient Rehabilitation Unit	200	Validity (correlation) between NDI and GPE between NPDS and GPE	GPE (5) (helped a lot=1, helped=2), one no change level (helped only a little=3), and two worsening levels (did not help=4, made things worse=5)	At the end of treatment 8 weeks
Ngo *et al*[31]	Patients with WAD. Most participants (69.6%) had grade II WAD.	Interviewed by person or by telephone in Ontario	46	Test–retest reliability	GPE (7) 1.General recovery question: completely better, much improved, slightly improved, no change, slightly worse, much worse, worse than ever 2.Change in neck pain question: very much better, better, slightly better, no change, slightly worse, worse, or very much worse	3–5 days
Shaheen *et al*[27]	Patients with neck pain lasting more than 3 months	Three primary health centres	70	Validity (correlation) between NDI-Ar and GROC	GROC (15) −7 (a very great deal worse) to 0 (about the same) to +7 (a very great deal better)	1 week

Continued

BMJ Open: first published as 10.1136/bmjopen-2019-033909 on 25 November 2019. Downloaded from http://bmjopen.bmj.com/ on May 13, 2025 at Department GEZ-LTA Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Table 1 Continued

Study	Population	Setting	Sample size	Properties evaluated	GROC evaluated (ranked categories)	Interval
Takeshita *et al*[28]	Patients with neck pain, cervical radiculopathy and/or cervical myelopathy	Variety of clinics and hospital settings	130	Validity (correlation) between NDI-J and GROC	PGIC (7) much better, better, slightly better, unchanged, slightly worse, worse and much worse	Over 8 weeks
Trouli *et al*[29]	Patients with neck pain	Primary healthcare clinic	68	Validity (correlation) between NDI-Gr and GROC	GROC (15) −7 (a very great deal worse) to −1 (almost the same, hardly any worse at all) and from 7 (a very great deal better) to 1 (almost the same, hardly any better at all)	Within 2 months, but 1 week for test–retest
Tuttle *et al*[30]	Patients with neck pain for more than 2 weeks	Private physiotherapy clinics	29	Validity (correlation) between NDI and GPE between PSFS and GPE between VAS and GPE between ROM and GPE	GPE (11) −5 is vastly worse and +5 is completely recovered	6 weeks
Young *et al*[37]	Patients presenting with mechanical neck pain	Outpatient physical therapy clinics.	91	Validity (correlation)	GROC (15) −7 ('a very great deal worse') to 0 ('about the same') to +7 ('a very great deal better')	3 weeks

AUC, area under the curve;GPE, global perceived effect; GROC, Global Rating of Change; NDI, neck disability index;NDI-GR, Greek version of Neck Disability Index; NDI-J, Japanese version of Neck Disability Index; NDI-P, Polish version of the Neck Disability Index; NDI-U, Neck Disability Index - Urdu language; NPAD, Neck Pain and Disability Scale;NPDS, Neck Pain Disability Scale; NPRS, Numeric Pain Rating Scale; PSFS, Patient-Specific Functional Scale; ROC, receiver operator characteristic; ROM, range of motion; VAS, Visual Analogue Scale; WAD, whiplash-associated disorder.

COSMIN risk of bias rating and quality appraisal of the included studies

Regarding the risk of bias, all studies were rated as very good (table 2). The quality of the studies ranged from 88% to 96% (table 3). The most common flaws were (1) lack of/inadequate sample size calculations; (2) missing data (ie, inadequate follow-up) and (3) inconsistencies between the data presented and hypothesis stated.

Reported GROC scales

The most commonly reported GROC scale (n=6 studies) was a 15-point scale with the most frequent anchors being '−7 (a very great deal worse) to 0 (about the same) to +7 (a very great deal better)'. A 7-point scale was reported in five studies, 11-point and 5-point scales were reported in two studies and a 9-point scale in one study. The anchors in those scales varied greatly and are presented in table 1. Only six studies[26 31–33 35 36] reported full details regarding the specific questions asked of the patients with neck disorder when a GROC scale was administered. Those questions that were reported are presented in table 4.

Reliability measures

Two studies were included that examined test–retest reliability of GPE for patients with WAD. Kamper *et al*[26] examined the (time interval) test–retest reliability of an 11-point GPE scale in 134 patients with chronic WAD and reported an ICC of 0.99 (95% CI: 0.99 to 0.99) at baseline, 0.96 (0.95 to 0.97) at 6 weeks and 0.92 (0.89 to 0.94) at 12 months (table 5). Ngo *et al*[31] assessed the test–retest reliability of a 7-point scale of GPE in patients with acute WAD at 3 to 5 days.[31] The ICC and 95% CIs were used to determine the test–retest reliability of the two versions of the perceived recovery questions using their original 7-item responses. Ngo *et al* also computed weighted kappa coefficients and 95% CI using quadratic weights to determine whether the distribution of responses influenced the reliability as measured by the ICC. An ICC for general recovery of 0.70 (0.60 to 0.80) and an ICC for neck pain questions of 0.80 (0.72 to 0.87) were found. A weighted kappa was also calculated (kappa=0.70 (0.42 to 0.98)) at 6 weeks for general recovery and at 6 weeks kappa=0.80 (0.51 to 1.0) for neck pain questions (table 5).

Validity measures

We found 14 studies that examined concurrent validity measures between GROC and another PRO.[22 23 25–27 30–32 34–38] Correlations of Pearson's and Spearman's coefficients between GROC and another PRO were ranging from very weak to very strong correlations. The validity measures are presented and summarised in table 6.

Meta-analysis and meta-regression of correlations between disability change scores and GROC scores

Five studies[23 25 34 37 38] of very good-to-excellent quality reported the Pearson correlation coefficients between neck disability change scores and the GROC scores were pooled together. We found that GROC was positively correlated

BMJ Open: first published as 10.1136/bmjopen-2019-033909 on 25 November 2019. Downloaded from http://bmjopen.bmj.com/ on May 13, 2025 at Department GEZ-LTA Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Table 2 Summary of psychometric properties reported in studies and COSMIN ROB and quality studies

Study	Psychometric properties reported	COSMIN ROB	COSMIN rating* (criteria)	Quality of studies (QACMRR)
Bjorklund *et al*[32]	Validity (correlation)	Very good	?	Excellent
Cleland *et al*[22]	Validity (correlation)	Very good	+	Excellent
Cleland *et al*[23]	Validity (correlation)	Very good	–	Excellent
Cook *et al*[24]	Sensitivity Specificity	Very good Very Good	+	Excellent
Farooq *et al*[38]	Validity (correlation)	Very good	+	Excellent
Guzy *et al*[25]	Validity (correlation)	Very good	?	Very good
Jorritsma *et al*[34]	Validity (correlation)	Very good	?	Excellent
Kamper *et al*[26]	Test–retest reliability	Very good	+	Excellent
Monticone *et al*[35]	Validity (correlation)	Very good	?	Excellent
Monticone *et al*[36]	Validity (correlation	Very good	?	Excellent
Ngo *et al*[31]	Test–retest reliability	Very good	+	Excellent
Shaheen *et al*[27]	Validity (correlation)	Very good	?	Excellent
Takeshita *et al*[28]	Validity (correlation)	Very good	?	Very good
Trouli *et al*[29]	Validity (correlation)	Very good	+	Excellent
Tuttle *et al*[30]	Validity (correlation)	Very good	?	Excellent
Young *et al*[37]	Validity (correlation)	Very good	?	Excellent

Criteria for good measurement properties: '+' sufficient; '–' insufficient; '?' indeterminate.
*The grading for the quality of the evidence based on the modified GRADE approach is not applicable.
COSMIN, Consensus-based Standards for the Selection of Health Measurement Instruments; QACMRR, Quality Appraisal for Clinical Measurement Research Reports Evaluation Form; ROB, risk of bias.

with disability change scores (r=0.53, 95% CI: 0.47 to 0.59, I^2=0%). Six studies[27–30 32 36] of very good-to-excellent quality reported the Spearman correlation coefficients between neck disability changes scores and the GROC scores and were pooled together. We found that GROC was moderately correlated with disability change scores (rho=0.56, 95% CI: 0.41 to 0.68, I^2=85%). The forest plots with correlation coefficients with 95% CIs are presented in figures 2–3. Our meta-regression showed that age was found as a significant factor in influencing Fisher's Z scores (β=–0.034, 95% CI: –0.05 to –0.01, p=0.001). The model explained 68% of the variance (R^2=0.68) (figure 4).

Area under the curve (AUC)—sensitivity and specificity

Cook *et al*[24] found that between-session NPRS pain changes were associated with greater than 3-point change on the GROC at 96 hours (AUC=0.76). The pain change associated with GROC was more specific (specificity=79.2%, range: 62.2–91.1) than sensitive (sensitivity=65.6%, range: 57.9 to 74.6). Those with a 36.7% between-sessions change in pain were also 7.3 times more likely to report an improvement of greater than 3-point change on the GROC than those who did not achieve a 36.7% change in pain (table 5).

DISCUSSION

This review has synthesised the current research from 16 studies that aimed to evaluate the psychometric properties of GROC scales for patients with neck disorders, with the goal to provide evidence for clinicians and researchers concerning its use within clinical practice and research. From the 16 included studies, only two studies[26 31] reported test–retest reliability statistics of the 7-ranked and 11-ranked categories of GPE scales for patients with WAD only. We were able to pool data from 12 studies regarding concurrent validity of GROC scales and neck disability change scores at one time point after the interventions. Themes influencing interpretation of the GROC were explored in a study[33] that evaluated the factors that contribute to how patients respond to a question on GPE. This study found that treatment process, biomechanical performance, self-efficacy and the nature of the condition may influence the responses on GPE, which is consistent with what we would expect for patients with neck pain. This suggests that change is a complex multifactorial global concept. A strength of GROC is that it is intended as a global assessment, and it can be assumed that it reflects the aspects of change important to the individual patient.

Reliability can be defined as the degree to which a measure produces consecutive results with the least amount of random error when the status of the population remains unchanged. The reliability of GPE displayed an excellent test–retest reliability of ICC>0.90 over an interval of 6 weeks and 12 months for patients with WAD. Conducting an assessment with a long test–retest interval (eg, 12 months) can provide challenges as there

BMJ Open: first published as 10.1136/bmjopen-2019-033909 on 25 November 2019. Downloaded from http://bmjopen.bmj.com/ on May 13, 2025 at Department GEZ-LTA Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Table 3 Quality Appraisal for Clinical Measurement Research Reports Evaluation Form

Study	Item evaluation criteria* 1	2	3	4	5	6	7	8	9	10	11	12	Total (%)	Quality summary
Bjorklund *et al*[32]	2	2	2	2	2	1	2	2	2	2	2	2	96	Excellent
Cleland *et al*[23]	2	2	2	2	1	2	2	2	2	2	2	2	96	Excellent
Trouli *et al*[29]	2	2	2	2	1	2	2	2	2	2	2	2	96	Excellent
Tuttle *et al*[30]	2	2	2	2	1	2	2	2	2	2	2	2	96	Excellent
Kamper *et al*[26]	2	2	2	2	1	2	2	2	2	2	2	2	96	Excellent
Cook *et al*[24]	2	2	2	2	1	2	2	2	1	2	2	2	92	Excellent
Jorritsma *et al*[34]	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
Cleland *et al*[22]	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
Monticone *et al*[35]	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
Monticone *et al*[36]	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
Ngo *et al*[31]	2	2	2	2	2	2	2	2	1	2	1	2	92	Excellent
Shaheen *et al*[27]	2	2	2	2	2	2	2	2	2	2	1	1	92	Excellent
Farooq *et al*[38]	2	2	1	2	2	2	2	2	1	2	2	2	92	Excellent
Young *et al*[37]	2	2	2	2	1	1	2	2	2	2	2	2	92	Excellent
Guzy *et al*[25]	2	2	1	2	1	2	2	2	1	2	2	2	88	Very good
Takeshita *et al*[28]	2	2	1	1	1	2	2	2	2	2	2	2	88	Very good

Total score=(sum of subtotals ÷ 24×100). If for a specific paper an item is deemed not applicable, then, total score = (sum of subtotals ÷ (2×number of applicable items)×100).
The subsection 6, asks for percentage of retention/follow-up. This subsection only applies to reliability test-retest studies.
Quality summary: poor (0%–30%), fair (31%–50%), good (51%–70%), very good (71%–90%), excellent (>90%).
*Item evaluation criteria: (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 standardised; (9) data were presented for each hypothesis; (10) appropriate statistics point estimates; (11) appropriate statistical error estimates; (12) valid conclusions and clinical recommendations.

is higher risk of individuals with WAD being symptomatically unstable.[9] Determining if patients are symptomatically stable can be achieved by administering another PRO such as the Single Assessment Numeric Evaluation (SANE)[39]; however, the 7-ranked and 11- ranked categories of GPE scales still demonstrated good stability properties at long test intervals (ie, of 6 weeks and 12 months).[26] Therefore, the measurements of the reliability parameters of the GPE may be very useful during longer test intervals in clinical trials.

The psychometric property of validity is defined as the degree to which a PRO measures what it is intended to

Table 4 Questions of GROC scales

Author	GROC (ranked categories)	Patients with neck disorders were asked:
Bjorklund *et al*[32]	GROC (7)	'Compared with before the treatment of the study started, my overall status is now' Compared with before the treatment of the study started, my status regarding my neck–shoulder problem is now'
Evans *et al*[33]	GPE (9)	'Overall, how much has your neck pain changed since you started treatment in the study?'
Kamper *et al*[26]	GPE (11)	'With respect to your whiplash injury how would you describe yourself now compared with immediately after your accident'
Monticone *et al*[35]	GPE (5)	'Overall, how much did the treatment you received help your fear of movement due to current neck pain?' 'Overall, how much did the treatment you delivered help your subject's fear of movement due to her/ his current neck pain?'
Monticone *et al*[36]	GPE (5)	'Overall, how much did the treatment you received help your neck problem?'
Ngo *et al*[31]	GPE (7)	'How well do you feel you are recovering from your injuries?' 'How do you feel your neck pain has changed since the injury?'

GPE, global perceived effect; GROC, Global Rating of Change.

BMJ Open: first published as 10.1136/bmjopen-2019-033909 on 25 November 2019. Downloaded from http://bmjopen.bmj.com/ on May 13, 2025 at Department GEZ-LTA Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Table 5 Summary of reliability properties of GROC scales

Study	Type of reliability	Reliability estimates	COSMIN	Quality of studies
Kamper *et al*[26]	Test–retest	ICC baseline: 0.99 (0.99–0.99) at 6 weeks: 0.96 (0.95–0.97) at 12 months: 0.92 (0.89–0.94).	Very good	Excellent
Ngo *et al*[31]	Test–retest	ICC at 6 weeks (general recovery): 0.70 (0.60–0.80) at 6 weeks (neck pain questions): 0.80 (0.72–0.87) Weighted kappa at 6 weeks (general recovery): 0.70 (0.42–0.98) at 6 weeks (neck pain questions): 0.80 (0.51–1.0) dichotomised response options for recovery (K statistics) 0.85 (0.64–1) when 'recovered' was defined 'completely better' 0.81 (0.64–0.99) when defined as 'completely better' or 'much improved' Dichotomised response options for change in neck pain questions (K statistics) 0.46 (0.20–0.74) when 'recovered' was defined as 'very much better' 0.80 (0.62–0.99) when defined as 'very much better' or 'better' Recall questions (K statistics) the kappa coefficient was 1 for participants who remembered their previous answers to the general recovery question; 0.88 (0.64–1) for those who did not remember and 0.50 (0.02–0.98) for participants who were not asked the question. The kappa coefficient was 1 for participants who remembered their previous answers to the change in neck pain question; 0.74 (0.41–1) for those who did not remember and 0.66 (0.22–1) for participants who were not asked the question.	Very Good	Excellent

GROC, Global Rating of Change; ICC, intraclass correlation coefficient.

measure. Pooled data from 11 studies overall suggest that post-treatment changes of validated disability outcome measures were moderately (Pearson's r=0.51, 95% CI: 0.43 to 0.58; Spearman's rho=0.56, 95% CI: 0.41 to 0.68) correlated to change in perceived effect (figures 2–3). This finding suggests that GROC scores taken at one point in time were related to scores in pain and disability in patients with neck disorders, as measured by standardised measures taken at two points in time. We identified one study[24] that found a 36.7% change in pain for within-session and between-session changes was associated with a 50% reduction in the NDI and an improvement of >3 levels on a 15-ordinal level GROC scale for patients with neck pain. This quantified predictive change value may have clinical utility for use in clinical practice.

Previous studies[9 40] have indicated serious concerns about the conceptual validity of the global rating of change. The review by Kamper *et al*[9] clearly showed that GROC was related to final status more than change and was least related to baseline health status. This result undermines the premise of what the global rating of change actually measures. For this reason, we conclude that the 0.50 pooled correlation across 12 studies between the GROC and other patient-reported outcome measure (PROM) change scores (eg, NDI scores) may reflect a relationship between follow-up status and change rather than supporting the contention that GROC actually measures change. This would also explain why only 25% of the variation in GROC change scores was explained by change scores from a PROM change score measured at two points in time. In all studies, participants completed the GROC scale at one time point after the intervention, and hence recall bias is a cause for concern. However, another potential factor for moderate correlations is that the PROMs that have been used as the comparator with GROC scores may not reflect priorities that are important to patients. That is, the field has largely been driven by assumptions that the GROC is a 'gold standard' for evaluating true change in a respondent's condition or status, and that all items on the comparator PROM are of equal importance to all people with that condition. The work presented herein challenges the valorisation of the GROC as a gold standard for change, and prior work has challenged the notions that all PROM items are

BMJ Open: first published as 10.1136/bmjopen-2019-033909 on 25 November 2019. Downloaded from http://bmjopen.bmj.com/ on May 13, 2025 at Department GEZ-LTA Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Table 6 Summary of validity properties of GROC scales

Study	Type of reliability	Validity estimates	COSMIN	Quality of studies
Björklund *et al*[32]	Spearman's correlation between the change scores of GROC and ProFitMap-neck GROC and NDI	rho=0.47 (p<0.05) rho=0.59 (p<0.05)	Very good	Excellent
Cleland *et al*[22]	Correlations (Pearson r) between change scores NDI and GROC PSFS and GROC	r=0.19 r=0.82	Very good	Excellent
Cleland *et al*[23]	Correlations (Pearson r) between change scores NDI and GROC NRS and GROC	r=0.58 r=0.57	Very good	Excellent
Cook *et al*[24]	ROC Within-session change Between-session change Between-session change of pain and GROC Sensitivity Specificity	AUC=0.61 AUC=0.76,>36.7% change in pain OR=7.3 (2.1, 24.7) 65.6% (57.9, 74.6) 79.2% (62.2, 91.1)	Very good	Excellent
Farooq *et al*[38]	Correlations (Pearson r) NDI-U	r=0.50	Very good	Excellent
Guzy *et al*[25]	Correlations (Pearson r) NDI vs GROC	2-week interval (r=−0.73) 4-week interval (r=−0.56)	Very good	Very good
Jorritsma *et al*[34]	Correlation between change scores of NPAD and GPE	r=0.49 (95 % CI 0.30 to 0.64)	Very good	Excellent
Monticone *et al*[35]	Correlations (Spearman) between change scores of the NeckPix and GPE	rho=0.69–0.82	Very good	Excellent
Monticone *et al*[36]	Correlation (Spearman) between change scores NDI-I and GPE NDPS and GPE	rho=0.71, p<0.01 rho=0.59, p<0.01	Very good	Excellent
Shaheen *et al*[27]	Correlations (Spearman's) NDI-Ar and GROC	rho=0.81, p<o.oo1	Very good	Excellent
Takeshita *et al*[28]	Correlations NDI and PGIC NDI-J and PGIC	Spearman (rho) rho=0.47, p<o.oo1 rho=0.59, p<o.oo1	Very good	Very good
Trouli *et al*[29]	Correlation (Spearman's) GROC vs Gr-NDI	rho=0.30, p=0.02	Very good	Excellent
Tuttle *et al*[30]	Correlations (Spearman's) NDI vs GPE (post 1, minus pre-1) NDI vs GPE (post 2, minus pre-1) NDI vs GPE (post 2, minus pre-2) PSFS vs GPE (post 1, minus pre-1) PSFS vs GPE (post 2, minus pre-1) PSFS vs GPE (post 2, minus pre-2) Pain intensity (post 1, minus pre-1) Pain intensity (post 2, minus pre-1) Pain intensity (post 2, minus pre-2) Total ROM (post 1, minus pre-1) Total ROM (post 2, minus pre-1) Total ROM (post 2, minus pre-2)	rho=0.17 rho=0.01 rho=0.03 rho=0.06 rho=0.03 rho=0.03 rho=0.00 rho=0.05 rho=0.01 rho=0.03 rho=0.01 rho=0.00	Very good	Excellent
Young *et al*[37]	Correlations (Pearson's) between change scores NDI and GROC	r=0.52 (p<0.01)	Very good	Excellent

Continued

BMJ Open: first published as 10.1136/bmjopen-2019-033909 on 25 November 2019. Downloaded from http://bmjopen.bmj.com/ on May 13, 2025 at Department GEZ-LTA Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Table 6 Continued

Study	Type of reliability	Validity estimates	COSMIN	Quality of studies
Monticone *et al*[36]	Correlation (Spearman) between change scores NDI-I and GPE NDPS and GPE	rho=0.71, p<0.01 rho=0.59, p<0.01	Very good	Excellent

AUC, area under the curve; COSMIN, Consensus-based Standards for the Selection of Health Measurement Instruments; GPE, global perceived effect; GROC, Global Rating of Change; NDI, neck disability index; NDI-Ar, Arabic Version of Neck Disability Index; NDI-I, Italian version of Neck Disability Index ; NDI-U, Urdu version of Neck Disability Index; NPAD, Neck Pain and Disability Scale; NPDS, Neck Pain Disability Scale; NRS, Numeric Rating Scale; PSFS, Patient-Specific Functional Scale; ROC, receiver operator characteristic; ROM, range of motion.

equally important.[9 41 42] It is therefore possible that the very constructs being evaluated require greater critical discourse before authors can say, with confidence, that one scale functions well or poorly based on its associations with another scale. Since no studies compared a retrospective global assessment of the GROC to pre-post single item global PROM for example, the SANE, we do not know the extent to which these two factors contributed to moderate correlation.

A unique aspect of this study was that it focused on global rating of change scales in a neck pain patient population. Our study appraisal suggests that future studies concerning GROC should include adequate sample sizes, maintain a rigorous follow-up and report appropriate statistical error estimates, since these were often inadequate. Various critical appraisal tools exist, and the perspectives and ratings may differ across instruments. COSMIN is just one methodology that can be used to synthesise or evaluate outcome measures and other methods might be equally valid or provide different perspectives. We used two different critical appraisal tools to evaluate quality from two perspectives. The COSMIN risk of bias assessments reflects the level of confidence in the conclusions and pooled estimates. The quality appraisal tool focuses on design issues in the studies and reflects gaps in research designs that should be considered in interpretation of current research and improved in future studies. Substantial heterogeneity was detected (I^2 >50%) in pooled Spearman's correlation coefficients which is a concern when pooling data. Sources of the observed heterogeneity were identified in our meta-regression results. Our univariate meta-regression analysis indicated that age across the studies explained 68% of the variance (figure 4). Other factors such as type of neck pain (with or without radicular symptoms), acute or chronic and sex did not explain the remaining heterogeneity (not statically significant). In our meta-regression, we used a patient level characteristic to identify the observed heterogeneity and therefore, our model may be vulnerable to aggregation bias. Furthermore, the scope of our literature search was focused on identifying full-text papers written only in English.

While this study included 16 studies, only 2 of these reported reliability statistics for GROC scales for patients with chronic WAD. Therefore, the applicability of our study is mostly limited to patients with chronic WAD. For validity measurements, GROC scales were mostly investigated by correlation analyses to evaluate the external responsiveness of another PRO measure over a specific time point. From our meta-analysis, we can be confident that the GROC scores were moderately correlated with neck disability change scores. However, more robust psychometric design studies to test the measurement properties of GROC scales as the primary outcome of investigation are highly needed. Future studies should aim to test to what extent the different range of items (eg, 7-level scale vs 11-level scale), the anchors (eg, much worse vs much better) may affect the measurement properties of GROC scales for patients with neck disorders. Also, it is important to indicate that most outcome measures are ordinal and assume that additive scores of ordinal items can be treated as interval

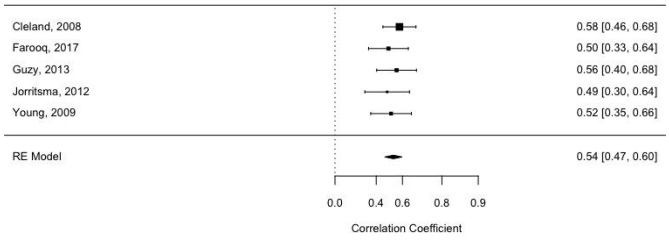

Figure 2 Meta-analysis of Pearson's correlation coefficients between neck disability change scores and GROC scores in patients with neck disorders based on five very good-to-excellent quality studies. GROC, Global Rating of Change.

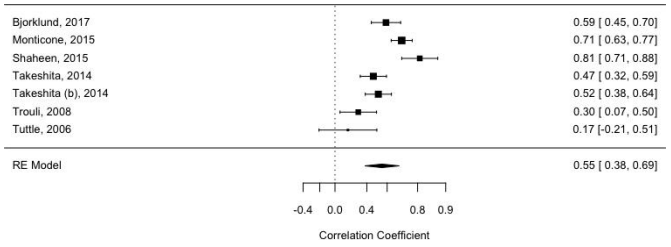

Figure 3 Meta-analysis of Spearman's correlation coefficients between neck disability change scores and GROC scores in patients with neck disorders based on six very good-to-excellent quality studies. GROC, Global Rating of Change.

BMJ Open: first published as 10.1136/bmjopen-2019-033909 on 25 November 2019. Downloaded from http://bmjopen.bmj.com/ on May 13, 2025 at Department GEZ-LTA Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

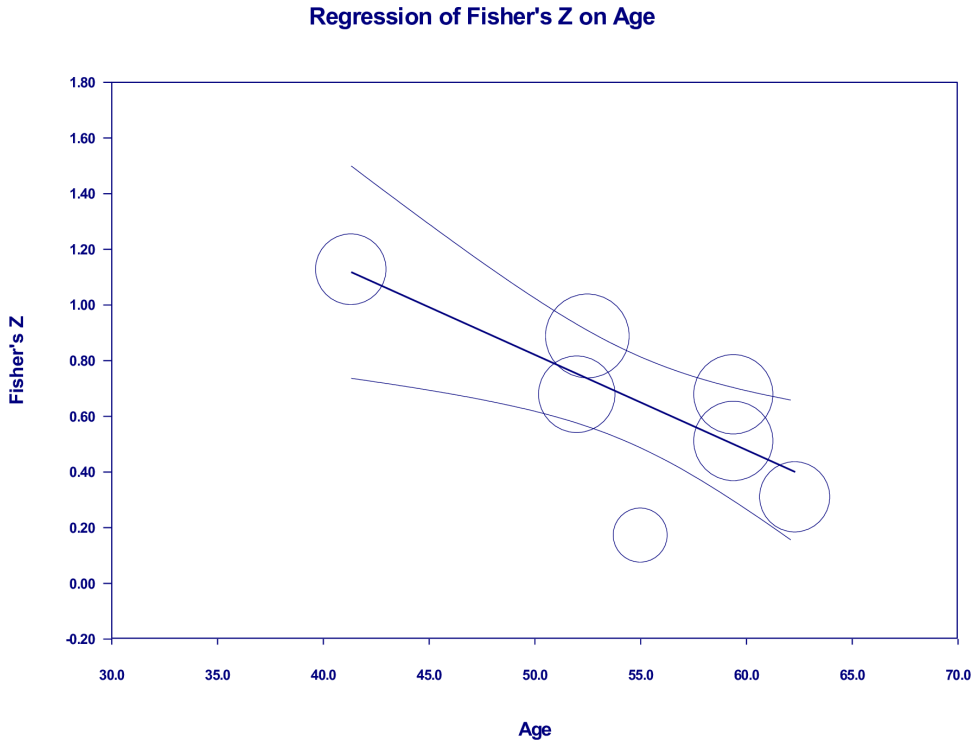

Figure 4 Random effects univariate meta-regression between age and the Fisher's Z estimates. Each circle represents a study and the size of the circle indicates the influence of that study on the model. The regression prediction is illustrated by the straight line and the curved lines represent the 95% CIs. Age explained 68% of the variance in the model (R^2=0.68).

level. This potentially could lead to scaling problems even in the face of strong psychometric properties. The main protection we have is to create new scales or retrofit existing scales based on Rasch analysis. Also, we acknowledge that the majority of work done on the GROC scales has been performed using statistical approaches that are most appropriate to linear rather than ordinal data.

CONCLUSIONS

This study found excellent quality evidence of very good-to-excellent test–retest reliability of GPE for patients with WAD. Evidence of very good-to-excellent quality studies found that GROC scores are moderately correlated to an external criterion PROM, measured pre–post treatment in patients with neck disorders. Studies addressing the optimal form of GROC scales for patients with neck disorders or comparing the GROC to other options for single-item global assessment of change were not found.

Twitter Pavlos Bobos @pavlosbob

Collaborators CATWAD co-authors: Michele Sterling, Anne Söderlund, Michele Curatolo, James M Elliott, David M Walton, Helge Kasch, Linda Carroll, Hans Westergren, Gwendolen Jull, Eva-Maj Malmström, Luke B Connelly, Joy C MacDermid, Mandy Nielsen, Pierre Côté, Tonny Elmose Andersen, Trudy Rebbeck, Annick Maujean, Sarah Robins, Kenneth Chen, Julia Treleaven.

Contributors PB contributed significantly to the conception and design of the study, data extraction, critical appraisal, interpretation of data and drafting of the manuscript. GN and RF were involved in literature search, critical appraisal and interpretation of data and drafting. GN 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. JM 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.

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

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

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

ORCID iD

Pavlos Bobos http://orcid.org/0000-0002-5098-4840

REFERENCES

1 Murray CJL, Atkinson C, Bhalla K, *et al*. The state of US health, 1990-2010: burden of diseases, injuries, and risk factors. *JAMA* 2013;310:591–608.
2 Fejer R, Kyvik KO, Hartvigsen J. The prevalence of neck pain in the world population: a systematic critical review of the literature. *Eur Spine J* 2006;15:834–48.
3 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* 2009;32:S46-60.
4 Bobos P, Nazari G, Palimeris S, *et al*. Contribution of health and psychological factors in patients with chronic neck pain and disability. A cross sectional study. *J Clin Diagnostic Res* 2017;12:YC04–7.

BMJ Open: first published as 10.1136/bmjopen-2019-033909 on 25 November 2019. Downloaded from http://bmjopen.bmj.com/ on May 13, 2025 at Department GEZ-LTA Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

5 MacDermid JC, Walton DM, Bobos P, *et al*. A qualitative description of chronic neck pain has implications for outcome assessment and classification. *Open Orthop J* 2016;10:746–56.
6 Treleaven J. Sensorimotor disturbances in neck disorders affecting postural stability, head and eye movement control--Part 2: case studies. *Man Ther* 2008;13:266–75.
7 Cohen SP, Epidemiology CSP. Epidemiology, diagnosis, and treatment of neck pain. *Mayo Clin Proc* 2015;90:284–99.
8 Bobos P, MacDermid JC, Walton DM, *et al*. Patient-Reported outcome measures used for neck disorders: an overview of systematic reviews. *J Orthop Sports Phys Ther* 2018;48:775–88.
9 Kamper SJ, Maher CG, Mackay G. Global rating of change scales: a review of strengths and weaknesses and considerations for design. *J Man Manip Ther* 2009;17:163–70.
10 Nazari G, Bobos P, MacDermid JC, *et al*. The effectiveness of Instrument-Assisted soft tissue mobilization in athletes, participants without extremity or spinal conditions, and individuals with upper extremity, lower extremity, and spinal conditions: a systematic review. *Arch Phys Med Rehabil* 2019;100:1726–51.
11 Bobos P, Nazari G, Szekeres M, *et al*. The effectiveness of joint-protection programs on pain, hand function, and grip strength levels in patients with hand arthritis: a systematic review and meta-analysis. *J Hand Ther* 2019;32:194–211.
12 Nazari G, Bobos P, MacDermid JC, *et al*. Psychometric properties of the Zephyr bioharness device: a systematic review. *BMC Sports Sci Med Rehabil* 2018;10:6.
13 Law MC, MacDermid J. *Evidence-based rehabilitation : a guide to practice*. Thorofare, NJ: Slack Incorporated, 2014.
14 Roy J-S, Desmeules F, MacDermid JC. Psychometric properties of presenteeism scales for musculoskeletal disorders: a systematic review. *J Rehabil Med* 2011;43:23–31.
15 Sańchez J, Rosner B. Rosner, B.: fundamentals of biostatistics, third edition. PWS-Kent, Boston 1990, XV, 655 pp., us $ 14.95, ISBN 0-534-91973-1. *Biom J* 1993;35.
16 Wuensch KL, Evans JD. Straightforward statistics for the behavioral sciences. *J Am Stat Assoc* 2006.
17 Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd Ed, 1988.
18 Mokkink LB, de Vet HCW, Prinsen CAC, *et al*. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Qual Life Res* 2018;27:1171–9.
19 Terwee CB, Mokkink LB, Knol DL, *et al*. 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.
20 Viechtbauer W. Conducting Meta-Analyses in *R* with the metafor Package. *J Stat Softw* 2010;36:1–48.
21 Higgins JPT, Thompson SG, Deeks JJ, *et al*. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557–60.
22 Cleland JA, Fritz JM, Whitman JM, *et al*. The reliability and construct validity of the neck disability index and patient specific functional scale in patients with cervical radiculopathy. *Spine* 2006;31:598–602.
23 Cleland JA, Childs JD, Whitman JM. Psychometric properties of the neck disability index and numeric pain rating scale in patients with mechanical neck pain. *Arch Phys Med Rehabil* 2008;89:69–74.
24 Cook C, Lawrence J, Michalak K, *et al*. Is there preliminary value to a within- and/or between-session change for determining short-term outcomes of manual therapy on mechanical neck pain? *J Man Manip Ther* 2014;22:173–80.
25 Guzy G, Vernon H, Polczyk R, *et al*. Psychometric validation of the authorized Polish version of the neck disability index. *Disabil Rehabil* 2013;35:2132–7.
26 Kamper SJ, Ostelo RWJG, Knol DL, *et al*. Global perceived effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. *J Clin Epidemiol* 2010;63:760–6.
27 Shaheen AAM, Omar MTA, Vernon H. Cross-Cultural adaptation, reliability, and validity of the Arabic version of neck disability index in patients with neck pain. *Spine* 2013;38:E609–15.
28 Takeshita K, Hosono N, Kawaguchi Y, *et al*. Validity, reliability and responsiveness of the Japanese version of the neck disability index. *J Orthop Sci* 2013;18:14–21.
29 Trouli MN, Vernon HT, Kakavelakis KN, *et al*. Translation of the neck disability index and validation of the Greek version in a sample of neck pain patients. *BMC Musculoskelet Disord* 2008;9:1–8.
30 Tuttle N, Laasko L, Barrett R. Change in impairments in the first two treatments predicts outcome in impairments, but not in activity limitations, in subacute neck pain: an observational study. *Aust J Physiother* 2006;52:281–5.
31 Ngo T, Stupar M, Côté P, *et al*. A study of the test-retest reliability of the self-perceived General recovery and self-perceived change in neck pain questions in patients with recent whiplash-associated disorders. *Eur Spine J* 2010;19:957–62.
32 Björklund M, Wiitavaara B, Heiden M. Responsiveness and minimal important change for the ProFitMap-neck questionnaire and the neck disability index in women with neck-shoulder pain. *Qual Life Res* 2017;26:161–70.
33 Evans R, Bronfort G, Maiers M, *et al*. "I know it's changed": a mixed-methods study of the meaning of Global Perceived Effect in chronic neck pain patients. *Eur Spine J* 2014;23:888–97.
34 Jorritsma W, Dijkstra PU, de Vries GE, *et al*. Detecting relevant changes and responsiveness of neck pain and disability scale and neck disability index. *Eur Spine J* 2012;21:2550–7.
35 Monticone M, Frigau L, Vernon H, *et al*. Responsiveness and minimal important change of the NeckPix© in subjects with chronic neck pain undergoing rehabilitation. *Eur Spine J* 2018;27:1324–31.
36 Monticone M, Ambrosini E, Vernon H, *et al*. Responsiveness and minimal important changes for the neck disability index and the neck pain disability scale in Italian subjects with chronic neck pain. *Eur Spine J* 2015;24:2821–7.
37 Young BA, Walker MJ, Strunce JB, *et al*. Responsiveness of the neck disability index in patients with mechanical neck disorders. *Spine J* 2009;9:802–8.
38 Farooq MN, Mohseni-Bandpei MA, Gilani SA, *et al*. Urdu version of the neck disability index: a reliability and validity study. *BMC Musculoskelet Disord* 2017;18:1–11.
39 Williams GN, Gangel TJ, Arciero RA, *et al*. Comparison of the single assessment numeric evaluation method and two shoulder rating scales. outcomes measures after shoulder surgery. *Am J Sports Med* 1999;27:214–21.
40 Schmitt J, Abbott JH. Global ratings of change do not accurately reflect functional change over time in clinical practice. *J Orthop Sports Phys Ther* 2015;45:106–11.
41 Chiarotto A, Ostelo RW, Boers M, *et al*. A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in patients with low back pain. *J Clin Epidemiol* 2018;95:73–93.
42 Ailliet L, Knol DL, Rubinstein SM, *et al*. Definition of the construct to be measured is a prerequisite for the assessment of validity. the neck disability index as an example. *J Clin Epidemiol* 2013;66:775–82.

BMJ Open: first published as 10.1136/bmjopen-2019-033909 on 25 November 2019. Downloaded from http://bmjopen.bmj.com/ on May 13, 2025 at Department GEZ-LTA Erasmushogeschool . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.